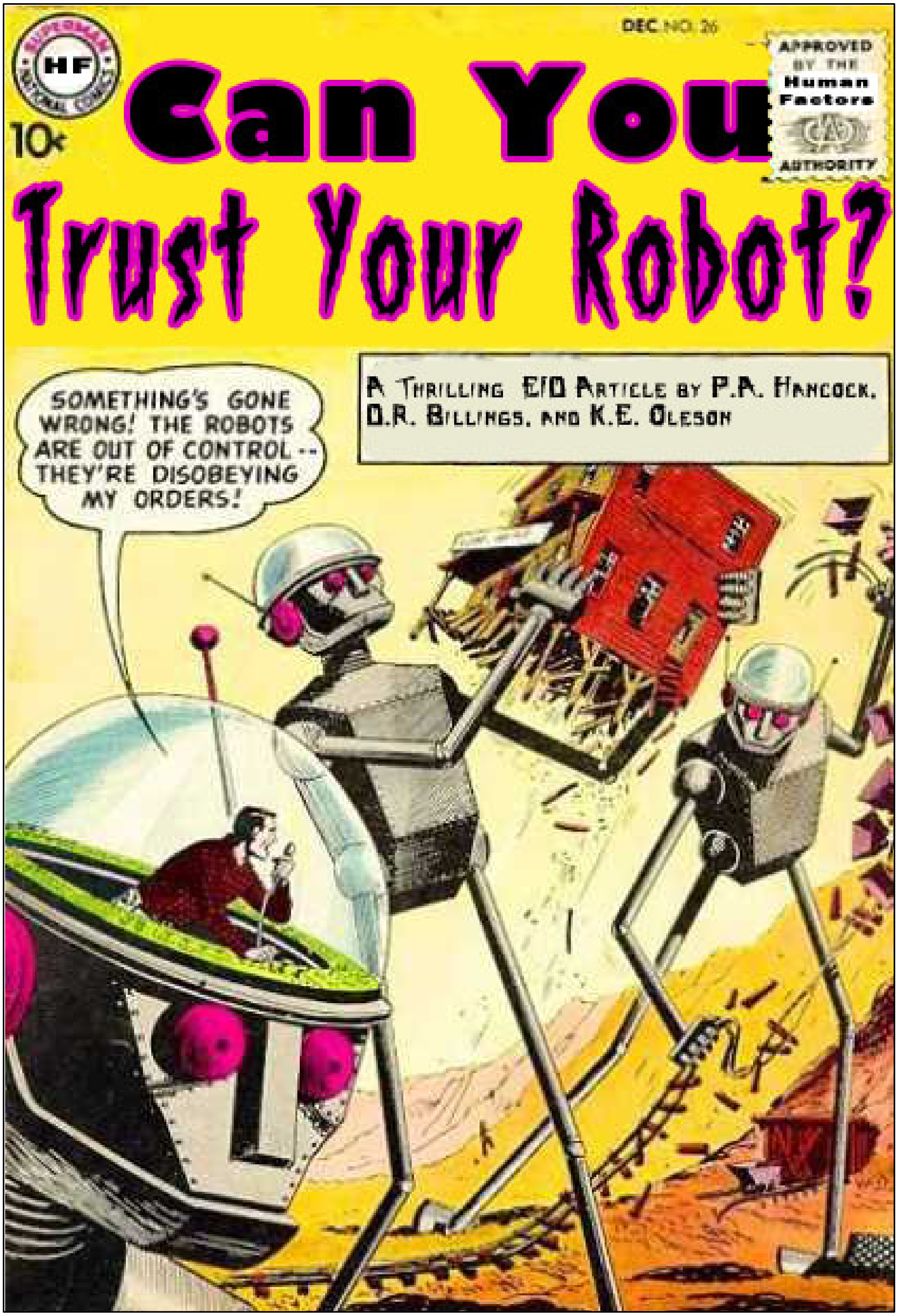

Can You Trust Your Robot?

A Thrilling E/O Article by P.A. Hancock, D.R. Billings, and K.E. Oleson

Can You Trust Your Robot?

BY P. A. HANCOCK, D. R. BILLINGS, & K. E. SCHAEFER

Modern robots may actively practice deception to achieve their goals, making trust one of the most crucial issues in their use and effectiveness.

