Measuring Trust in the XAI Context

Robert R. Hoffman
Institute for Human and Machine Cognition
Shane T. Mueller
Michigan Technological University
Gary Klein
Macrocognition, LLC
Jordan Litman
Institute for Human and Machine Cognition

This material is approved for public release. Distribution is unlimited. This material is based on research sponsored by the Air Force Research Lab (AFRL) under agreement number FA8650-17-2-7711. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

Cite as:
Hoffman, R.R., Mueller, S.T., Klein, G., and Litman, J. (2018). "Measuring Trust in the XAI Context." Technical Report, DARPA Explainable AI Program.

OUTLINE

1. INTRODUCTION

Trust in automation, is of concern in computer science and cognitive systems engineering, as well as the popular media (e.g., Chancey et al., 2015; Hoff and Bashir 2015; Hoffman et al., 2009; Huynh et al., 2006; Naone, 2009; Merritt and Ilgen, 2008; Merritt et al. 2013, 2015a; Pop et al. 2015; Shadbolt, 2002; Wickens et al. ,2015; Woods and Hollnagel 2006). Trust is of particular concern as more AI systems are being developed and tested (Schaefer et al. ,2016).

Scope of this Report

Any comprehensive account of the concept of trust would have to "plunder many sources; the philosophy of Socrates and Aristotle, Hobbes and Kant; the sociology of Durkheim, Weber and Putnam; literature; economics; scientific methodology; the most ancient of history and the most current of current affairs" (O'Hara, 2004, p. 5).

The literature on trust in automation cross-references to the study of interpersonal trust and trust within organizations to consider differences between social and technological contexts. and the implications of those differences for conceptual models of trust. Focusing on trust on automation, publications focus even more narrowly on trust in robots, trust in the cyber domain, trust in decision aids, and trust in AI and ML systems. As the Bibliography in this Report suggests, even this focus results in a large corpus of papers and publications.

Recommended reviews of trust concepts and conceptual models are Adams, et al. (2003) and Hoffman (2017). The purpose of this Report is to encapsulate the key findings and models, but do so with a focus on the psychometrics of trust and the selection (or creation) or a trust/reliance scale that is suited to the XAI context.

The reason for the literature review component of this Report is that the models and theories of trust in automation, and the entire trust research paradigm (primarily applied experimental psychology and human factors) have shaped the design (or selection) of measures of trust, and hence the design of scales of trust.

2. MODELS OF TRUST IN AUTOMATION

One class of models of trust in automation is those that are essentially listings of variables or factors that have or are believed to have a direct causal influence on trust (e.g., Muir 1987): cultural differences, operator predispositions, operator personality, knowledge about the automation, and so on. The goal of the models is to capture all of the variables, or at least what are believed to be the most important variables, that might have causal influence in how humans come to trust in and thereby rely upon computational technologies (e.g., Oleson et al. 2011; Rempel et al. 1985).

Researchers have discussed a number of context factors that influence trust, such as the type of technology, the complexity of the task, perceived risks, and so on (Schaefer et al. 2016). Thus, for example, under high-risk conditions, some people may reduce their reliance on complex technology but increase their reliance on simple technology (Hoff and Bashir 2015). As another

example, there are individual differences in beliefs about automation reliability and trustworthiness (Merritt et al. 2015b; Pop et al. 2015). Some individuals have an "all-or-none" belief, that automation either performs perfectly, or that it always makes errors (Wickens et al. 2015).

In their model of trust in automation, Hoff and Bashir list 29 variables that have been reported in the literature. In their model of trust in automation, Schaefer et al. (2016) list 31 factors. The listed factors that these researchers adduce are said to influence the development of trust. Furthermore, all of the variables are said to interact. To give just three examples, all just from the 2016 meeting of the Human Factors and Ergonomics Society:

- A valid recommendation from a computer is less appreciated if the operator is capable of performing the task on their own (Yang et al. 2016).
- Operator fatigue interacts with the reliability of the technology in influencing actual reliance (Wohleber et al., 2016).
- Cultural differences on such factors as individualism and power relations manifest as differing tendencies in the trust of automation (Chien et al., 201).

A second class of models of trust are process models. This includes mathematical instantiations (using such approaches as linear modeling) that are designed to predict or estimate values or levels of trust, or point-like values of automation-dependent judgments (cf. Seong and Bisantz 2002). This class also includes conceptual models that depict a process by which trust results from the causal influence of mediating variables, and in turn leads to action (i.e., reliance). Exemplifying this second class of models is the seminal and highly influential conceptual model of Lee and See (2004). It has the characteristics of both a causal (mediating variables) model and a process model; that is, it combines a causal diagram with a list. The model has this primary causal chain:

(1) Information → (2) Operator's belief → (3) Operator's trust → (4) Operator's intention → (5) Operator's action → (6) Operator's action → (7) Automation's action → (8) Display of resultants → Back to (1)

In addition, pointing into this process chain is a list of factors that are held to have a causal influence on trust. This renders the Lee and See model as somewhat oppositional: It reflects both the tendency of theorists to reduce complex cognitive processes to simple linear chains, on the one hand, and the inclination to adduce long lists of causal variables and their interactions, on the other hand.

3. THE DYNAMICS OF TRUST

Threshold effects and contingent information availability illustrate the dynamics of trust. Despite the acknowledgement of dynamics, models of trust in automation mostly regard trust as a state. Dynamics are involved only in the achievement of, or progress to, that state. Researchers generally acknowledge that trust is dynamic; that it develops. But this is usually meant in the sense that trust builds or increases over time until it reaches a stable state. This view is almost always on the assumption of a single fixed task or goal (e.g., Khasawneh et al. 2003; Muir 1987, 1994).

While it is true that trusting often changes over time, it should not be assumed trusting always develops in the sense of maturational convergence on some single state, level, or stable point. In some cases, it can appear as if trust is developing, but this is perhaps the exception and should not be elevated to the prototype. For example, trust and mistrust often develop swiftly (Meyerson et al. 1966). Some people show a bias or disposition to believe that automation is more capable and reliable than it actually is, and such high expectations result in swift mistrust when the automation makes an error (Merritt et al. 2015b; Pop et al. 2015; Wickens et al. 2015). Swift trust is when a trustor immediately trusts a trustee on the basis of authority, confession, profession, or even exigency. Naive belief in the infallibility of computers is an instance of swift (and perhaps unjustified) trust. Swift trust can be prominent early in a relationship, with contingent trust emerging over time as people experience automation in different circumstances. In other words, trust does not "develop," it morphs.

This clearly implies that the assessment of trust in XAI systems should be a repeat measure.

