The Dynamics of Trust in Cyberdomains

Robert R. Hoffman, Institute for Human and Machine Cognition
John D. Lee, University of Wisconsin–Madison
David D. Woods, The Ohio State University
Nigel Shadbolt, University of Southampton
Janet Miller, US Air Force Research Laboratory
Jeffrey M. Bradshaw, Institute for Human and Machine Cognition

All economic, social, and legal interactions are based on assumptions that individuals can verify identities; that they can rely on rules, institutions, and normative practice; and that they can be assured that their private space will remain protected. Monitoring, managing, verifying, auditing, and enforcing these assumptions are difficult online.

How are “cydentities” (cyber-identities) affirmed when they are constantly changing and are sometimes human, sometimes machine, and sometimes virtual? How can individuals and organizations derive persistent and valuable digital identities? How does digital activity change when threats to privacy and trust assumptions arise? How can trust be repaired? Is trust actually stronger once repaired? Is it trust we actually want from our digital systems, or is it accountability and the opportunity for recourse or redress?

Many such empirical and technological questions are raised by cyberstories that have become well known.1-4 Some of the more widespread cyberattacks are phishing and e-mail scams. Research from Get Safe Online (www.getsafeonline.org), published 9 February 2009, found that 23 percent of Internet users in the United Kingdom had either been a victim of a phishing scam in the preceding 12 months or knew someone who had been affected by the crime. Recent months have witnessed denial-of-service cyberattacks such as those on Facebook and Twitter.5 A Google search reveals numerous Web sites that provide advice about avoiding scams at Web sites.6 “Spoofs abound on the Internet. Web sites about hoaxes cover bogus science and technology … and report hoax computer viruses.”7 Thus, we see the advent of “trust evaluation tools” and protocols for a common language for online trust.8

An account in the Washington Post of the recent cyberattacks on South Korea and the United States points to how difficult it is to protect from attacks and even to identify the perpetrators.9

Cyberspace is a primary medium for the way the Air Force does business, whether it is used for command, control, communications … Almost everything I do is either on the Internet, an Intranet, or some type of network—terrestrial, airborne or spaceborne … Yet, everyone out there knows that hackers can potentially get into my network and slow down or corrupt it or cause me to lose faith in the networks or shut them down completely.10

The Lesson and Its Implication

For some decades now, cybersecurity has been a continuous game of catch-up. Clever hackers find some new scheme for malware, the malware is sent out and does its thing, the malware is sniffed out, and then the security people come up with a new software or hardware patch. This “clever game” is not likely to end. Indeed, attempts to improve software and hardware must continue.

But the real lesson is this: the window of vulnerability never closes. We must ask, therefore, what else might be done in addition to playing the clever
game? With this question we find issues that are directly pertinent to human-centered computing, both in leveraging what we know about humans, and in leveraging technology to amplify humans.

The cultural, social, psychological, and computational are merging into a Networld. Trust in and through technology will likely mediate the effectiveness of software and hardware in maintaining security. In this essay, we examine some human-centering issues for the Networld, placed primarily for convenience into five categories: antitrust in technology, a consensus on what “trust” is, interpersonal trust versus trust in automation, trusting as a dynamic process, and resilience engineering for the active management of trust.

Antitrust
Previous essays in this department have discussed how so-called intelligent technology triggers frustration and other types of negative affect. Frustration leads users to create kludges and workarounds.11,12 Antitrust—the skeptical assessment of a technology-reliant work system—is another form of negative affect that can trigger an effort to adapt. In a study of weather forecasters the experts were asked, “Do you trust your technology?” Frequently, they immediately responded, “Never!”13 This is antitrust: confidence (to the point of certainty) that the technology will choke, will frustrate, and will trigger a need for workarounds. It is no surprise, therefore, that psychologists and human-factors engineers have devoted considerable effort to understanding the circumstances of automation underuse, misuse, and even abuse.14,15

This suggests a combinatoric of justified versus unjustified trust and justified versus unjustified distrust. The calibration of trust in automation involves finding the sweet shifts within this constantly morphing space.16 For instance, novice users sometimes assume that computers are infallible. While they can adapt to the fallibilities of a human teacher, they are stymied when the computer gives what seems to be wrong feedback. People sometimes need help to understand how and when to shift from unjustified trust closer to justified distrust. On the other hand, experienced domain practitioners can become jaded, and are more likely to benefit from information that helps them calibrate their trust by shifting them from unjustified distrust to justified trust.