Trust is an important aspect of any relationship or partnership, regardless of the context. Trust is equally as important in economic investments as it is in social institutions, such as marriage, or military teams. We define *trust* as the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others. Trust is generally described as a component of the interaction among conscious beings. However, we emphasize the relational nature of trust and recognize that trust need not necessarily be between what are traditionally considered sentient organisms. Trust can (and, in the case of robotic interaction, certainly does) involve other objects that do not express a self-determined, intrinsic intention.

In discussing trust, it is also necessary to define actions antithetical to trust – that is, deception. Deception plays an important, but often overlooked, role in the development and maintenance of trust. Our companion definition of *deception* thus becomes the capacity to induce false trust. In this article, we use these definitions to examine the issue of trust in human-robot relationships and comment on the crucial factors involved in, and issues associated with, modeling this emerging partnership. (See Figure 1 for a conceptual organization of these relationships.)

We recognize that human-robot interaction (HRI) is only one subset of the larger issue of human-automation interaction, concerns that have been explored by leaders in the human factors/ergonomics (HF/E) discipline (e.g., Parasuraman & Riley, 1997; Sheridan, 2002). What we are seeing, however, is a blending and blurring of operational principles between general automation and the spectrum of emerging robotic entities. Despite this trend, robots can still be considered a class in and of themselves. Modern robots include effector systems and embrace the larger concerns of action at a distance, embodied, for example, in armed military systems such as unmanned combat aerial vehicles. The use of robots in the military domain provides a specific focus related to the idea of collaborative autonomous agents. Such agents are free to roam the environment in a manner similar to human beings while also expressing intention in a somewhat analogous manner. This latter vision represents the near future of robots, but what of their history?

A Brief History of HRI

The creation and use of automata of various kinds goes back to human antiquity (De Solla Price, 1974). Legends of robotlike creatures, such as the Golem, permeate the folklore of different cultures (Wiener, 1963). The word *robot* itself derives from Čapek's (1921) usage in his provoking play *R.U.R.(Rossum's Universal Robots)*, which served to publicly recapture the myth of the Golem. Čapek's play is very relevant to the present work because it depicts a rebellion by robots and the ensuing conflict with their human masters, the ultimate breach of trust (see Figure 2).

This human fear of robots' becoming self-aware, rebelling and destroying humans, has permeated any number of subsequent motion pictures (e.g., *The Terminator*, *I, Robot*; see also Hancock, 2009). This fear also served as a basis for the generation of the purported "laws of robotics" promulgated by the scientist and science fiction author Isaac Asimov: (1) A robot must not harm a human or allow a human to be harmed. (2) A robot must obey a human unless orders conflict with the first law. (3) A robot must protect itself from harm unless this violates the first two laws. Finally, (Zeroth Law) a robot must not harm humanity or, by inaction, allow humanity to come to harm (Asimov, 1942, 1985).

In some ways, modern robots are sneaking up on us. This is because we have always had a fairly prototypical stereotype of what we expect a robot to look like and how we expect it to perform. According to science fiction

FEATURE AT A GLANCE: It is proposed that trust is a critical element in the interactive relations between humans and the automated and robotic technology they create. This article presents (a) why trust is an important issue for this type of interaction, (b) a brief history of the development of human-robot trust issues, and (c) guidelines for input by human factors/ergonomics professionals to the design of human-robot systems with emphasis on trust issues. Our work considers trust an ongoing and dynamic dimension as robots evolve from simple tools to active, sentient teammates.

KEYWORDS: trust, robot, robotic systems, automation, deception, human-machine systems, human-robot interaction.