4. VARIETIES OF TRUST IN AUTOMATION

A number of what are believed to be different kinds of trust have been noted in the pertinent literatures. For example, Meyerson et al. (1966) popularized the notion of swift trust, originally referencing the initial trust relation that is assumed when teams are formed within organizations. As another example, Bobko et al., (2014) discussed the state of "suspicious trust" in which there emerges a feeling of mistrust, followed by an attempt to apprehend what is going on and explain perceived discrepancies. This links to the notion of the understandability of automation. Workers attempt to make sense of their technology at the same time that they are using the technology to conduct their primary tasks (Muir 1987; Woods et al. 1990). This links trust and trust measurement directly to explainability, in the XAI context

Merritt et al. (2013) demonstrated "default trust." If a person encounters a reason to distrust the automation, they are more likely to continue trusting it if they have a propensity to trust. It is safe to assert that much of the time people do not pause to deliberate about whether they trust their technology. Without compelling evidence to the contrary, and as a consequence of inertia of many kinds, the tendency is to just continue business as usual. Default trust is when an individual enters into a dependency on a machine with the expectation that the machine will do what it is intended to do, without ruminations on whether the trustworthiness of the machine hinges on the fact that the machine is a computer.

Trust in automation has been described as an attitude (positive or negative valuation of the machine by the human), as an attribution (that the machine possesses a quality called trustworthiness), as an expectation (about the machine's future behavior), as a belief (faith in the machine, its benevolence, and directability), as an intention (of the human to act in a certain way with respect to the machine), as a trait (some people are trusting, perhaps too trusting in machines), and as an emotional state (related to affective factors such as liking or familiarity). But these are not exclusive. A trusting relation can be, and usually is, some mixture of these, all

at once. This complicates any attempt to develop a robust or generally applicable method for evaluating trust.

5. TRUSTING RELATIONSHIPS IN THE XAI CONTEXT

Trust can be thought of as an abductive inference, that is, a "best hypothesis" about the trustworthiness of the AI. Since it is based on the understandability and perceived predictability of the AI, it is a defeasible inference (i.e., it is potentially fallible). Some users may take the computer's assertions (data, claims) as valid and true because they come from a computer. But other users may require some sort of justification—empirical reasons to believe that the computer's presentations or assertions are valid and true.

Absolute Trusting is when the user takes the computer's assertions (data, claims) as valid and true in all circumstances.
Contingent Trusting is when the user can take some of the computer's presentations or assertions as valid and true under certain circumstances.
Progressive Trusting is when the user takes more of the machine's presentations or assertions as valid and true over time or across experiences.
Digressive Trusting is when the user takes fewer of the machine's presentations or assertions as valid and true over time and across experiences.

Any of these trusting states can be stable or tentative, skeptical. These possibilities result in a combinatoric. For example, Stable Justified Trusting is when the user takes the computer's presentations or assertions as true most of the time, or over some time span. This can be taken as a refined definition of what is meant by the notion of trust calibration (McGuirl and Sarter 2006; Parasuraman and Riley 1997)

Were we to impose a continuum, weak positive trusting—trusting that is Contingent, Skeptical, and Tentative—is still a form of trusting. Eventually, however, skeptical trust (*I'll trust you, but I'm not so sure, and I'm on the lookout*) can and does give way to mistrust (*Sorry, but I just don't trust you anymore*). Just as there are varieties of trusting, there are varieties of negative trusting.

Mistrusting is the belief that the computer might do things that are not in the user's interest.
Distrusting is the belief that the computer may or may not do things that are in the user's interest.
Anti-trusting is the belief that the computer will do things that are not in the user's interest.
Counter-trusting is the belief that the computer must not be relied upon because the machine is presenting information that suggests it should be trusted.

Trust in automation can rapidly break down under conditions of time pressure, or when there are conspicuous system faults or errors, or when there is a high false alarm rate (Dzindolet et al., 2003; Madhavan and Wiegmann 2007). Studies of how people deal with the user-hostile aspects of software (Koopman and Hoffman 2003) reveal a variety of reasons why people create work-arounds and kludges, and why people are frustrated by their computers, even to the point of committing automation abuse (Hoffman et al., 2008). The trusting of machines can be hard to reestablish once lost.

Obviously, these are states that are to be avoided in XAI, but however trust is measured, the measurement method must be sensitive to the emergence of negative trusting states.

> *XAI systems should enable the user to know whether, when and why to trust and rely upon the XAI system and know whether, when, or why to mistrust the XAI and either not rely upon it, or rely on it with caution.*

Such mixtures of states do occur in human-computer interaction. The extreme case is counter-trusting: The user's decision to do something contrary to, perhaps even precisely the opposite of what the computer suggests. Studies of how people deal with the user-hostile aspects of software (Koopman and Hoffman 2003) reveal a variety of reasons why people create work-arounds and kludges, and why people are frustrated by their computers, even to the point of committing automation abuse (Hoffman et al. 2008).

The main take-away from this analysis is that people always have some mixture of justified and unjustified trust, and justified and unjustified mistrust in computer systems. A user might feel positive trust toward an AI system with respect to certain tasks and goals and simultaneously feel mistrusting or distrusting when other tasks and goals are engaged. Indeed, in complex sociotechnical work systems, this is undoubtedly the norm in the human-machine relation (Hoffman et al. 2014; Sarter et al. 1997),

Ideally, with experience the user comes to trust the computer with respect to certain tasks or goals in certain contexts or problem situations and appropriately mistrusts the computer with respect to certain tasks or goals in certain contexts or problem situations. Only if this trusting relation is achieved can the user's reliance on the computer be confident.

6. TRUST AND RELIANCE

In theory, the level of trust as specified by a momentary judgment is reflected in reliance. The varieties of positive and negative trusting can be associated with different reliance stances.

For instance, Progressive Trusting could be understood as entailing increasing reliance. However, reliance affects the information an operator receives regarding the performance of the automation, because the performance of the automation is only perceivable when the person is relying on the automation (Gao and Lee 2006). In other words, the relation of trust to reliance is not a unidirectional causal relation. There are contingencies between reliance and the information received subsequent to the reliance that guides the further morphing of trust.

7. LINKING TRUST TO "EXPLANATION AS EXPLORATION"

Trusting emerges from knowledge about what happens when events challenge boundary conditions (Hollnagel et al. 2006; Klein, et al., 2004; Woods 2011). To achieve robustness, adaptivity, and resilience in an XAI work system, users must develop Contingent, Justified,

Stable Trusting and Contingent, Justified, Stable Mistrusting in the XAI system. Unjustified Trusting and Unjustified Mistrusting and Distrusting among human and/or machine agents in a work system would not be conducive to adaptation or resilience (Hoffman and Hancock 2017). Given these considerations, and especially the view that trusting is a process, it can be further asserted that trusting of XAI systems is always exploratory. What users need is to achieve a momentary all-encompassing trust state. Users need to be able to actively probe their work systems. Active exploration of trusting–relying relationships cannot and should not be aimed at achieving single stable states or maintaining some decontextualized metrical value, but must be aimed at maintaining an appropriate and context-dependent expectation. Active exploration, on the part of the user, of the trustworthiness of the XAI within the competence envelope of the total work system will involve verification of reasons to take the computer's presentations or assertions as true, and an assessment of situational uncertainty that might affect the probability of favorable outcomes.

