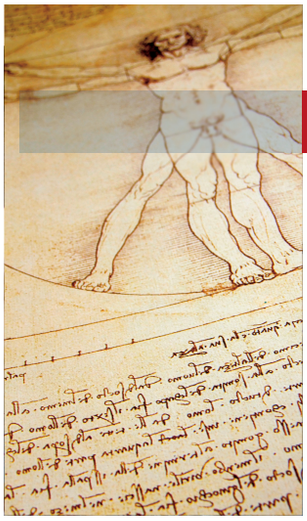

Editors: **Robert R. Hoffman, Jeffrey M. Bradshaw, and Kenneth M. Ford,**
Florida Institute for Human and Machine Cognition; rhoffman@ihmc.us

Trust in Automation

Robert R. Hoffman, Matthew Johnson, and Jeffrey M. Bradshaw, *Florida Institute for Human and Machine Cognition*
Al Underbrink, *Sentar*

Historically, technology assessments such as the Unmanned Systems Roadmap[1] from the US Department of Defense (DoD) have asserted that greater autonomy is the pathway to more effective military systems. However, this path hasn't been as straightforward as hoped—indeed, some are seriously questioning traditional views about the nature of autonomous systems and how they function in relation to humans and human environments. For example, a US Army Research Laboratory analysis of fratricide incidents involving the Patriot Missile system concluded that complex technologies increase the need for operator expertise, rather than reduce it.[2] In a significant step that reversed years of precedent in federally funded autonomy research, the 2012 Defense Science Board report, *The Role of Autonomy in DoD Systems*,[3] recommended that military technology procurement programs abandon the focus on supervisory control and "levels of autonomy" and develop a reference framework that emphasizes the importance of human-computer collaboration. An important element in developing such a framework is a better understanding of trust in automation.

Concern with issues of trust in automation emerged in parallel with the inroads computers made in such areas as supervisory control and industrial robotics. In controlling complex processes, such as nuclear power, it has been widely assumed that "human operators are not to be trusted."[4] On the other hand, concern about trust in automation is also understandable, given that deployed technologies are limited in their understandability. Trust issues challenge macrocognitive work at numerous levels, ranging from decision making at the policy level, to capability at the mission and organizational levels, to confidence at the level of cognitive work, to reliance on technology on the part of individual operators.

A previous installment of this department discussed the notion of trust in automation, focusing on technology in cyberdomains.[5] This installment broadens the subject matter to macrocognitive work systems in general. The analysis we present could contribute to an ontology that might suggest measures and techniques for what we call "active exploration for trusting" (AET). Our goal is to contribute to the development of a methodology for designing and analyzing collaborative human-centered work systems, a methodology that might promote both trust "calibration" and appropriate reliance.[5,6]

Trust is a complex and nebulous concept, so analysis to some degree of detail is necessary for us to avoid reductive thinking. Trust is closely related to myriad other concepts, and we run the risk of getting fooled by the fact that we happen to have this word "trust" in our languages. We could also get lost in philosophical hornets' nests as we attempt to forge an analytical framework. We can understand trust in automation to some extent by analogy to interpersonal trust,[7] but this analogy can also be misleading.

Interpersonal Trust vs. Trust in Machines

Interpersonal trust has been defined as a trustor's willingness to be vulnerable to a trustee's actions based on the expectation that the trustee will perform a particular action important to the trustor.[7] Research has shown that interpersonal trust depends on several factors, including perceived competence, benevolence (or malevolence), understandability, and directability—the degree to which the trustor can rapidly assert control or influence when something goes wrong.[8,9] Any one of these factors, or dimensions, could be more or less important in a given situation.