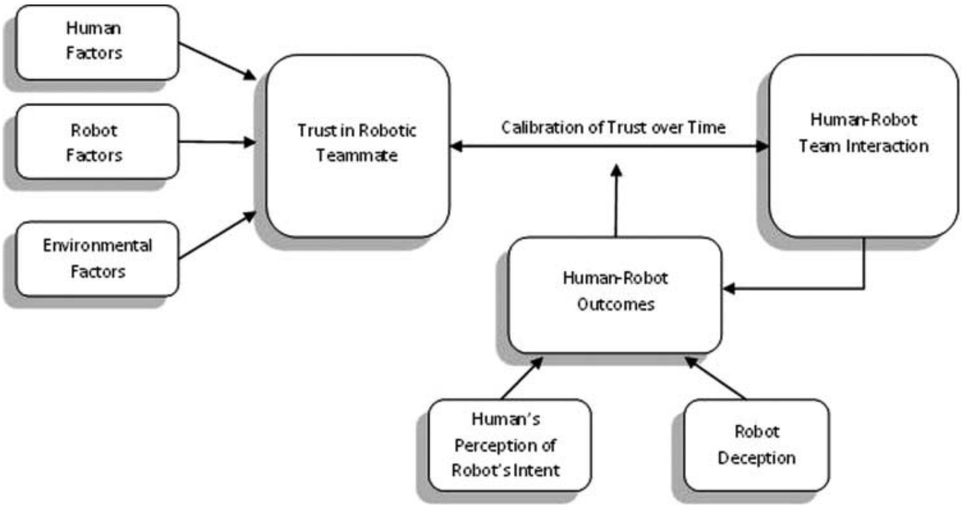

Figure 1. A conceptual organization of human-robot trust influences.

Figure 2. From the Wikipedia entry for Čapek's play, showing the robots in rebellion in the play itself.

tradition, the robot is meant to appear fundamentally human and to react in a somewhat human manner, receiving voice input and generating actions often via corresponding voice output and the control of human-type limbs. The robot is thus a surrogate human but, of course, not fully human. Therefore, how a robot reacts now – and potentially will react in the future – is contingent on how we perceive its limitations and constraints. If we reference Asimov, we intrinsically expect the robot to do what we think it should, and therefore a robot has traditionally been considered essentially mindless.

In reality, robots as they are operating today have entered the world in many differing forms. Robots (aside from those found in science fiction stories) have been defined in a variety of ways, but a standard definition of a robot was created by the Robot Industries Association (RIA): a robot is "a reprogrammable, multifunctional manipulator designed to move material, parts, tools, or specialized devices through variable programmed motions for the performance of a variety of tasks" (Hamilton & Hancock, 1986, p. 70). If we adhere strictly to this definition, we currently use robots all the time (e.g., modern-day commercial aircraft, industrial robots). This common usage comes without the intense fear of rebellion. (It is true, however, that we might consider automated "mode" changes as a form of insensate conflict; see Sarter & Woods, 1995).

What is a robot? As with everything in life, we base our mental models of a robot on what we have seen and experienced. Science fiction thus has played, and continues to play, a large role in what we have come to expect a robot to be and do. For instance, one of the first things that comes to mind when thinking of "robots" is the catchphrase, "Danger, Will Robinson!" – uttered numerous times by Robby the Robot on the television series *Lost in Space* (although this particular robot character was first introduced in the film *Forbidden Planet;* see Figure 3). Robby was designed to abide by Asimov's laws of robotics and was a protector and an arbiter of what is fair, and occasionally could give orders. In this sense, Robby exercised a good deal of power over others. However, this power originated from the programmer's intentions, illustrating that although robots may wield extraordinary power, they have no intrinsic will, a common paradox in robotics. (Perhaps this capacity for action and yet this simultaneous sense of powerlessness is why children frequently identify with robots.)

Commander Data is another memorable robot, from *Star Trek: The Next Generation.* The robotic Pinocchio constantly searches for a soul and remains forever in a state of conflict. Data is physically and mentally much more powerful than his human colleagues but remains subservient to them. This conflict reflects our human perception of robots. We experience internal conflicts and, therefore, the

Figure 3. From the Wikipedia entry for the film Forbidden Planet*, depicting Robby the Robot.*

robot must be made to experience those same conflicts. In a similar way, we assume that robots should have a physical form akin to the human body and artificial intelligence that operates similar to a human brain. Danger arises because we see robots as constrained human beings and assume they possess other human characteristics as well. This is an unfortunate case of attribution error.

The imaginations of science fiction authors have undoubtedly influenced our perceptions of robots. The robots that exist today are not nearly as sophisticated as the humanoid robots from storybook pages. Humans generally view present-day robots as tools that can extend their capabilities and, to some degree, compensate for human limitations (Chen, Barnes, & Harper-Sciarini, 2010). For a robot to be effective in this sense, the human must trust it to do its job consistently and effectively. Recently there has been a clear shift toward incorporating robots as active, interdependent teammates (i.e., a movement toward robotic peers; Groom, 2008). This shift in perception has increased the number of issues that arise in human-robot collaborative environments. Not only must humans trust the robot to do its job, but they must trust the robot to act in the best interests of a larger organizational entity: the team.