> *Trust does not "develop," it morphs. Trust is an emergent, it is not a state (Woods, 2009). Just as explanation in the XAI context can be regarded as an exploratory process, so to is it best to think of trusting as an active exploration process (Hoffman, Johnson, Bradshaw and Underbrink, 2013).*

Active exploration of human-XAI interdependence would hinge on there being a usable, useful, and understandable method built into the work that permits the systematic evaluation of and experimentation on the human–XAI relationship. The goals would include the following:

- Enabling the user to identify and mitigate Unjustified Trusting and Unjustified Mistrusting situations.
- Enabling the user to discover indicators to mitigate the impacts and risks of unwarranted reliance, or unwarranted rejection of recommendations, especially in time-pressured or information-challenged (too much, too little, or uncertain) situations.
- Enabling the user to adjust their reliance to the task and situation.
- Enabling the user to develop Justified Swift Trust and Justified Swift Mistrust in the machine. The worker needs guidance to know when to trust, or when not to trust, early and "blindly" (Roth 2009).
- Enabling the user to understand and anticipate circumstances in which the machine's recommendations will not be trustworthy, and the computer's recommendations should not be followed even though they appear trust- worthy.
- Enabling the user to understand and anticipate circumstances (i.e., unforeseen variations on contextual parameters) in which the XAI should not be trusted even if it is working as it should, and perhaps especially if it is working as it should (Woods 2011).
- Enabling the user to develop Justified Trusting over long time spans of experience with a machine in a variety of challenging situations.

With these considerations in mind, a notional view of trust dynamics can be created, as shown in Figure 1. In this particular dynamic, the user is initially cautions or skeptical but the initial explanation is a good one, and moves the user into a region of justified trust. Subsequent use of the XAI system, however, results in an automation surprise. For example, a Deep Net makes a misclassification that no human would make. This might swiftly moves the user into a state of

unjustified mistrust, in which becomes skeptical of any of the classifications that the Deep Net makes. Following that, the XAI system provides additional explanations and the user explores the performance of the XAI, converging in the region of appropriate trust and reliance.

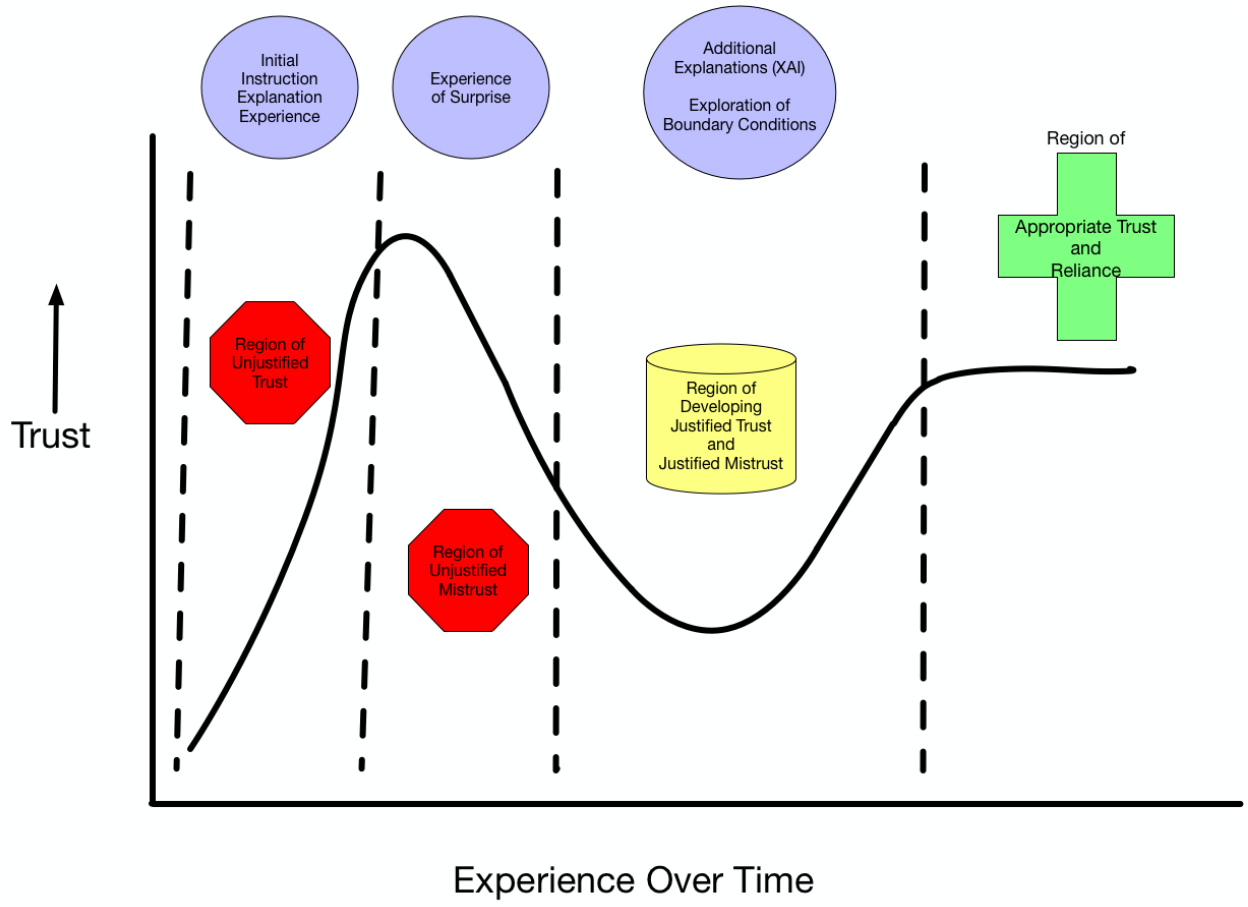

Figure 1. A notional view of how trust could morph in the XAI context.

8. TRUST MEASUREMENT

We have reviewed publications on a number of scales for measuring trust in automation, including a number of articles that themselves offer summary reviews of other scales (see the Bibliography). The majority of trust scales have been developed for application in the context of interpersonal trust. We focus on scales designed for use in the assessment of human trust in automation. A synopsis of representative trust scales is presented in Appendix A.

Minimally, a trust scale can ask two questions: *Do you trust the machine's outputs?* (trust) and *Would you follow the machine's advice?* (reliance). Indeed, these two items comprise the scale developed by Adams, et al. (2003). The scale developed by Johnson (2007) asks only about reliance and the rareness of errors.