Research has demonstrated that such factors do pertain to trust in automation. However, trust

1541-1672/13/$31.00 © 2013 IEEE
Published by the IEEE Computer Society

IS-28-01-Hcc.indd 84 1/14/13 3:52 PM

in automation involves other factors that relate specifically to technology's limitations and foibles. These factors include reliability, validity, utility, robustness, and false-alarm rate.[4,10–13] Complicating the matter further is the emergence of *cognitive agents*, which some claim are neither machine nor human.[14]

The time frame over which people can gain or lose trust in automation might be similar to that of interpersonal trust. Trust in automation can break down rapidly under time pressure or when conspicuous system faults or errors exist.[15,16]

Once lost, trust in automation, like interpersonal trust, can be hard to reestablish. However, people can be more forgiving of trust breeches from humans than from machines.[14] There's another asymmetry: "swift trust" between humans can occur because of a confession, which is an assertion that is immediately credible (perhaps on the basis of authority); that leaves the trustee vulnerable by admission of some mistake, weakness, or misjudgment; and that conveys a shared intent with regard to the topics of trust. Machines can't do this sort of thing ... yet.

Much research on trust in automation involves small-world studies using college students as subjects and simulated decision aids or scaled problems, so we must be cautious about generalizing the research results to "people," which most researchers do. Nevertheless, some have claimed that miscalibration is the norm—that is, people in general place unjustified trust in computer systems (*overtrust*, or taking advice because it comes from a computer that is called an "expert system"). Circumstances also arise in which people, in general, don't place enough trust in computer systems (*under-reliance*, or failure to rely on useful technological capabilities). This might be especially true for individuals who haven't had much exposure to technology.

Our conclusion is to tap into the interpersonal analogy—which might be unavoidable because of how we humans understand all this—but be cautious about it. We have some additional premises.

First, trust as a phenomenon is complex. Almost anything at one level (Am I achieving my mission goals?) can depend on something on another level (Is that warning indicator faulty?).

Second, trust is dynamic. Neither trusting (as a relation) nor trustworthiness (as an attribution) is a static state. Relations develop and mature; they can strengthen, and they can decay. Even when periods of relative stability seem to occur, trust will depend on context and goals. "Trust" is how we lump together a complex of multiple processes that are parallel and interacting. Processes often have no clear starting or stopping point. Given these inherent dynamics, we prefer to refer to trusting rather than to trust.

Once lost, trust in automation, like interpersonal trust, can be hard to reestablish.

Finally, workers in macrocognitive work systems always have some complex of justified and unjustified trust, and justified and unjustified mistrust in the technologies that mediate their interactions with the world. This is especially true for "intelligent" systems.[14,17]

Modes of Trust in Intelligent Systems

Trust in automation is limited to the degree that evidence from an operator's past experience does or doesn't provide adequate warrant for predicting how the machine will behave in novel situations. If adequate trust and mistrust signatures for every situation were always available, we could remedy this problem—but such expectations are unrealistic. Instead, trusting has what we might think of as multiple "modes."

Often, people don't pause to deliberately think about whether or how they trust their technology; they have what might be called *default trust*. For example, you might not think about whether the external drive will automatically back up your laptop overnight. After all, it's always done so. Mostly. On occasion, you might worry about a possible loss of data. But generally, we just move on and shrug these concerns off. On the other hand, trust in technology is sometimes very deliberative. In some situations, people think carefully about whether to trust a machine, and what the conditions on that trusting might be.

Then there is expertise. In a weather forecasting expertise project, forecasters were asked about their trust in technology. The answer often consisted of two components: a sarcastic chuckle, and a statement along the lines of "No, never. I always look for confirming and disconfirming evidence in what the data and the processing are showing."[18] Experts have sufficient experience with their technology to calibrate their trust, moving within the space defined by unjustified trust and unjustified mistrust, and within the space defined by justified trust and justified mistrust. That is, they have sufficient experience with the

IS-28-01-Hcc.indd 85 1/14/13 3:52 PM

technology to understand its competence envelope.

For so-called *fused data*, experts like to be able to "drill down" because machine processing always hides things as much as it reveals them.[19,20]

In addition, sensors might be miscalibrated, data might not come from a trusted source, data could have expired, and so forth. For example, in aviation weather forecasting, an expert who's anticipating severe weather will examine multiple data types to validate his or her interpretation of the primary radar data, and will deliberately seek evidence that severe weather might not develop. The expert will consider how the radar algorithms are biased under different circumstances.