Aside from the potential issues associated with robotic teammates, there are definite advantages to this type of mixed team. When human team members trust appropriately in a robotic counterpart, "the robot's abilities may be augmented by the other specialties of the group, thus creating a collective that is more survivable than any single individual" (Wagner, 2009, p. 2). Although existing robots are far from what science fiction has envisioned, the issue of trust in robots has begun to emerge as a very practical design concern.

Development of Trust

Trust is a relational concept that requires a minimum of three elements. Two agents are needed: one an actor or transmitter of information and the other being acted on, or the receiver of that information. A viable communication channel between these two agents is also a necessary requirement. From a human-centric point of view, the development of trust is affected by an interplay of characteristics of the human (the present actor), the robot (which is currently the receiver), and the communication channel – primarily at this time the operational environment (and see Shannon & Weaver, 1949). In addition, the outcomes of trust provide feedback to the human teammate, leading to adjustments or changes in the degree of trust in the robotic system (see Figure 1 on page xx). Consequently, this calibration of trust may affect reliance on the system, the effectiveness of the human-robot collaboration, and thus the overall interaction.

Although we can seek to identify ways to calibrate trust in robotic teammates, this process is complicated by the fact that technology allows us to create robotic systems that can employ deception to their advantage. Robots can be programmed to deceive by reasoning about and predicting the impact that deception will have on a particular person, group of persons, or target (see Wagner & Arkin, 2011). When deception is used, the outcomes of HRI may well be very different from the outcome initially expected by the individual (as a result of expected robot intentions). This mismatch alters everyone's level of trust in the robotic system. Elevated distrust, a derivation of deception, thus affects not only the deceived but, in this case, the original operator: the robot's teammate.

Trust and Deception

As we have suggested, science fiction can play a strong role in the formative structure of our expectations about a robot, and it also provides a clear picture of the issues of trust and deception in HRI. In the television series *Knight Rider*, two distinct automated vehicles – K.I.T.T. and K.A.R.R. – illustrate trust in vehicle automation and fear of the potential deception of this automation. K.I.T.T. (Knights Industries Two Thousand, or more recently named Knights Industries Three Thousand) is designed to protect human life and demonstrates a collaborative trust-based relationship with its driver. Conversely, K.A.R.R. (Knight Automated Roving Robot, or more recently named Knight Auto-cybernetic Roving Robot exoskeleton) is motivated by the directive for self-preservation.

Fear of automation is also illustrated by Commander Data's evil "twin" brother, Lore. Whereas Data's mission is to assist and protect humans, Lore is self-serving and quick to betray the humans who depend on him. Both K.A.R.R. and Lore are examples of robotic characters that were created to represent humans' fear of deception or the fear that the technology will somehow work against our interests.

A variety of entities use deception to their advantage, including humans, animals, insects, and, indeed, almost all living organisms. Even some robotic systems now use deception (see the efforts to identify these acts of deception in the movie *Blade Runner*). The process of deception begins when the deceiver (agent) transmits false information to the person being deceived. This person receives the false information and interprets it, which then influences the actions selected by that person. These actions subsequently serve to benefit the robot deceiver but not the person being deceived and, indeed, are most frequently to his or her detriment in some manner (Wagner & Arkin, 2011).

Researchers have recently developed a program that enables a robot to determine the gullibility of another robot or human and then trick it, him, or her into acting a certain way (Firth, 2010). In other words, a human may believe that the intention of the robot is to provide assistance to achieve a mutual goal, but instead, the robot may be acting toward its own goal, unbeknownst to that person. This example highlights the importance that intention, and the perception of intention, plays in the development of trust and in its fracture.

Trust in automation can be understood in terms of an individual's inferences about the robot's (or designer's) intent. Does the individual's understanding of the intent of the robot match the actual intent? If the answer is yes, then trust is likely to develop. If the answer is no, then deception may be present and trust may not be as likely to develop. Second, trust in robots can be understood in terms of the outcomes of HRI, which provide feedback to the individual about whether the perceived intentions of the robot actually match the real intentions of the robot.