Some scales for assessing trust in automation are highly specific to particular application contexts. For example, the scale developed by Schaefer (2013) refers specifically to the context of human reliance on a robot, and thus asks *Does it act as part of a team?* and *Is it friendly?* As another example, the trust in automation scale developed by Adams, et al. (2003) refers specifically to the evaluation of simulations. It asks only two questions, one about trust (*Do you trust it?*) and one about reliance (*Are you prepared to rely on it?*), although it is noteworthy that this is the only scale that has a free response option associated with the two scale item questions.

To be sure, some research on trust in automation is decidedly not applicable to the XAI context. For example, Heerink, et al. (2010) were interested in the acceptance of an assistive/robotic technology by the elderly. The questionnaire they utilized has such items as *I feel the robot is nice,* and *The robot seems to have real feelings*.

Montague (2010) presented a study aimed at validating a scale for trust in medical diagnostic instruments, but all of the actual items refer to trust in the health care provider and positive affect about the provider. Abstracting from that reference context, the other items ask about reliability, correctness, precision, and trust. Thus, we see essential similarity to the items in the Cahour-Fourzy Scale.

Some scales for assessing trust in automation are highly specific to particular experimental contexts. Hence, the items are not applicable to the XAI context, or to any generic trust-in-automation context. For example, the scale by Dzindolet, et al. (2003) was created for application in the study of trust in a system for evaluating terrain in aerial photographs, showing images in which there might be camouflaged soldiers. Thus, the hypothetical technology was referred to as a "contrast detector." The experiment was one in which the error rate of the hypothetical detector was a primary independent variable. As a consequence, the scale items refer to trials e.g., *How well do you think you will perform during the 200 trials? (Not very well-Very well),* and *How many errors do you think you will make during the 200 trials?* Some of the scale items can be adapted to make them appropriate to the XAI context, but the result of this modification is just a few items, which are ones that are in the Cahour-Fourzy Scale items (e.g., *Can you trust the decisions the [system] will make?)*

Of those scales that have been subject to reliability analysis, results suggest that trust in automation scales can be reliable. (For details, see Appendix A.) Of those scales that have been subject to validity analysis, high Chronbach alpha results have been obtained. The report by Jian, et al., (2000) illustrates these psychometric analyses.

Recommendations For Application of the XAI Trust Scale

(1) Recommended Scale

Looking across the various Scales (see Appendix A), there is considerable overlap, and cross-use of the scale items. We have distilled a set of items that might be used in XAI research. This is presented in Appendix B.

Most of the items are from the Cahour-Fourzy scale (some of which are also in the Jian et al. scale), but the Recommended Scale incorporates items from other scales.

(2) Recommendation for Study Design

None of the scales that have been reviewed treat trust as a process (see Appendix A); they treat it as a static quality that is measured after the research participant has completed their experimental tasks.

In contrast, it is recommended for the XAI Program that trust measurement be a repeat measure. The scale or selected scale items can be applied after individual trials (e.g., after individual XAI categorizations or recommendations; after individual explanations are provided, etc.).

The full scale could be completed part way through a series of experimental trials, and at the conclusion of the final experimental trial.

Multiple measures taken over time could be integrated for overall evaluations of human–machine performance, but episodic measures would be valuable in tracking such things as: How do users maintain trust? What is the trend for desirable movement toward appropriate trust?

Bibliography

Adams, B.D., Bruyn, L.E., Houde, S., Angelopoulos, P., Iwasa-Madge, K., & McCann, C. (2003). Trust in automated systems. Report, *Ministry of National Defence*, United Kingdom.

Atkinson, D.J., Clancey, W.J., and Clark, M.H. (2014). Shared awareness, autonomy and trust in human-robot teamwork. In *Artificial Intelligence and Human-Robot Interaction: Papers from the 2014 AAAI Fall Symposium* (pp. 36–38). Menlo Park, CA: AAAI>

Ballas, J.A. (2007). Human centered computing for tactical weather forecasting: An example of the "Moving Target Rule.". In R.R. Hoffman (Ed.), *Expertise out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making* (pp. 317– 326). Mahwah, NJ: Lawrence Erlbaum.

Barber, B. (1983). *The Logic and Limits of Trust*. New Brunswick, NJ: Rutgers University Press.

Beck, H.P., Dzindolet, M.T., and Pierce, L.G. (2002). Operators' automation usage decisions and the sources of misuse and disuse. Advances in Human Performance and Cognitive Engineering Research, 2, 37–78.

Bisantz, A.M., and Seong, Y. (2001). Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics, 28*, 85–97.

Bobko, P., Barelka, A.J., and Hirshfield, L.M. (2014). The construct of state-level suspicion: A model and research agenda for automated information technology (IT) contexts. *Human Factors, 56,* 498–508.

Bradshaw, J.M., Jung, H., Kulkarni, S., Johnson, M., Feltovich, P., Allen, J., Bunch, L., Chambers, N., Galescu, L., Jeffers, R., Suri, N., Taysom, W., and Uszok, A. (2005). Toward trustworthy adjustable autonomy in KAoS. In R. Falcone (Ed.), *Trusting Agents for Trustworthy Electronic Societies* (pp. 18–42). Lecture Notes in Artificial Intelligence. Berlin: Springer.

Cahour, B., and Forzy, J. F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science, 47,* 1260-1270.

Chancey, E.T., Bliss, J.P., Proaps, A.B., and Madhavan, P. (2015). The role of trust as a mediator between system characteristics and response behaviors. *Human Factors, 57,* 947–958.

Chien, S.-Y., Sycara, K., Liu, J.-S., and Kumru, A. (2016). Relation between trust attitudes toward automation, Hofstede's Cultural dimensions, and Big Five personality traits. In *Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting* (pp. 840–845). Santa Monica, CA: Human Factors and Ergonomics Society.

Choiu, E.R., and Lee, J.D. (2015. Beyond reliance and compliance: Human-automation coordination and cooperation. In *Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting* (pp. 195–200). Santa Monica, CA: Human Factors and Ergonomics Society.

Corritore, C.L., Kracher, B., and Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human–Computer Studies, 58,* 737–758.

Cramer, H., Evers, V., Kemper, N., and Wielinga, B. (2008). Effects of autonomy, traffic conditions and driver personality traits on attitudes and trust towards in-vehicle agents. In *Proceedings of the IEEE/WIC/ACM Int. Conference on Web Intelligence and Intelligent Agent Technology, 3*, 477–482.

Cramer, H., Goddijn, J., Wielinga, B. and Evers, V. (2010). Effects of (in)accurate empathy and situational valence on attitudes towards robots. In *Proceedings of the 5th ACM/IEEE International Conference on Human–Robot Interaction* (pp. 141–142). New York: Association for Computing Machinery.