Beyond justified mistrust, as people become more familiar with using technology, they can develop what might be called *negative trust*. This isn't quite the same as mistrust. Experience teaches people that technology will be buggy, that it will break down, that it will force the user to develop workarounds,[21] and that it will in some ways make the work inefficient.[5]

Modes of trust need to include some notion of *absolute* versus *contingent* trust. This is one circumstance in which the analogy to interpersonal trust breaks down. Sometimes, people have absolute or unconditional trust in a close relative to "do the right thing," for example. Although such trust is situation-dependent, it's rock solid within those pertinent situations (for instance, I trust my spouse to be true to me, but I don't trust my spouse to be able to pull me up to safety when I'm hanging off the edge of a cliff). The only form of trust in automation that is absolute is negative trust: people, at least everyone we've polled, are certain that any given machine will ultimately fail to work properly. Apart from this, trust in machines is always conditional or tentative—that is, the machine is trusted to do certain things, for certain tasks, in certain contexts.[4,17]

Some specification of what trust is about (actions, resources, and so on), the conditions or circumstances under which trust and reliance are in effect, and why the trust is in effect, will all be necessary for active exploration—that is, to establish, evaluate, and maintain trusting relationships in a macrocognitive work system. Trust that is absolute, verified, and reliable would, of course, be an unachievable ideal. Trust that's highly contingent, partially refuted, and very tentative (I'll trust you only in this circumstance, and only for now, because I have refuting evidence) would be a situation requiring close attention.

Trust that is absolute, verified, and reliable would, of course, be an unachievable ideal.

In analyzing macrocognitive work, we must specifically consider these and other modes as they play into the data a machine presents and actions it takes. Based on our premise that trusting is a process, we would infer that trusting is always exploratory, with the key variable being the amount of exploration that's possible and seems necessary.[22]

Outstanding Challenges

The active exploration of trusting-relying relationships can't and shouldn't aim at achieving some single stable state or maintaining some single metrical value; instead, it must aim to maintain an appropriate expectation. Active exploration by a human operator of the trustworthiness of the machine within the total work system's competence envelope will involve verifying reasons to take the machine's presentations or assertions as true, and verifying reasons why directives that the operator gives will be carried out.

We make no strong assumptions about whether or how a machine might somehow evaluate its own trustworthiness. Short of working some such miracle, we ask instead whether an AET computational system might support context- and task-dependent exploration of trusting. From a human-centered computing perspective, facilitating the active exploration of trusting should help the worker accomplish the macrocognitive work's primary goals. For this, a usable, useful, and understandable method must be built into the cognitive work that permits an operator to systematically evaluate and experiment on the human-machine relationship.

This entails numerous specific design challenges.

How can an AET system let the operator identify unjustified trust or unjustified mistrust situations? For a given work system, what circumstances define appropriate trust? What are the signatures that might suggest a mismatch? How can the system design enable an operator to identify early indicators to mitigate the impacts and risks of relying on or rejecting recommendations, especially in a time-pressured situation? How can a system design quantify the magnitude of mismatches?

How can an AET system mitigate unjustified trust? Once a mismatch

IS-28-01-Hcc.indd 86 1/14/13 3:52 PM

is identified, how can the machine convey trust and mistrust signatures in a way that helps the operator adequately calibrate their trusting and adjust their reliance to the task and situation? Following that, how might a machine signal the operator about unwarranted reliance?

How can an AET system mitigate unjustified mistrust? How might an operator mitigate the impact mistrust signatures have in circumstances where the mistrust is unjustified? How might a machine encourage justified trusting to promote appropriate reliance?

How can an AET system promote justified swift trust in the machine? The operator needs guidance to know when to trust early and "blindly"—and when not to. Can a machine promote the development of swift trusting while enabling the operator to maintain trust calibration?[23] In interpersonal trust relations, trust can be achieved rapidly if the trustee makes him- or herself vulnerable by making a confession. Can or how might confession-based swift trust be carried over to trust in automation? Could we achieve this via a method in which the technology's competence envelope is made explicit in descriptions of what the technology can't do or can't do well in different circumstances?[24]

How can an AET system mitigate the consequences of trust violations? How can a macrocognitive work system recover, rapidly, in circumstances where actions have been taken or decisions have been made on the basis of information or machine operations that had subsequently been found to be untrustworthy?