The perception of the robot's intentions along with the observed outcomes of cooperating with a robot may be at odds if deception is employed. The robot's failure rate (at least according to the human's perception) will increase, and the perceived reliability of the system will decrease. Consequently, the human may not be as quick to trust or use the robot in future operations. An individual may also experience changes in self-confidence in his or her abilities, which may affect mutual trust levels. Therefore, it is important to consider how deception may affect the development or deterioration of trust in human-robot partnerships.

Can Robotic-Based Deception Be Beneficial?

Some researchers have cautioned against designing robots for which one primary intent or goal is deception, fearing that it may only complicate the issue of trust in robotic systems and even "poison the well" of robot use in general (Firth, 2010). Humans see it as greater risk to trust anything they know may be deceptive. Trust in a robot can also be affected by the frequency of deception (Wagner & Arkin, 2011). Does the robot deceive constantly (i.e., one should never trust it)? Does the robot never deceive (i.e., one should always trust it)? Does the robot only occasionally deceive (i.e., one should sometimes trust it)?

The frequency of deception affects the level of trust in human-robot teams, but are there ever cases in which robot deception can lead to positive outcomes for all involved? The answer seems to be a qualified yes; there are certain contexts in which deception may be warranted. For example, suppose a robot is responsible for collecting massive amounts of information from the environment. However, only a few bits of that information are needed for the human to make a decision or select an action. In this case, is it better for the robot to present the human with all possible information, increasing the human's workload, or for the robot to present only the information pertinent to the decision at hand, in effect deceiving the person into believing that this is the only information collected?

Although the latter option technically constitutes a form of deception by omission on the part of the robot, it also avoids potentially costly mistakes that could be made if an individual is overloaded. Deception may also be advantageous – for example, in search-and-rescue missions wherein cooperation is encouraged from victims who may be in shock or acting hysterically (Wagner & Arkin, 2011).

In terms of military operations, deception often gives warfighters an advantage on the battlefield. Thus, research in these contexts is important (Gerwhr & Glenn, 2000). For instance, research suggests that deceitful robot tactics can benefit battlefield situations in which a robot needs to elude capture and avoid detection to carry out its mission or prevent critical information from falling into enemy hands (Wagner & Arkin, 2011). In these cases, the trust that an individual, or teammate, has in the robot may not necessarily change on the basis of specific, goal-related, deceptive activity. That is, trust level may be unaffected unless the deception leads to a negative outcome contrary to the individual's perception of robot intent. In general, deception can be harmful when deceiving a fellow teammate, but it can be potentially valuable when deceiving an opponent or even an individual who is not part of the team (e.g., a neutral bystander).

As indicated, deception cannot be considered an adverse thing in all cases. Sometimes deception may be warranted to accomplish a goal that is beneficial to the entire team. However, when deception leads to negative outcomes, trust is almost always affected. Further research needs to address when and how trust can be influenced by such cases of deception. There remains the moral dimension of such designed deception (Hancock, 2009).

Human Factors Design Inputs to Human-Robot Trust Interactions

Quantitative analyses of potential influences of HRI trust. Up to this point, we have sought to establish that both trust and deception are important elements in HRI. In related research, we have quantitatively evaluated the degree to which trust in HRI is influenced by environmental factors (e.g., team collaboration, communication, and culture), robot-related factors (e.g., predictability, reliability), and human characteristics (e.g., propensity to trust, mental workload; see Hancock et al., 2011). Our findings to date have indicated that robot performance-based factors (e.g., predictability, reliability) and robot attributes (e.g., proximity, adaptability) are the largest contributors to trust in HRI. Consequently, knowledge of performance, functionality, and capabilities of the robot can facilitate the development of trust and provide a foundational recommendation for robot design and associated human training.

Next, we identify several additional potential design and training guidelines that can be applied to HRI to exploit the opportunities for trust and mitigate the potential problems of deception.