What the cyberresearch community needs is a set of powerful principles to guide people in reaching for desired states on such polarities as antitrust (unjustified mistrust) versus sceptical trust (justified trust); and contingent trust (conditional trust) versus unconditional trust (faith).

We also need to try to define some terms, as impossible as that often is.

A Consensus on What “Trust” Is
The concept of trust has been a topic of analysis in many disciplines, including philosophy (especially ethics), sociology, management science, and psychology. As Kieron O’Hara has noted, any comprehensive account 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.”20

Nevertheless, we find it interesting that, of all the fuzzy and abstract concepts inhabiting terra cognita (and ripe for debate in human-centered computing), there actually seems to be a consensus on how to define the multifaceted concept of trust.16,21,22

Trust has aspects of

• an attitude (of the trustor about the trustee),
• an attribution (that the trustee is trustworthy),
• an expectation (about the trustee’s future behavior),
• a feeling or belief (faith in the trustee, or a feeling that the trustee is benevolent, or a feeling that the trustee is directable),
• an intention (of the trustee to act in the trustor’s interests), and
• a trait (some people are trusting and more able to trust appropriately).

This is antitrust: confidence (to the point of certainty) that the technology will choke, will frustrate, and will trigger a need for workarounds.

Trust calibration is captured in one of the laws of macrocognitive work systems,17,18 which we dub Mr. Weasley’s Law, after that fictional character’s admonition of his wizarding daughter, “Never trust anything that can think for itself, if you can’t see where it keeps its brain.”19 Formally stated, the law is this:

Mr. Weasley’s Law: Workers in macrocognitive work systems develop unjustified trust and unjustified mistrust in their technology and the work system as a whole when the factors governing the technology’s activity are not visible.

This is antitrust: confidence (to the point of certainty) that the technology will choke, will frustrate, and will trigger a need for workarounds.

Trust calibration is captured in one of the laws of macrocognitive work systems,17,18 which we dub Mr. Weasley’s Law, after that fictional character’s admonition of his wizarding daughter, “Never trust anything that can think for itself, if you can’t see where it keeps its brain.”19 Formally stated, the law is this:

Mr. Weasley’s Law: Workers in macrocognitive work systems develop unjustified trust and unjustified mistrust in their technology and the work system as a whole when the factors governing the technology’s activity are not visible.
There are many perspectives on how trust plays a role in human relations,\textsuperscript{21,22} in decentralized sensing and networked systems,\textsuperscript{23,24} Analysts generally converge on the concept that \textit{trust reflects an assessment of the trustee's capabilities and competencies to respond to uncertain situations to meet common goals}. The outcome of the assessment is not always a single trust/do-not-trust judgment, but rather an assessment of what particular roles and tasks the trustee can be counted upon to accomplish successfully. The assessment is contingent; it is dynamic and evolves (or degrades) as events occur in the world and information and outcomes feed back to influence the trustor-trustee relation.

Trust can be thought of as a family of relations, certainly not a single relation. For instance, trust can be about different things: beliefs (information, data, knowledge), resources (such as valuables), or actions. Trust can be directional (the trustor trusts the trustee) or reciprocal (each party is both trustor and trustee). Trust can be general or contingent. Table 1 expresses some of the many relations.

There is also a consensus on the danger of anthropomorphism in generalizing ideas or research findings about interpersonal trust to the domain of trust in automation.\textsuperscript{14,16}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
\textbf{Type of trust} & \textbf{Beliefs for interpersonal trust} & \textbf{Beliefs for trust in automation} & \textbf{Trustor certainty that the trustee (or automation) will carry out the directives} \\
\hline
Unconditional trust (faith) & The trustor takes the trustee's assertions as true. & The trustor is confident that the automation is directable. The trustor takes the automation's actions as correct. & Certain. \\
\hline
Skeptical trust & The trustor takes some of the trustee's assertions as possibly true. & The trustor takes some of the automation's actions as possibly correct. & Somewhat certain. \\
\hline
Circumscribed trust & The trustor takes the trustee's assertions as true, for the time being, or with respect to a certain class of activities. & The trustor takes the automation's actions as correct for the time being, or correct with respect to a class of functions. & Certain for the time being, or certain with respect to certain activities. \\
\hline
Contingent trust & The trustor takes the trustee's assertions as true, depending on the circumstances. & The trustor takes the automation's actions as correct, depending on the circumstances. & Certain depending on the circumstances. \\
\hline
Antitrust & The trustor takes all of the trustee's assertions as false and potentially misleading. & The trustor might take some of the automation's actions as possibly correct, but also anticipates some to be wrong. & The trustor is certain that the technology will choke, will frustrate, and will trigger a need for workarounds. \\
\hline
Swift trust & The trustor has to take the trustee's assertions as true, because of urgent circumstances, often on the basis of trustee authority or position. & The trustor has to take the automation's actions as correct, because of urgent circumstances. & Somewhat certain. \\
\hline
Swift antitrust & The trustor takes the trustee's assertions as false and potentially misleading, because of emerging events or circumstances that reveal the trustee's genuine intentions. & The trustor takes any of the automation's actions as possibly wrong. & The trustor is certain that the technology will provide misleading information. \\
\hline
\end{tabular}
\caption{Some of the possible trust relationships.}
\end{table}