How can an AMT system promote justified swift mistrust? Certainly, it's good when swift mistrust emerges because the machine is making mistakes. The swift development of mistrust that is justified can be crucial, and might be anticipated by identifying early indicators—what we might call *mistrust signatures*. Inevitably, circumstances will arise in which automated recommendations won't be trustworthy—and the operator won't know that. Circumstances will also arise in which an operator shouldn't follow automated recommendations, even when they appear trustworthy. How can a system design mitigate the impacts and risks in a time-pressured situation of relying on or rejecting good recommendations from the automation? Certainly, circumstances exist (that is, unforeseen variations on contextual parameters) in which even the best software should in fact not be trusted, even if it's working as it should, and perhaps especially if it's working as it should.[22]

How can an AET system promote appropriate trust calibration when the situation is novel, and achieving primary task goals hinges on developing a new plan or new method on the fly? How can work system design (and training) encourage the varied interaction strategies that will accelerate learning in rare circumstances? Can we overcome the "this worked last time" attitude when context warrants a change?

How can an AET system include a useful and usable traceback capability so that the cognitive work is observable? A trustworthiness traceback capability must exist to support hindsight analyses of factors or events that contributed to increases or decreases in trust measurements. In a retrospective analysis, the active trust exploration system must support visualization of the complete data path and potential state change points.

Finally, we must consider three general entailments.

How can an AET system support the exploration of the work system competence envelope to allow calibrated trusting to emerge? This would involve enabling the operator to explore future possibilities in terms of how the measurements or data might modulate the machine operations. Specifically, operator inputs could modify software agent policies in a capability for trust-dependent task allocation. Furthermore, active exploration could include interrogating the machine to disconfirm trust hypotheses. This too might involve operator input regarding trust parameters or estimates, to influence subsequent machine operations.

How can an AET system simplify trust analysis? This final question is most important. Should, or how should the operator be able to collapse across these complex modes, measures, and dimensions to generate alternative ways of scaling trust and trustworthiness based on priorities or circumstance? The AET system must not only help the operator formulate the right questions when evaluating trust but simplify this process when the operator is potentially overwhelmed.

The final point is crucial: All of what we've expressed here means that *we must escape the traditional distinction between the operational context and the experimentation context*, especially given the ever-changing nature of the challenges that confront macrocognitive work systems. In an AET system as envisioned here, the operator can not only be made aware of trust and mistrust signatures but can also actively probe the technology (probing the world through the technology) to test hypotheses about trust, and then use the results to adjust subsequent human-machine activities (that is, reliance).

IS-28-01-Hcc.indd 87 1/14/13 3:52 PM

Of these three general entailments, the first two are design challenges; the last is a challenge for procurement. These are complex problems that we can't solve by taking an approach that all we need is more widgets.[2]