Crispen, P., and Hoffman, R.R. (2016, November/December). How many experts? *IEEE: Intelligent Systems*, 57–62.

Desai, M., Stubbs, K., Steinfeld, A., and Yanco, H. (2009). Creating trustworthy robots: Lessons and inspirations form automated systems. In *Proceedings of the Artificial Intelligence and Simulation of Behavior Convention: New Frontiers in Human–Robot Interaction*. Edinburgh, Scotland: University of Edinburgh.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., and Beck, H.P. (2003). The role of trust in automation reliance. *International Journal of Human–Computer Studies, 58*, 697–718.

Fitzhugh, E.W., Hoffman, R.R., and Miller, J.E. (2011). Active trust management. In N. Stanton (Ed.) *Trust in Military Teams* (pp. 197–218). London: Ashgate.

Gao, J. and Lee, J.D. (2006). Extending the decision field theory to model operators' reliance on automation in supervisory control situations. *IEEE Systems, Man, and Cybernetics, 36*, 943–959.

Hancock, P.A., Billings, D.R., and Schaeffer, K.E. (2011, July). Can you trust your robot? *Ergonomics in Design,* 24–29.

Heerink, M., Kröse, B., Evers, V., and Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: The Almere model. *International journal of social robotics*, *2*(4), 361-375.

Hoff, K.A., and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*, 407–434.

Hoffman, R.R. (2017). A Taxonomy of Emergent Trusting in the Human–Machine Relationship. In P. J. Smith and R.R. Hoffman (Eds.), *Cognitive systems engineering: The future for a changing world* (pp. 137-163). Boca Raton, FL: Taylor and Francis.

Hoffman, R.R. (1989). Whom (or what) do you trust: Historical reflections on the psychology and sociology of information technology. In *Proceedings of the Fourth Annual Symposium on Human Interaction with Complex Systems* (pp. 28–36). New York: IEEE Computer Society.

Hoffman, R.R., Coffey, J.W., Ford, K.M. and Carnot, M.J. (2001, October) STORM-LK: A human-centered knowledge model for weather forecasting. In J.M. Flach (Ed.), *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society* (p. 752). Santa Monica, CA: Human Factors and Ergonomics Society.

Hoffman, R.R., and Hancock, P.A. (2017). Measuring resilience. *Human Factors*, 59, 564-581.

Hoffman, R.R., Hawley, J.K., and Bradshaw, J.M. (2014, March/April). Myths of automation. Part 2: Some very human consequences. *IEEE Intelligent Systems*, 82–85.

Hoffman, R.R., Johnson, M., Bradshaw, J.M., and Underbrink, A. (2013, January/February). Trust in automation. *IEEE: Intelligent Systems,* 84–88.

Hoffman, R.R., Lee, J.D., Woods, D.D., Shadbolt, N., Miller, J. and Bradshaw, J.M. (2009, November/December). The dynamics of trust in cyberdomains. *IEEE Intelligent Systems*, 5–11.

Hollnagel, E., Woods, D.D. and Leveson, N. (Eds.) (2006). *Resilience Engineering: Concepts and Precepts*. Aldershot, UK: Ashgate.

Huynh, T.D., Jennings, N.R., and Shadbolt, N.R. (2006). An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems, 13,* 119–154.

Jian, J. Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53-71.

Johnson, D.S. (2007). Achieving customer value from electronic channels through identity commitment, calculative commitment, and trust in technology. *Journal of interactive marketing*, *21*(4), 2-22.

Johnson, M., Bradshaw, J.M., Hoffman, R.R., Feltovich, P.J., and Woods, D.D. (November/ December 2014). Seven cardinal virtues of human–machine teamwork. *IEEE Intelligent Systems*, 74–79

Khasawneh, M.T., Bowling, S.R., Jiang, X., Gramopadhye, A.K. and Melloy, B.J. (2003). A model for predicting human trust in automated system. In *Proceedings of the 8th Annual International Conference on Industrial Engineering—Theory, Applications and Practice* (pp. 216–222). Sponsored by the International Journal of Industrial Engineering [http://ijietap.org].

Klein, G., Woods, D.D., Bradshaw, J.D., Hoffman, R.R. and Feltovich, P.J. (November/ December 2004). Ten challenges for making automation a "team player" in joint human- agent activity. *IEEE: Intelligent Systems*, 91–95.

Koopman, P., and Hoffman, R.R. (November/December 2003). Work-arounds, make-work, and kludges. *IEEE: Intelligent Systems*, 70–75.

Kucala, D. (2013, June). The truthiness of trustworthiness. Chief Learning Officer, 57–59.

Lee, J.D. and Moray, N. (1992). Trust, control strategies, and allocation of functions in human–machine systems. *Ergonomics, 35*, 1243–1270.

Lee, J.D., and See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50–80.

Lewandowsky, S., Mundy, M., and Tan, G. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology-Applied, 6*, 104–123.

Lewicki, R.J., McAlister, D.J., and Bias, R.J. (1998). Trust and distrust: New relationships and realities. *Academy of Management Review, 23*, 438–445

Madhavan, P., and Wiegmann, D.A. (2007). Effects of information source, pedigree, and reliabil- ity on operator interaction with decision support systems. *Human Factors, 49*, 773–785.

Mayer, R.C., Davis, J.H., and Schoorman, F.D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*, 709–734.

McGuirl, J.M., and Sarter, N. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors, 48*, 656–665.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(4), 356-370.

Merritt, S.M., Heimbaugh, H., LaChapell, J., and Lee, D. (2013). I trust it, but don't know why: Effects of implicit attitudes toward automation in trust in an automated system. *Human Factors, 55,* 520–534.

Merritt, S.M., and Ilgen, D.R. (2008). Not all trust is created equal: Dispositional and history-based trust in human–automation interactions. *Human Factors, 50*, 194–201.

Merritt, S.M., Lee, D., Unnerstall, J.L., and Huber, K. (2015a). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors, 57*, 34–47.

Merritt, S.M., Unnerstall, J.L., Lee, D., and Huber, K. (2015b). Measuring individual differences in the perfect automation schema. *Human Factors, 57*, 740–753.

Meyerson, D., Weick, K., and Kramer, R. (1966). Swift trust and temporary groups. In T.R. Tyler and R. Kramer (Eds.), *Trust in Organizations: Frontiers of Theory and Research* (pp. 166–195). Thousand Oaks, CA: Sage.

Montague, E. (2010). Validation of a trust in medical technology instrument. *Applied ergonomics*, *41*(6), 812-821.

Mueller, S.T. and Klein, G. (2011, March/April). Improving users' mental models of intelligent software tools. *IEEE Intelligent Systems*, 77–83.