\textbf{Interpersonal Trust versus Trust in Automation}

Interpersonal trust is both fundamentally and subtly different from trust in automation. Trust between people typically involves expectations of intent and reciprocity. But while such basic aspects of social relationships might seem irrelevant to human-automation relationships, people often behave as if technology is a social actor.\textsuperscript{25} A direct comparison of trust in humans to trust in automation showed the dynamics of trust to be qualitatively similar.\textsuperscript{26} For instance, an aspect of interpersonal trust is directability—that is, that the trustor can direct the trustee. Directability also contributes to human-computer relations.\textsuperscript{27}

The feeling that the automation is a partner or a personality is sometimes enhanced by graphic personas. These can promote unjustified trust, and can be an undercurrent in automation abuse (that is, wanting an uncooperative machine to feel pain). The starkest differences between interpersonal trust and trust in automation have to do with the many and powerful effects of menus, graphic objects, and the like on automation misuse and underuse—effects that have no direct analog in interpersonal trust.\textsuperscript{28}
How trust (in either a human or a technological agent) develops depends on context and experience. Early evidence that a trustee might be untrustworthy (or that automation might have a high false-alarm rate) can subsequently make it hard for the trustor to develop trust in the trustee (or in the automation). On the other hand, if the trustee (or the automation) provides explanations of the untrustworthy behavior (or why the automation makes false alarms), the effect of fallible performance can be mitigated.

These empirical findings link trust concepts to two additional laws of macrocognitive work systems. First, people are explanation generators. If they are not given sufficient information to make satisfying explanations, they will make explanations anyway.

The Cognitive Vacuum Law: Workers develop mental models of the macrocognitive work system, including the technology. Thus, in most cases, a worker’s mental model of the automation is likely incomplete and not entirely veridical with respect to the designer’s intent. The potential for error is increased when the designer’s intent actually mismatches the worker’s goals and needs. This latter situation is often brought about by the reliance on a designer-centered design approach instead of human-centered design approach. Operators will develop a level of trust in the automation, but it can be over- or under-trust if the designers do not fill the vacuum to make the automation trustworthy. This leads us to another law:

Law of Surrogate Systems: Macro-cognitive work systems embody the stances, agendas, and goals of the designers.

Across all the research on trust in automation, a finding that percolates up consistently is that information allowing the user to understand what the automation does, why it does what it does, and what the designer’s intent is significantly promotes trust in the automation. Such information promotes a better understanding of the automation and leads to an appropriate level of trust, which in turn mediates between the feeling that the automation is trustworthy and the actual intention to go ahead and rely on the automation. Without this information, trust or distrust will still develop, but it will likely be disconnected from the automation’s actual capability.

A third perspective on trust is to think of trust using a verb rather than a noun form. What Table 1 points to is not just that trust is a family of relations, but that any given trustor-trustee relation is a dynamic thing.

The dynamics of trust are complex. Factors that contribute to this complexity include, first, threshold effects relating the level of trust to changes in reliance, and, second, contingencies between reliance and the information that guides the evolution of trust.

Threshold effects are reflected in the tendency to maintain a fixed level of reliance even as the level of trust changes, resulting in a dichotomous pattern of reliance. This pattern of reliance can in turn affect the information an operator has regarding the automation’s performance. In some cases, the automation’s performance is only perceivable when the person is relying on the automation. Threshold effects and contingent information availability complicate the dynamics of trust, and both can undermine the calibration of trust.