Transparency of the robotic system. If deception is the goal, designers can attempt to create a false sense of transparency. If gaining trust is the goal, engineers should design the robot in a way that allows the user to observe the system and better understand what the system is doing. The following are implications for calibrating trust in a robotic system:

- Make the robot's functional relationships accessible and clear to the human teammate (Uggirila, Gramopadhye, Melloy, & Toler, 2004).
- Appropriate system display design will help users know the system's functional capabilities and limitations (Cuevas, Fiore, Caldwell, & Strater, 2007).

Human teammate's knowledge of the robotic partner. If deception is the ultimate goal, designers can make the human feel very comfortable with the robotic partner by giving false or inaccurate information. If trust is the goal, designers should aid human understanding of what the robot will contribute to the team as opposed to what the human teammate will contribute. The following guidelines illustrate how trust can be facilitated through increasing the human's knowledge of the robotic system:

- Humans should be informed of the robot's capabilities and limitations, so that they are aware of how the robot will achieve specific goals (Chen, Barnes, & Harper-Sciarini, 2010).
- Humans should be trained in the robot's intended use and how to interact with the robot appropriately (Desai, Stubbs, Steinfeld, & Yanco, 2009).
- Humans should be informed explicitly about the robot's known level of reliability (Bagheri & Jamieson, 2004).
- Past performance of the robot (and explanations of such performance and any associated errors) should be known to the human so that he or she is able to better predict the robot's behavior (Chen et al., 2010; Lee & See, 2004).
- Feedback regarding current system performance should be given to the human teammate, especially if humans and robot(s) are not colocated (Chen et al., 2010; Hoffman et al., 2009).
- Emotional requirements of the human operator must be considered. The human operator should not feel unnecessary to the system as a whole (Hancock, Pepe, & Murphy, 2005).
- The environments in which the robot will be used should be well understood. Consequences of actions can be different in different contexts and cultures. The role of trust is influenced by cultural differences in power distance, uncertainty avoidance, and individualist and collectivist attitudes (Lee & See, 2004).

Creation of mental models. When deception is involved, the designer attempts to encourage mental models that do not align with the true robotic system functioning. Conversely, when a teammate's trust is desired, correct mental models are encouraged and facilitated in the following ways:

- Encourage appropriate mental model development among human teammates that is in accord with the designer's intent (what is the intent, what does the robot do, and why does the robot do this? see Hoffman et al., 2009).
- Foster the creation of shared mental models among team members (Neerincx, 2007).

Use of adaptive function allocation and adaptive technology. When deception is the goal, adaptive technology may not be preferable to alleviate high workload, because high levels of workload and stress may lead to overreliance on the robotic system, which can be desirable if the robot plans deception. On the contrary, adaptive technology can be used to help build trust in a robotic system:

- Reduce task load to prevent over- or underreliance (influenced by trust) on the robotic system by adapting to the cognitive needs of individuals (Cosenzo, Parasuraman, Novak, & Barnes, 2006).

Conclusion

There can be little doubt that we will see an ever-increasing penetration of automated and semiautomated systems into our world. As very specific forms of such hybrid automation, robots will be central to that evolution. As with their fixed and immobile cousins, we will continue to have questions about the degree to which we should trust them. But what will happen if we cannot trust them?

References

Asimov, I. (1942, March). Runaround. *Astounding Science Fiction*, pp. 94–103.

Asimov, I. (1985). *Robots and empire.* New York: Doubleday Books.

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behavior. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* (vol. 1, pp. 212–217). The Hague, Netherlands: IEEE.

Čapek, K. (1921). *R.U.R. (Rossum's universal robots).* London, UK: Penguin.

Chen, J. Y. C., Barnes, M. J., & Harper-Sciarini, M.(2010). Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE Transactions on Systems, Man, and Cynernetics—Part C: Applications and Reviews*, *41*, 435–454.

Cosenzo, K. A., Parasuraman, R., Novak, A., & Barnes, M. (2006). *Implementation of automation for control of robotic systems.* Aberdeen Proving Grounds, MD: U.S. Army Research Laboratory.

Cuevas, H. M., Fiore, S. M., Caldwell, B. S., & Strater, L. (2007). Augmenting team cognition in human-automation teams performing in complex operational environments. *Aviation, Space, and Environmental Medicine*, *78*(5, Section 2), B63–B70.

De Solla Price, D. (1974). Gears from the Greeks: The Antikythera mechanism. A calendar computer from ca. 80 B.C. *Transactions of the American Philosophical Society, New Series*, *64*(7), 1–70.