References

1. *Unmanned Systems Roadmap 2011–2036*, Office of the Undersecretary of Defense for Acquisition, Technology, and Logistics, Dept. of Defense, 2007.
2. J.K. Hawley, "Not By Widgets Alone," *Armed Forces J.*, Feb. 2011; www.armedforcesjournal.com/2011/02/5538502/.
3. *The Role of Autonomy in DoD Systems*, task force report, Defense Science Board, US Dept. of Defense, July 2012.
4. T. Sheridan, "Computer Control and Human Alienation," *Technology Rev.*, vol. 83, 1980, pp. 61–73.
5. R.R. Hoffman et al., "The Dynamics of Trust in Cyberdomains," *IEEE Intelligent Systems*, Nov./Dec. 2009, pp. 5–11.
6. E.W. Fitzhugh, R.R. Hoffman, and J.E. Miller, "Active Trust Management," *Trust in Military Teams*, N. Stanton, ed., Ashgate, 2011, pp. 197–218.
7. R.C. Mayer, J.H. Davis, and F.D. Schoorman, "An Integrative Model of Organizational Trust," *Academy of Management Rev.*, vol. 20, no. 3, 1995, pp. 709–734.
8. J.M. Bradshaw et al., "Toward Trustworthy Adjustable Autonomy in KAoS," *Trusting Agents for Trustworthy Electronic Societies*, LNAI, R. Falcone, ed., Springer, 2005.
9. C.L. Corritore, B. Kracher, and S. Wiedenbeck, "Online Trust: Concepts, Evolving Themes, a Model," *Int'l J. Human-Computer Studies*, vol. 58, 2003, pp. 737–758.
10. J.D. Lee and K.A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, vol. 46, no. 1, 2004, pp. 50–80.
11. 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, no. 3, 1996, pp. 429–460.
12. R. Parasuraman and V. Riley, "Human and Automation: Use, Misuse, Disuse, Abuse," *Human Factors*, vol. 39, no. 2, 1997, pp. 230–253.
13. Y. Seong and A.M. Bisantz, "The Impact of Cognitive Feedback on Judgment Performance and Trust with Decision Aids," *Int'l J. Industrial Ergonomics*, vol. 38, no. 7, 2008, pp. 608–625.
14. E.J. de Visser et al., "The World Is Not Enough: Trust in Cognitive Agents," *Proc. Human Factors and Ergonomics Soc. 56th Ann. Meeting*, Human Factors and Ergonomics Soc., 2011, pp. 263–268.
15. P. Madhavan and D.A. Wiegmann, "Effects of Information Source, Pedigree, and Reliability on Operator Interaction with Decision Support Systems," *Human Factors*, vol. 49, no. 5, 2007, pp. 773–785.
16. M.T. Dzindolet et al., "The Role of Trust in Automation Reliance," *Int'l J. Human-Computer Studies*, vol. 58, no. 6, 2003, pp. 697–718.
17. T.B. Sheridan and W. Verplank, *Human and Computer Control of Undersea Teleoperators*, tech. report, Man-Machine Systems Laboratory, Dept. of Mechanical Eng., Mass. Inst. of Technology, 1978.
18. R.R. Hoffman et al., "A Method for Eliciting, Preserving, and Sharing the Knowledge of Forecasters," *Weather and Forecasting*, vol. 21, no. 3, 2006, pp. 416–428.
19. D.D. Woods and N.B. Sarter, "Capturing the Dynamics of Attention Control from Individual to Distributed Systems: The Shape of Models to Come," *Theoretical Issues in Ergonomic Science*, vol. 11, no. 1, 2010, pp. 7–28.
20. G.A. Klein and R.R. Hoffman, "Seeing the Invisible: Perceptual-Cognitive Aspects of Expertise," *Cognitive Science Foundations of Instruction*, M. Rabinowitz, ed., 1992, pp. 203–226.
21. P. Koopman and R.R. Hoffman, "Work-Arounds, Make-Work, and Kludges," *IEEE Intelligent Systems*, Nov./Dec. 2003, pp. 70–75.
22. D.D. Woods, "Reflections on 30 Years of Picking Up the Pieces After Explosions of Technology," *AFRL Autonomy Workshop*, US Air Force Research Laboratory, Sept. 2011.
23. E.M. Roth, "Facilitating 'Calibrated' Trust in Technology of Dynamically Changing 'Trust-Worthiness,'" *Working Meeting on Trust in Cyberdomains*, Inst. for Human and Machine Cognition, 2009.
24. S.T. Mueller and G.A. Klein, "Improving Users' Mental Models of Intelligent Software Tools," *IEEE Intelligent Systems*, Mar./Apr. 2011, pp. 77–83.

Robert R. Hoffman is a senior research scientist at the Florida Institute for Human and Machine Cognition. Contact him at rhoffman@ihmc.us.

Matthew Johnson is a research scientist at the Florida Institute for Human and Machine Cognition. Contact him at mjohnson@ihmc.us.

Jeffrey M. Bradshaw is a senior research scientist at the Florida Institute for Human and Machine Cognition. Contact him at jbradshaw@ihmc.us.

Al Underbrink is senior analyst with Sentar. Contact him at al.underbrink@sentar.com.

cn *Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.*

IS-28-01-Hcc.indd 88 1/14/13 3:52 PM