There are also interesting findings on the repair of trust. The role of apology and forgiveness as a means to repair trust have been extended to online transaction systems. This line of inquiry entails the possibility that...
Opportunities and Dangers

Cybersecurity and defense issues, including social deception, misdirection, influence, and manipulation, span all the venues of macrocognitive work in the Networld. How can we analyze networks to detect poorly calibrated trust, or increases and declines in trust? All networks are vulnerable to malicious attack, motivating our earlier comment, that the window of vulnerability never closes.

A direct implication of this is that the vulnerability of macrocognitive work systems to malware must be continuously evaluated and experimented on within the operational context where networks are being used, generated, grown, and adapted. How can net work continue when one believes that the network has been corrupted? Indeed, clever adversaries seek to maintain networks’ overall integrity so that they can manipulate data or applications and go undetected. For the individual net worker, how do we calibrate trust and distrust in technology when the technology is potentially compromised? For macrocognitive work systems, how do we create work methods and work processes for active management of trustworthy communities and work groups?

We might change our perspective from that of victim to that of influencer, from defense to offense. If you remove the highly connected individual nodes in a network you can do much damage, but you can also do damage by removing nodes that are “weakly” connected to other nodes. To concretize this, consider law enforcement combating organized crime, which by tradition has tight and secure networks. If the law

To achieve resilience, the technology and work methods must be created to support directability, responsiveness, reciprocity, and responsibility.

and machines adapt when events challenge their boundary conditions. To achieve resilience, the technology and work methods must be created to support directability, responsiveness, reciprocity, and responsibility. Workers must appropriately trust those parts of the work system that will respond adaptively to disrupting events, events that alter plans and activities in progress. Coordinating the adaptive responses makes decentralized systems resilient. Thus, trust in work systems can be thought of as confidence in this ability of different units at different echelons to act resiliently. Low levels of trust among units of a work system could play a critical role in maintaining a resilient system by signaling the need for additional resources or reconfiguration.

In macrocognitive work systems, trust can also be thought of as the expect of reciprocity from others. The parties involved in joint activity enter into a “basic compact,” an agreement (often tacit) to facilitate coordination, work toward shared goals, and prevent coordination breakdowns. In a reciprocal cooperative relationship, one human or machine agent relies on another to reciprocate in the future by taking an action that might give up some benefit in order to make both agents better off than they were at the starting point. Coordination and synchronization in distributed systems require reciprocity; otherwise, distributed systems are brittle and exhibit one or more maladaptive behavior patterns.

Both responsiveness and reciprocity emphasize anticipation, which introduces a forward temporal dimension of trust interactions. Both concepts point out that strategies change with changes in trust. For example, when one risks counting on others but anticipates little reciprocity or responsiveness, the result will be unstable, and the responsible party will shift to more conservative, independent strategies.

All the workers in macrocognitive work systems benefit if they can coordinate their activities, but coordination has costs. The challenge is to sustain the commitment to coordination and counteract any tendency to take advantage of others so as to benefit only one role. Decision making always occurs in the context of an expectation that one might be called to account for decisions. Expectations about what are considered adequate accounts, and the consequences for people whose
accounts are judged inadequate, are critical parts of a cycle of accountability. Breakdowns in responsiveness or reciprocity encourage role retreat and break the basic compact.

Models of resilient work systems emphasize how interdependent activities are co-adapted to each other and to changing short-term and long-term conditions. We need new control architectures to manage these resilient, distributed, multi-echelon, human-automation systems.

Finally, the issue of information accountability is arousing new interest in computer science. As we noted at the start of this essay, it is difficult to guarantee privacy when the technologies for information storage, aggregation, and analysis develop so rapidly. We live in an increasingly open information environment, in an increasingly linked Networld. Perhaps we need our technologies to be accountable such that the use of information is apparent, thus making it possible to determine whether a use is appropriate or legitimate in a particular context.

The laws of macrocognitive work systems, which we have presented in this and previous essays in this department, can be thought of as signposts along the path to resilience. The approach of designing for resilience offers some guidance and foundation for the active management of trust in cyberdomains.

References


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

John D. Lee is a professor in the Department of Industrial and Systems Engineering at the University of Wisconsin–Madison. Contact him at jdlee@engr.wisc.edu.

David D. Woods is a professor in the Cognitive Systems Engineering Laboratory at the Ohio State University. Contact him at woods.2@osu.edu.

Nigel Shadbolt is a professor of artificial intelligence in the School for Electronics and Computer Science at the University of Southampton, UK. Contact him at nrs@ecs.soton.ac.uk.

Janet Miller is senior cognitive systems engineer at the US Air Force Research Laboratory. Contact her at janet.miller3@wpafb.af.mil.

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

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