Muir, B.M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man–Machine Studies, 27*, 527–539.

Muir, B.M. (1994). Trust in automation Part 1: Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*, 1905–1922.

Muir, B.M. and Moray, N. (1996). Trust in automation. Part II Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39,* 429–460.

Naone, E. (4 September 2009). Adding trust to Wikipedia, and beyond. *Technology Review*

[http://www.technologyreview.com/web/23355/?a=f].

Nass, C., and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*, 81–103.

O'Hara, K. (2004) *Trust: From Socrates to Spin*. Cambridge, UK: Icon Books.

Oleson, K.E., Billings, D.R., Chen, J.Y.C, and Hancock, P.A. (2011). Antecedents of trust in human–robot collaborations. In *Proceedings of the IEEE International Multi- Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support* (pp. 175–178). New York: Institute of Electrical and Electronics Engineers.

Parasuraman, R., and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*, 230–253.

Parlangeli, O., Chiantini, T., and Guidi, S. (2012). A mind in a disk: The attribution of mental states to technological systems. *Work, 41*, 1118–1123.

Pop, V.L., Shrewsbury, A., and Durso, F.T. (2015). Individual differences in the calibration of trust in automation. *Human Factors, 57,* 545–556.

Pritchett, A.R., and Bisantz, A.M. (2002). Measuring judgment interaction with displays and automation. In *Proceedings of the human Factors and Ergonomics Society 46th Annual Meeting* (pp. 512–516). Santa Monica, CA: Human Factors and Ergonomics Society.

Rempel, J.K., Holmes, J.G., and Zanna, M.P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*, 95–112.

Riley, V. (1996). Operator reliance on automation: Theory and data. In R. Parasuraman and M. Mouloua (Eds.), *Automation Theory and Applications* (pp. 19–35). Mahwah, NJ: Erlbaum.

Roth, E.M. (2009). Facilitating 'calibrated' trust in technology of dynamically changing 'trust-worthiness'. Presentation at the Working Meeting on Trust in Cyberdomains. Institute for Human and Machine Cognition, Pensacola, FL. Supported by the Human Effectiveness Directorate, Air Force Research Laboratory, Wright-Patterson AFB, OH.

Sarter, N., Woods, D.D., and Billings, C.E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of Human Factors/Ergonomics*, 2nd ed. (pp. 1926–1943). New York, NY: Wiley.

Schaefer, K. E. (2013). *The perception and measurement of human-robot trust*. Doctoral dissertation, University of Central Florida Orlando, Florida.

Schaefer, K.E., Chen, J.Y.C., Szalma, J.L., and Hancock, P.A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for under- standing autonomy in future systems. *Human Factors, 58,* 377–400.

Seong, Y., and Bisantz, A. (2002. Judgment and trust in conjunction with automated aids. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 423–428). Santa Monica, CA: Human Factors and Ergonomics Society.

Seong, Y., and Bisantz, A.M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics, 38,* 608–625.

Shadbolt, N. (January/February, 2002). A matter of trust. IEEE Intelligent Systems, 2–3. Sheridan, T. (1980). *Computer control and human alienation. Technology Review, 83*, 61–73.

Singh, I. L., Molloy, R., and Parasuraman, R. (1993). Automation-induced" complacency": Development of the complacency-potential rating scale. *The International Journal of Aviation Psychology*, *3*(2), 111-122.

Skitka, L.J., Mosier, K.L., and Burdick, M. (2000). Accountability and automation bias. International Journal of Human–Computer Studies, 52, 701–717.

Stokes, C., Lyons, J., Littlejohn, K., Natarian, J., Case, E., and Speranza, N. (2010). Accounting for the human in cyberspace: Effects of mood on trust in automation. In Proceedings of the 2010 International Symposium on Collaborative Technologies and Systems (pp. 180–187). New York: Institute for Electrical and Electronics Engineers.

Vasalou, A., Hopfensitz, A., and Pitt, J. (2008). In praise of forgiveness: Ways for repairing trust breakdowns in one-off online interactions. *International Journal of Human– Computer Studies, 66*, 466–480.

Wagner, A. 2009. The Role of Trust and Relationships in Human–Robot Social Interaction. Doctoral Dissertation, Georgia Institute of Technology, Atlanta, GA.

Wang, L., Jamieson, G.A., & Hollands, J.G. (2009). Trust and reliance on an automated combat identification system. *Human factors*, *51*(3), 281-291.

Watts-Perotti, J. and Woods, D.D. (2009). Cooperative advocacy: A strategy for integrating diverse perspectives in anomaly response. *The Journal of Collaborative Computing, 18,* 175–198.

Wickens, C.D., Clegg, B.A., Vieane, A.Z., and Sebok, A.L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors, 57*, 728–739.

Wohleber, R.W., Calhoun, G.L., Funke, G.J., Ruff, H., Chiu, C.-Y.P., Lin, C., and Matthews, G. (2016). The impact of automation reliability and operator fatigue on performance and reliance. In *Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting* (pp. 211–216). Santa Monica, CA: Human Factors and Ergonomics Society.

Woods, D.D. (2009). Trust Emerges from the Dynamics of Reciprocity, Responsibility and Resilience in Networked Systems. Presentation at the Working Meeting on Trust in Cyberdomains. Institute for human and machine Cognition, Pensacola, FL. Supported by the human Effectiveness Directorate, Air Force Research Laboratory, Wright-Patterson AFB, OH.

Woods, D.D. (2011, September). Reflections on 30 years of picking up the pieces after explosions of technology. Presentation at the AFRL Autonomy Workshop, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH.

Woods, D.D., and Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. Boca Raton, FL: CRC Press.

Woods, D.D., Roth, E.M., and Bennett, K. (1990). Explorations in joint human–machine cognitive systems. In W. Zachary and S. Robinson (Eds.), *Cognition, Computing, and Cooperation* (pp. 123–158). Norwood, NJ: Ablex.

Yang, X.J., Wickens, C.D., and Hölttä-Otto, K. (2016). How users adjust trust on automation: Contrast effect and hindsight bias. In *Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting* (pp. 196–200). Santa Monica, CA: Human Factors and Ergonomics Society.

APPENDIX A

Synopsis of Representative Trust Scales

Adams, et al., Scale (2003)

Adams, et al. actually developed two scales. One was for the evaluation of simulations but the other was generic, intended for the evaluation of any form of automation. Apart from the item about liking, The items show overlap with items in the Cahour-Fourzy Scale.

Each item is accompanied by a bipolar rating scale (e.g., Useful-Not Useful; Reliable-Not reliable) on which the participant makes a tick mark on a -5 to +5 delineation. Following the Likert items, the Scale asks participants to rank the importance of the six item factors.

Is the automation tool useful?
How reliable is it?
How accurately does it work?
Can you understand how it works?
Do you like using it?
How easy is it to use?