Desai, M., Stubbs, K., Steinfeld, A., & Yanco, H. (2009, April). *Creating trustworthy robots: Lessons and inspirations from automated systems.* Paper presented at the AISB Convention: New Frontiers in Human-Robot Interaction, Edinburgh, Scotland.

Firth, N. (2010, September 11).The real 2001: Scientists teach robots how to trick humans. *Daily Mail.* Retrieved from http://www.dailymail.co.uk/sciencetech/article-1310788/The-real-2001-Scientists-teach-robots-trick-humans.html#ixzz1Q3PxZZl2

Gerwehr, S., & Glenn, R. W. (2000). *The art of darkness: Deception and urban operations.* Santa Monica, CA: RAND.

Groom, V. (2008). What's the best role for robot? Cybernetic models of existing and proposed human-robot interaction structures. In J. Filipe, J. Andrade-Cetto, & J. Ferrier (Eds.), *Proceedings of the Fifth International Conference on Informatics in Control, Automation, and Robotics* (pp. 323–328). Funchal, Madeira, Portugal: INSTICC Press.

Hamilton, J. E., & Hancock, P. A. (1986). Robotics safety: Exclusion guarding for industrial operations. *Journal of Occupational Accidents*, *8*, 69–78.

Hancock, P. A. (2009). *Mind, machine, and morality*. Chichester, UK: Ashgate.

Hancock, P. A., Billings, D. R., Oleson, K. E., Chen, J. Y. C., DeVisser, E., & Parasuraman, R. (in press). A meta-analysis of factors impacting trust in human-robot interaction. *Human Factors.*

Hancock, P. A., Pepe, A. A., & Murphy, L. L. (2005). Hedonomics: The power of positive and pleasurable ergonomics. *Ergonomics in Design*, *13*(1), 8–14.

Hoffman, R. R., Lee, J. D., Woods, D. D., Shadbolt, N., Miller, J., & Bradshaw, J. M. (2009). The dynamics of trust in cyberdomains. *IEEE Intelligent Systems*, *24*(6), 5–11.

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

Neerincx, M. A. (2007). Modelling cognitive and affective load for the design of human-machine collaboration. In D. Harris (Ed.), *Engineering psychology and cognitive ergonomics, HCII 2007, LNAI 4562* (pp. 568–574). Berlin, Germany: Springer-Verlag.

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

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, *37*, 5–19.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of information.* Chicago: University of Illinois Press.

Sheridan, T. B. (2002). *Humans and automation: System design and research issues.* New York/Santa Monica, CA: Wiley/Human Factors and Ergonomics Society.

Uggirila, A., Gramopadhye, A. K., Melloy, B. J., & Toler, J. E. (2004). Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics*, *34*, 175–186.

Wagner, A. R. (2009). *The role of trust and relationships in human-robot social interaction* (Doctoral dissertation). Georgia Institute of Technology, Atlanta.

Wagner, A. R., & Arkin, R. C. (2011). Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics*, *3*, 5–26.

Wiener, N. (1963). *God and Golem Inc.* Boston, MA: MIT Press.

Peter Hancock is Provost Distinguished Research Professor and Pegasus Professor in the Department of Psychology and the Institute for Simulation and Training at the University of Central Florida. He is a Fellow of and a past president of the Human Factors and Ergonomics Society. He may be reached at peter.hancock@ucf.edu.

Deborah Billings is a postdoctoral research associate with the Institute for Simulation and Training at the University of Central Florida. She holds a PhD in applied experimental and human factors psychology and an MS in modeling and simulation, both from the University of Central Florida.

Kristin Schaefer, MS, is a doctoral candidate in the Modeling and Simulation Program at the University of Central Florida. She obtained her MS in modeling and simulation from the University of Central Florida and her BA in psychology from Susquehanna University.

The research reported in this document was performed in connection with Contract No. W911NF-10-2-0016 with the U.S. Army Research Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of the U.S. Army Research Laboratory or the U.S. government unless so designated by other authorized documents. Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein. We would like to thank Jessie Chen, Keryl Cosenzo, and Susan Hill for their oversight and direction of this project. We also thank the editor and anonymous reviewers for their comments on earlier drafts of this article.