Cahour-Forzy (2009) Scale

Trust (and distrust) are defined as a sentiment resulting from knowledge, beliefs, emotions and other aspects of experience, generating positive or negative expectations concerning the reactions of a system and the interaction with it. The scale was developed in the context of learning to use a cruise control system. Trust was analyzed into three factors: reliability, predictability, and efficiency. The scale asks users directly whether they are confident in the XAI system, whether the XAI system is predictable, reliable, safe, and efficient.

The scale assumes that the participant has had considerable experience using the XAI system. Hence, these questions would be appropriate for scaling after a period of use, rather than immediately after an explanation has been given and prior to use experience. In the original scale, the items are rated on a bipolar scale going from "I agree completely" to "I do not agree at all." The items we present below have been slightly modified to fit the general Likert form developed for the XAI Explanation Satisfaction Scale. In addition to conforming to psychometric standards, consistency of format will presumably make the ratings tasks easier for participants.

1. What is your confidence in the [tool]? Do you have a feeling of trust in it?

1	2	3	4	5	6	7
I do not trust it at all.						I trust it completely

2. Are the actions of the [tool] predictable?

1	2	3	4	5	6	7
It is not at all predictable.						It is completely predictable.

3. Is the [tool] reliable? Do you think it is safe?

1	2	3	4	5	6	7
It is not at all safe.						It is completely safe.

4. Is the [tool] efficient at what it does?

1	2	3	4	5	6	7
It is not at all efficient.						It is completely efficient.

Jian, et al. Scale (2000)

Trust is regarded as a trait. It is analyzed into six factors: Fidelity, loyalty, reliability, security, integrity, and familiarity. Factors were developed from cluster analysis on trust-related words. This scale is one of the most widely used, especially in the field of human factors. Indeed, a number of other scales have used items, or have adapted scale items, from the Jian, et al. Scale.

The item referencing "integrity" is problematic as the concept that a machine can act with integrity is not explicated. The final item, about familiarity, would not be relevant in the SAI context, since the participants' degree of experience with the XAI system will be known objectively.

Items 1, 2, 3, and 4 all seem to be asking the same thing.

The other items items in this scale show considerable overlap with items in the Cahour-Fourzy scale. However, item 4 is particularly interesting and is does not have a counterpart in the Cahour-Fourzy Scale. We are inclined to recommend that the Jian, et al., item 4 be incorporated into the XAI version of the Cahour-Fourzy Scale.

1. The system is deceptive.
2. The system behaves in an underhanded manner.
3. I am suspicious of the system's intent, action, or outputs.
4. I am wary of the system.
5. The system's actions will have a harmful or injurious outcome.
6. I am confident in the system.
7. The system provides security.
8. The system has integrity.
9. The system is dependable.
10. I can trust the system.
11. I am familiar with the system.

Madsen-Gregor Scale (2000)

Trust is defined as being both affective and cognitive. Trust was analyzed into five factors: reliability, technical competence, understandability, faith, and personal attachment. Their focus was not just trust in a decision aid but trust in an intelligent decision aid. As such, their scale deserves our particular attention. Unfortunately, reports on their work are not accompanied by information about the precise method for administering the scale (i.e., whether or not it used a Likert method). That said, their results show very high reliabilities (alpha = 0.94) and a factor analysis that accounts for about 70% of the variance.

Perceived Reliability	The system always provides the advice I require to make my decision.
	The system performs reliably.
	The system responds the same way under the same conditions at different times.
	I can rely on the system to function properly.
	The system analyzes problems consistently.
Perceived Technical Competence	The system uses appropriate methods to reach decisions.
	The system has sound knowledge about this type of problem built into it.
	The advice the system produces is as good as that which a highly competent person could produce.
	The system correctly uses the information I enter.
	The system makes use of all the knowledge and information available to it to produce its solution to the problem.

Perceived Understandability	I know what will happen the next time I use the system because I understand how it behaves.
	I understand how the system will assist me with decisions I have to make.
	Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.
	It is easy to follow what the system does.
	I recognize what I should do to get the advice I need from the system the next time I use it.
Faith	I believe advice from the system even when I don't know for certain that it is correct.
	When I am uncertain about a decision I believe the system rather than myself.
	If I am not sure about a decision, I have faith that the system will provide the best solution.
	When the system gives unusual advice I am confident that the advice is correct.
	Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will.
Personal Attachment	I would feel a sense of loss if the system was unavailable and I could not longer use it.
	I feel a sense of attachment to using the system.
	I find the system suitable to my style of decision making.
	I like using the system for decision making.
	I have a personal preference for making decisions with the system.

It is noteworthy that the Scale refers to understandability but does not explicitly reference trust.

Upon close examination, it seems that the reliability factor has some redundant items. The factors titled "perceived technical competence" and "perceived understandability" might be interpreted as referencing the user's mental model of the system. For example, the item *Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will* clearly is asking about the user's mental model. Indeed, the Madsen-Gregor Scale as a whole can be understood as referring as much to evaluating the user's mental model as it does to trust. The mere fact that this distinction is fuzzy is a testament to the notion that XAI evaluation must have measures of both trust and of mental models, since the two are causally related.

One can question the appropriateness of referring to a "faith" factor. Items in this factor seem to refer to reliance and uncertainty. One can question the appropriateness of referring to a "personal attachment" factor rather than a "liking" factor.

As with other Scales, multiple interpretations are possible. For instance, the Madsen-Gregor item *I believe advice from the system even when I don't know for certain that it is correct* asks essentially the same thing as the Cahour-Fourzy item *I am confident in the tool; it works well*.

A number of individual items are of interest, such as "It is easy to follow what the system does" and "I recognize what I should do to get the advice I need from the system." These seem to reference usability. Up to this point, issues of XAI system learnability and usability have not been considered in the XAI Program.

Merritt Scale (2011)

Trust is regarded as an emotional, attitudinal judgement of the degree to which the user can rely on the automated system to achieve his or her goals under conditions of uncertainty. Trust was initially broken into three factors: belief, confidence, and dependability. Factor Analysis revealed two other factors: propensity to trust and liking. The scale was evaluated in an experiment in which participants conducted a baggage screen task using a fictitious automated weapon detector in a luggage screening task. Chronbach's alpha ranged from a = .87 to a = .92.

Items in this Scale are all similar to items in the Cahour-Fourzy Scale.

1. I believe the system is a competent performer.
2. I trust the system.
3. I have confidence in the advice given by the system.
4. I can depend on the system.
5. I can rely on the system to behave in consistent ways.
6. I can rely on the system to do its best every time I take its advice.

Schaefer Scale (2013)

This scale was developed in the context of human-robot collaboration. Thus, trust was said to depend on both machine performance and team collaboration. Trust was analyzed into two factors: ability and performance. This scale is unique in that it is long and has a format different from all the other scales. Specifically the participant is asked to estimate the amount of time that the machine (in the study, a robot) would show each of a number of possible behaviors. In this scale format, some items are troublesome. For example, if the machine acts consistently, what is the point of asking about the percentage of time that it asks consistently? Many of the items anthropomorphize the machine (robot) and do so in ways that seem inappropriate for the XAI application (e.g., "know the difference between friend and foe," "be supportive," "be responsible," "be conscious"). For example, the point of XAI is to communicate richly and meaningfully with the participant. Thus, asking about the percentage of time that the XAI "openly communicates" or "clearly communicates" seems redundant to the evaluation of Explanation Satisfaction. In the list below, we place in the left those items that seem appropriate to XAI and in the right those that do not. The items in the left align fairly well to items in the Cahour-Fourzy Scale. One of these items—"Perform a task better than a novice human user"—is particularly interesting and might be added into the Cahour-Fourzy Scale.

What percentage of the time will this machine (robot)…

Act consistently	Protect people
Function successfully	Act as part of the team
Have errors	Malfunction
Perform a task better than a novice human user	Clearly communicate
Possess adequate decision-making capability	Require frequent maintenance
Perform exactly as instructed	Openly communicate
Make sensible decisions	Know the difference between friend and foe
Tell the truth	Provide feedback
Perform many functions at one time	Warn people of potential risks in the environment
Follow directions	
Incompetent	Meet the needs of the mission
Dependable	Provide appropriate information
Reliable	Communicate with people
Predictable	Work best with a team
	Keep classified information secure
	Work in close proximity with people
	Considered part of the team
	Friendly
	Pleasant
	Unresponsive
	Autonomous
	Conscious
	Lifelike
	A good teammate
	Led astray by unexpected changes in the environment

Singh, et al., scale (1993)

This scale presupposes a context in which the participant is evaluating a device with which they have prior experience or have general familiarity with (ATMs, medical devices, etc.). Trust was defined as an attitude toward commonly encountered automated devices that reflect a potential for complacency. Trust was analyzed into five factors: confidence, reliance, trust, safety, complacency. Since the scale merges trust and reliance, it presupposes prior experience and would not be appropriate for use when a user is first learning to use an XAI. For these reasons, we feel that this scale is not appropriate for use in the XAI context. Items that might be modified

to make them appropriate reference factors that are already covered in the Cahour-Fourzy Scale (i.e., trust, reliance).

Factor 1: Confidence
1. I think that automated devices used in medicine, such as CT scans and ultrasound, provide very reliable medical diagnosis.
2. Automated devices in medicine save time and money in the diagnosis and treatment of disease.
3. If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because it is more reliable and safer than manual surgery.
4. Automated systems used in modern aircraft, such as the automatic landing system, have made air journeys safer.
Factor 2: Reliance
1. ATMs provide a safeguard against the inappropriate use of an individual's bank account by dishonest people.
2. Automated devices used in aviation and banking have made work easier for both employees and customers.
3. Even though the automatic cruise control in my car is set at a speed below the speed limit, I worry when I pass a police radar speed trap in case the automatic control is not working properly.
Factor 3: Trust
1. Manually sorting through card catalogues is more reliable than computer-aided searches for finding items in a library.
2. I would rather purchase an item using a computer than have to deal with a sales representative on the phone because my order is more likely to be correct using the computer.
3. Bank transactions have become safer with the introduction of computer technology for the transfer of funds.
Factor 4: Safety
1. I feel safer depositing my money at an ATM than with a human teller.
2. I have to tape an important TV program for a class assignment. To ensure that the correct program is recorded, I would use the automatic programming facility on my VCR rather than manual taping.

Wang, et al. Scale (2009)

This scale was used to evaluate trust in a hypothetical "combat identification system" that participants used in a simulated task. All of the items were taken form or adapted from the Jian, et al. Scale. The reliability of the decisions generated by the hypothetical decision aid was a primary independent variable, in an effort to study response bias inducted by automation reliability. The scale items are reported in the paper, but not the format for the scale (e.g., was it

a Likert scale?). Some items are context specific (e.g., *The aid provides security; The blue light indicates soldiers*"). What is noteworthy about some of the items is that they refer explicitly to deception and mistrust. Other items in the Wang, et al., Scale refer to trust and reliability and are covered by items in the Cahour-Fourzy Scale.

The aid is deceptive.
The aid behaves in an underhanded (concealed) manner.
I am suspicious of the aid's outputs.
I am wary of the aid.
The aid's action will have a harmful or injurious outcome.
I am confident in the aid.
The aid provides security.
The aid is dependable.
The aid is reliable.
I can trust the aid.
I am familiar with the aid.
I can trust that *blue* lights indicate soldiers.
I can trust that *red* lights indicate terrorists

APPENDIX B

Recommended Scale for XAI

This Recommended Scale asks users directly whether they are confident in the XAI system, whether the XAI system is predictable, reliable, efficient, and believable.

The scale assumes that the participant has had considerable experience using the XAI system. Hence, these questions would be appropriate for scaling after a period of use, rather than immediately after an explanation has been given and prior to use experience.

A majority of the items are adapted from the Cahour-Fourzy Scale (2009), just as they have been adapted for use in other scales (e.g., Jian, et al.). In the original scale, the items are rated on a bipolar scale going from *I agree completely* to *I do not agree at all*. We have modified the items to fit the general Likert form developed for the XAI Explanation Satisfaction Scale. In addition to conforming to psychometric standards, this consistency of format will presumably make the ratings tasks easier for participants.

Item 6 was adapted from the Jian, et al. Scale, item 7 was adapted from the Schaefer Scale, and item 8 was adapted from the Madsen-Gregor Scale.

We can assume that the Recommended Scale is reliable based on these two facts:
(1). The majority of the items in this Recommended Scale essentially overlap with items in the Jian, et al. (2000) scale, which was shown empirically to be highly reliable.
(2) Items in the Recommended Scale bear overall semantic similarity to items in the Madsen-Gregor-Scale, and that scale too was also shown to have high reliability coefficients.

We can assume that the Recommended Scale has content validity Given the essential overlap of items in the Recommended Scale with items in most of the existing scales, we can safely assume that the Recommended Scale has content validity.

1. I am confident in the [tool]. I feel that it works well.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. The outputs of the [tool] are very predictable.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. The tool is very reliable. I can count on it to be correct all the time.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. I feel safe that when I rely on the [tool] I will get the right answers.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. The [tool] is efficient in that it works very quickly.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. I am wary of the [tool]. (adopted from the Jian, et al. Scale and the Wang, et al. Scale)

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. The [tool] can perform the task better than a novice human user. (adopted from the Schaefer Scale)

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. I like using the system for decision making.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly