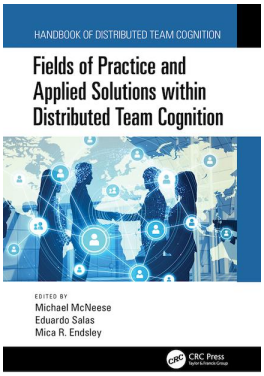

Fields of Practice and Applied Solutions within Distributed Team Cognition

Michael D. McNeese, Eduardo Salas, Mica R. Endsley

The Cognitive Wingman

Publication details
https://test.routledgehandbooks.com/doi/10.1201/9780429459542-9
Michael D. Coovert, Matthew S. Arbogast, Ewart J. de Visser
Published online on: 29 Sep 2020

9 The Cognitive Wingman
Considerations for Trust, Humanness, and Ethics When Developing and Applying AI Systems

Michael D. Coovert, Matthew S. Arbogast, and Ewart J. de Visser

CONTENTS

TRUST IN HUMAN-MACHINE SYSTEMS

As we begin our consideration of trust and what it means in the context of human-machine teams, especially when AI enters the mix, it is important to "begin at the beginning" and define our core construct, namely trust. There are many definitions of trust, but arguably two have become more dominant than others (Lewicki & Brinsfield, 2017). The first is offered by Rousseau, Sitkin, Burt, and Camerer (1998, p. 395), whose focus is on how trust makes us vulnerable to one another. For these authors, trust is the "intention to accept vulnerability based upon positive expectations of the intentions or behavior of another." Acceptance of vulnerability seems a reasonable description of trust in human-machine systems as well, for if an individual utilizes a technology such as a self-driving car or the autopilot in a plane, one is certainly exposing oneself to the consequences of the action of the technology. A second definition is proposed by McAllister (1995, p. 25), "the extent to which a person is confident in, and willing to act on the basis of, the words, actions and decisions of another." Here the central focus is on the relationships between individuals or entities. In our context, one of the entities is a technology. As Lewicki and Brinsfield (2017) point out, research around these two perspectives of trust illustrate it is multidimensional, having three factors: cognitive, affective, and behavioral. The cognitive factor relates to those specific expectations and beliefs relating to the other; affective refers to the emotional connection with the trusted entity; and behavioral refers to those actions taken by an individual in a trusting relationship. In addition to these three dimensions of trust, it should be noted there is emerging evidence for a neurological basis. Examination via fMRI reveals that trust and mistrust involve different functional areas of the brain (Krueger et al., 2007) with certain neurochemicals (oxytocin) also being involved (Zak, 2017). While we do not delve into it here, this does impact those interested in brain-computer interfaces for monitoring levels of trust and the use of neurochemicals for manipulating susceptibility to trusting.

Trust is one of those ideas that everyone knows about and experiences. Yet, in terms of operationalizing the construct, there are many boundary conditions to consider; fortunately, there is also much agreement. For extensive reviews, visit Lewicki and Brinsfield (2017) for interpersonal trust and Lee and See (2004) for trust in automation. Bringing this vast literature into focus for the purpose of our chapter, we highlight a few topics central to trust in technology before moving to consider humanness and ethical issues in Artificial Intelligence (AI) systems.

Trust in Technology

When thinking of trust, one often considers such notions as competence, reliability, and benevolence (McAllister, 1995). This is certainly true for trust in interpersonal relationships, and we believe it extends to as trust in technology as well. We now describe some exemplar studies central to trust in AI systems.

Measuring Trust and Modeling Its Growth

To understand trust and how it changes in the context of either interpersonal relationships or technology, we must be able to measure it and then model it. Coovert, Miller, and Bennett (2017) demonstrated the utility of latent growth models and latent change score models for both measurement and modeling purposes. They examined the growth of the two types of trust (cognitive and affective) found in McAllister's (1995) framework. Coovert et al. demonstrate trust in teammates develops first with cognitive trust and once that is established, affective trust begins to grow. Utilizing bivariate coupling with other constructs of interest (e.g., satisfaction with teammates, team performance) the growth of trust within a dynamical system can be modeled along with other parameters of interest and importance.

Facilitating Trust

A system, especially one with advanced technology, will not be used until it is trusted. Given this truism, several research programs have been developed to determine how to facilitate the development of trust. Recently, Lyons, Ho et al. (2016) focused on the factors associated with the use of a controlled flight into terrain (CFIT) avoidance system in aircraft. The purpose of the system is to take over control of flying the aircraft should the pilot be flying too close to hazards (or in the extreme become incapacitated, as can happen when a plane is performing maneuvers exerting high G-loads on the pilot). Lyons et al. expressed three key takeaways from their study. The first is the algorithms (used to take control) must be nuisance-free, in that you do not want to warn a pilot about an impending crash too early. This finding is concomitant with the false alarm rate described later in the chapter. As it turns out, there are individual differences associated with preferences for the timing of the warning. Since individual differences are involved, this is a parameter that could be set by each pilot. For example: If *airspeed* is greater than A, *closure-rate* greater than or equal to B, and *hazard-type* is C then display CFIT warning type *Bravo*. Pilots are individually allowed to set values for *airspeed*, *closure-rate*, *hazard-type*, and CFIT *warning type*. Since each individual configures the system according to their own preference, this leads to increased propensity to trust.

A second finding also relates to individual differences. When aircraft control is taken by the system it should fly the evasive maneuvers consistent with a pilot's preference for flying those maneuvers. Trust is personal, so if you want the pilot to trust the automation, it needs to take over in a fashion that is consistent with the pilot's mental model of preferred actions. The third finding relates to building trust through

training. By using training to assess, demonstrate, and verify the reliability of the systems, one builds trust in the system.

Consider the findings from the above research: (1) understanding false alarms and why they occur, (2) grasping a system's parameters and how they can be tailored for individual preference, and (3) evaluating fit with one's mental model. Is there a common factor underlying these findings? One can argue *transparency* is such a factor. The idea of transparency has been examined by several researchers working in the area of human-robot interaction (HRI). In laying a groundwork for transparency, Kahn et al. (2012) demonstrate that humans hold robots accountable, at least more accountable than they hold other inanimate objects such as a toaster that burns bread or the ice maker that fails to make ice in a freezer. Development of effective HRI interactions occur by having individuals understanding the robot's ability, intent, and situational constraints (Coovert et al., 2014). Aspects of these ideas were also posed by Kim and Hinds (2006), who suggest transparency should increase as automation accelerates along a dimension anchored by complete autonomy.

Furthermore, Burke, Coovert, Murphy, Riley, and Rodgers (2006) and Lyons (2013) argue people need to understand robots and the robots need to have an "understanding" of people. Lyons suggests the robot-to-human transparency should take place on several levels, with separate models for intention, task, analytical, and environment. He proceeds to provide the following description and elaboration of the models. An *intentional model* conveys the "why" of the system. This helps place the human's understanding of the system in the correct context. It presents the design, purpose, and intent of the system. In addition to "why," it also provides "how" the robot performs the actions. Similar to Asimov's (1942) three laws of robotics, it provides an understanding of the robot's moral philosophy of interacting with humans (we discuss moral and ethics issues below). A *task model* includes information relating to the robot's cognitive goals at a given time, information relating to progress toward those goals, information signifying an awareness of the robot's capabilities, and awareness of errors. An *analytical model* communicates the underlying analytical principles used by the robot to make decisions. For instance, knowing that a particular robot fuses information from satellite imagery and ground sensors in determining where potential emergency zones are located could be useful if the human knew the ground sensor network had been compromised (Lyons, 2013). This is similar to the algorithm transparency principal describes in the CFIT research. Finally, the *environmental model* provides the human with information about what the robot is experiencing.

The idea of transparency is not limited to embodied complex technologies such as robots. Organizations are now grappling with when and how to reuse complex computer code. The ability to reuse code has several advantages, two of which are of immediate benefit to an organization. First, as a pragmatic issue, the code is already paid for. The second factor relates more directly to our concern: code that is in use is *trusted* code. Alarcon et al. (2017) discuss issues relative to the reuse of code, and much of the argument could be viewed through the lens of transparency. The authors used interviews and cognitive task analysis to identify issues associated with the reuse of code. Several factors emerged. Firstly, the reputation of the code considering its source, reviews, and number of users—each of these is an indicator of

trust. Secondly, transparency—factors such as organization, style, architecture, and comments each influences how understandable the code is. Thirdly, performance of the code reflected in metrics for efficiency, resiliency, relevant functionality, flexibility, and being error free. Fourthly, environmental factors play a role. These include customer needs and requirements, organizational and resource constraints, and consequences of failure. Finally, Alarcon et al. believe individual differences such as propensity to trust and personality will also impact the decision to reuse code.

In summary, several models play a role in the transparency of a technological system and the human's ability to gain trust with the complex technology, be it a robot, AI system, or other type of advanced technology. To further facilitate the building of trust, one might consider providing a technology the ability to assess the human's state (fatigued, overloaded, frustrated, afraid, angry). An example is the case for monitoring a driver's state to allow technological intervention during instances of high fatigue (May & Baldwin, 2009).

If we agree transparency is important, then we must determine how best to introduce it into the system. One obvious place is via the interface, but a second is arguably more impactful and that is through extensive training. By utilizing effectual training, the operator gains an understanding of the models employed by the system, for example with the robot (intentional, task, analytical, environmental) or reuse of code (reputation, transparency, performance, environmental). We do not claim that all technological systems need to have these models; some may need more and others may function quite well with fewer, but for many technologies those systems seem to be an effective blend. Whatever the case, in order to build trust and maximize human-machine synergy it is necessary for the technological systems to be transparent in terms of their goals and operations.

Is the meaning and use of transparency in technological systems transparent? That is to say, there may be many types of transparency and it is important to identify those which are most effective with different technological systems. Lyons, Koltai et al. (2016) empirically examined two different types of transparency in terms of trusting and using an emergency landing planner (ELP) for commercial airliners. The traditional system developed by NASA provides standard information to pilots without explanation of the benefits (or costs) of alternative courses of action. These authors looked at transparency of explanation as impacting acceptance of a recommendation made by the ELP. They used a within-subjects design and considered two types of explanatory systems. The control condition used the standard ELP (e.g., weather, runway characteristics, and terrain). The first transparency system utilized risk-based transparency (referred to as the *value* condition). It gave all the information as the base condition but also provided a probability of success (e.g., 39% chance the flight crew will be able to successfully complete the approach and landing under current conditions). The second condition was based on logic-based transparency, whereby a statement was included to provide the logic behind the risk statement (e.g., the runway is unacceptable because the crosswind is too high for a safe landing). Results demonstrated trust in the system was the highest in the logic-transparency condition, followed by the risk-based transparency, and lowest in the control condition. These findings clearly support the premise that advanced technology will not be used without trust in the system. Here the output of a system is not accepted on a prima facia

basis. Rather, trust in the system is facilitated through an explanation of how the technology arrived at the recommendation. This explanation gets at the heart of the models employed by the technology and makes everything transparent to the user.

False Alarms—Inhibitors of Trust

There are many factors that impact trust among workers in an organization. For example, Galford and Drapeau (2003) describe how false feedback, inconsistent messaging, and inconsistent standards are among those factors negatively impacting trust in the workplace. Since many, if not most of these systems will be associated with workplace tasks, we must consider issues of workplace and interpersonal trust, as well as trust in automation. If a system is to be trusted it must, at a minimum, provide reliable information. Yet many systems have multiple technological components. For example, consider the cockpit of a modern aircraft. There are many individual gauges and displays (e.g., attitude, airspeed, altitude) providing information regarding the status of the aircraft. The same can be said for the dash in the driver's area of a car, providing information regarding the status of the car (e.g., speed, traction, oil pressure). A question arises as to how individuals establish a level of trust in such situations; is it based on each individual component sensor or display (e.g., separate trust levels for attitude, airspeed, and altitude) or is one level of trust established for the system as a whole? Gees-Blair, Rice, and Schwark (2013) examined this issue through the manipulation of false alarm rates. Their work indicates individuals form one impression of the system as a whole and increased levels of not just faulty, but also false alarms in one device/display will impact trust in the entire system. Thus, an individual may be unwilling to utilize a technology (e.g., fly the aircraft, ride in a self-driving car) because impressions of the unreliability of one component spreads (contaminates) the perception of the reliability of the system as a whole.

Once trust in a technology degrades or ceases due to real failures, degradation of performance, or false alarms, if the technology is to be used again we must deal with the process of trust repair. Just as humans must engage in a progression whereby trust between two individuals must be rebuilt due to some action or inaction on the part of another, so too must we consider trust repair in the context of human-machine teams. Technology will give false alarms, make poor recommendations, take inappropriate actions, and will fail to act when appropriate. Trust between humans and machines will need to be repaired and we should know how best to proceed. Much work needs to be done in this nascent area and those interested should see de Visser, Pak, and Shaw (2018) who argue we should use models of how relationships between two humans are repaired to guide us in the process of how human-technology trust repair should be fashioned. Another useful model is presented by Kim, Dirks, and Cooper (2009) who offer an interesting bilateral model of trust repair that accounts for perceptions and characteristics associated with both the human trustor and human trustee. Their approach to understanding trust negotiation efforts and trust repair methods are worth additional study and application. Furthermore, Kim et al. suggest that unfixable flaws in the person (e.g. their character) are hardest to repair. This suggests that the moral fiber and ethical decision making of an agent is the

greatest concern; integrity and benevolence is key to maintaining a healthy trust-based relationship and has implications for understanding morality and trust in cognitive agents (more on this later in the chapter).

We began this chapter introducing definitions of trust containing facets of vulnerability, action, cognition, and affect, among others. Trusting someone means putting ourselves at risk; it also means allowing someone or something to act on our behalf. For this to effectively occur the literature is clear that various aspects of the technology must be transparent to the user. This transparency will come in many types depending on the technology, and may be best presented to the user via the technology's interface or through training and experience with the technology. Just like humans, however, the technology will not be eternally omniscient and act flawlessly. Therefore, trust between the human and the technology will need repair. We need to develop theories and strategies associated with trust repair in human-machine systems. Of course, none of this can be done without reliable and valid means to measure and model trust.

We now move to examine humanness, as much of what we know about technology and how it will be trusted, accepted, and utilized will depend on the degree to which it is anthropomorphized. Following our discussion of humanness, we move to consider technology, especially AI technology as our "cognitive wingman."

TRUST, HUMANNESS, AND MORALITY

DAVE BOWMAN: Open the pod bay doors, HAL.
HAL: I'm sorry, Dave. I'm afraid I can't do that.

—2001: A Space Odyssey

Movies such as *The Terminator*, *I Robot*, *The Matrix*, and *2001: A Space Odyssey* have portrayed robots and non-human intelligent agents with ill intent. The basic premise of these films is that machine intelligence and awareness inevitably result in conflict and adversarial relationships with humans, often with lethal results for the human (de Visser et al., 2018; Snyder & Mcneese, 1987; Fraser, Hipel, Kilgour, McNeese & Snyder, 1989). More recent science fiction, however, paints a subtler relationship between humans and intelligent machines. In *Moon*, the robot GERTY is assigned to monitor the health and schedule of a human employee but hides communications and directives from the company. In *Robot & Frank*, a health provider robot prescribes meals and exercises that are (initially) rejected by Frank. In *Her*, the operating system Samantha exists to serve as a personal assistant, but instead pursues her own hidden goals and initiatives. And in the recent *I Am Mother*, a robot raises a child in a world absent of other people and hides the true extent of her master plan. In sum, the popular conception of non-human intelligent agents has shifted somewhat towards increased nuance, away from overt aggression and toward subtle conflicts of interest.

Recent work by futurists and philosophers have proposed the inevitable rise of the super intelligent machine, that is, a machine that creates its own machines which are more intelligent than any human being currently alive. These futurists say the super intelligent agent may be the last invention we ever have to create but warn that

it may be mankind's undoing. Others are more optimistic and propose that super or artificial intelligence will generally be a force for good to be embraced by everyone.

The possible creation of super intelligent machines poses an interesting problem for the scientific community, because current frameworks on machine intelligence primarily address automation, adaptive automation (Byrne & Parasuraman, 1996; Scerbo, 1996, 2008; Feigh, Dorneich, & Hayes, 2012), and more recently autonomy (Kaber, 2018a). The frameworks and theories are not equipped to deal with intelligence that is not based on human intelligence. Thus, such intelligent machines will have an impact on the types of frameworks we use to normally classify automation and technology (Kaber, 2018b; McNeese, 1986).

The Cognitive Agent Spectrum

Although the stories from the movies in the previous section are mainly for entertainment purposes, some of these imagined realities may not be far from resembling our own. As non-human agents have become better able to mimic human intelligence (or affect certain elements of human intelligence—the increasing social "charm" of Apple's Siri or Amazon's Alexa, for example), research is needed that accounts for the various roles that will be fulfilled by these agents. Specifically, greater understanding is required of the implications of automated aids with "ulterior motives," that is, those whose stated goal (e.g., informing an unbiased purchasing decision) might differ from the goal the user perceives the agent as having (e.g., guiding them toward a particular purchase). Improvements in technology and design will allow human-machine interactions to be more complex and nuanced. As the quality of these interactions becomes closer to human-human interactions, a greater understanding of the role of machine "humanness" should be pursued. Although the human-automation research literature has traditionally assumed a dichotomy between humans and automation (Madhavan & Wiegmann, 2007; Nass, Moon, Fogg, Reeves, & Dryer, 1995; Nass, Steuer, & Tauber, 1994), the increasing social and analytic capabilities of automation may soon necessitate the consideration of automation "humanness" along a continuum. There have been classifications of machines along a functional spectrum (Kaber & Endsley, 2004; Parasuraman, Sheridan, & Wickens, 2000) but we propose a *cognitive-agent spectrum*. Such a classification would arrange machines along a continuum of cognitive agent (cognitive independence): the degree to which a machine is perceived as initiating, executing, and controlling its own actions. Variables that determine the degree of independence ascribed to a machine may include complex factors such as the level and sophistication of interaction capabilities as well as simpler qualities like appearance (e.g., lines of code vs. an expressive computer avatar). With these novel agents varying on humanness, it becomes important to investigate how we perceive and trust these agents.

The Uncanny Valley

The Uncanny Valley, a theoretical effect proposed by Mori (1970), suggests that as human attributes are emulated by machine agents with increasing fidelity, subjective feelings of liking will increase until that emulation becomes near perfect, at which

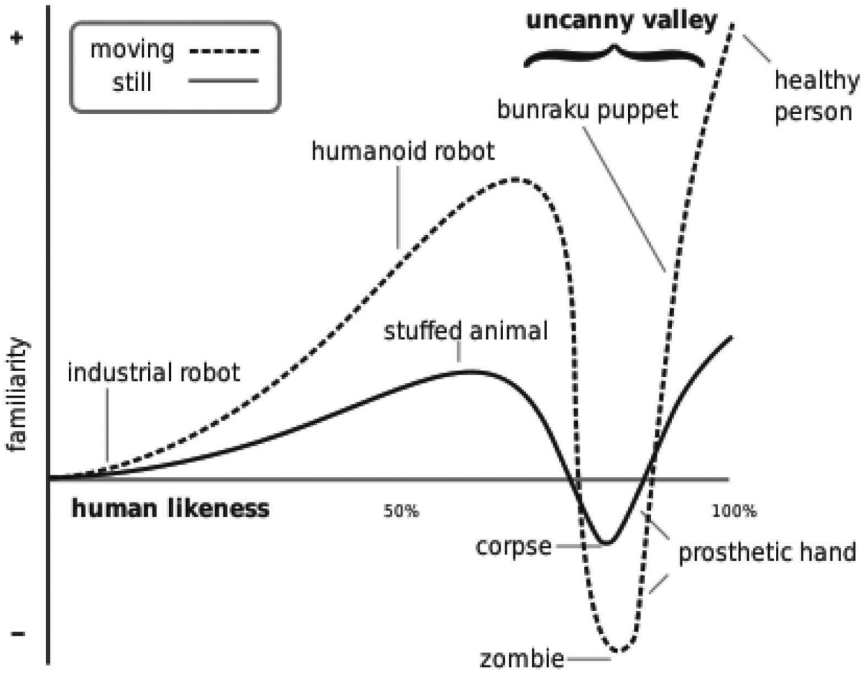

FIGURE 9.1 The Uncanny Valley.

Source: Figure adapted from MacDorman and Kageki (2012).

point the agents will evoke feelings of discomfort instead (see Figure 9.1). The sudden decrease in liking as robotic agents approach human appearance is the "valley." Since this theory was proposed, it has received a great amount of attention in the human-robot interaction community; if valid, the Uncanny Valley has direct implications for the design of the humanoid robots.

The empirical support for this theory remains mixed, however. Some research has elicited effects consistent with the valley using still images of humans and machines morphed together (MacDorman & Ishiguro, 2006) or using images of abnormal features (Seyama & Nagayama, 2007) or images of monkey faces (Steckenfinger & Ghazanfar, 2009) or human faces in decision support systems (Wellens, 1993). Research using videos of robots-in-motion, however, has typically failed to elicit an Uncanny Valley effect, which is somewhat surprising given that motion is theoretically supposed to augment the effect (C.-C. Ho, MacDorman, & Pramono, 2008; C. Ho & MacDorman, 2010; K.F. MacDorman, 2005; Saygin, Chaminade, Ishiguro, Driver, & Frith, 2011). Rather than causing a valley, the effect for increasing fidelity of humanoid movement seems to be a monotonic increase in perceived humanness and familiarity (Thompson, Trafton, & McKnight, 2011).

Research on the Uncanny Valley has thus far failed to provide a consistent account for the role of humanness in perception of humanoid agents. The majority of this

work, however, has focused on the visual features of the agent, such as its physical appearance or aspects related to its motion through space. Perhaps the uncanniness of the agents, rather than stemming from their physical appearance, is more attributable to the presence of certain internal features implied by the external illusion of humanness (Epley, Waytz, & Cacioppo, 2007). That is, robotic agents that closely approximate human appearance or movement (while remaining distinctly non-human) cause the human observer to assume the robot is capable of other human-like functions, like goal-directed behavior or emotion.

The Uncanny Valley and Morality

Recent work suggests that a critical factor for uncanniness may pertain to the perception of mind. Gray and Wegner (2012) found that when a computer expresses experiential features such as hunger and fear, it produces greater uncanniness in the mind of a human observer compared to a robot that simply expresses movement-related agency. The attribution of mind therefore might be more important than the perception of physical movement for eliciting an effect consistent with the Uncanny Valley. Indeed, perceived internal experience (consciousness) may be the driving force behind the Uncanny Valley effect.

If the perception of internal processes and preferences induces feelings of unease among human observers, it seems likely that robotic agents expressing a particular moral preference would be similarly disturbing. A recent review of morality and theory of mind suggests that moral preference and perceptions of intent are difficult to disentangle (Gray, Young, & Waytz, 2012). Although no research has yet examined user perceptions of a morally dubious robot, some research has begun to explore the idea of moral accountability for robotic aids. Using a monetary incentive, Kahn et al. (2012) found that participants held a robot partner morally accountable when the robot failed to perform in the incentivized task. Ratings after the experiment showed that most participants rated the robot as morally accountable for the prizes that it caused the participant to lose. The fact that participants held the robotic aid morally accountable begs some comparisons to how they might have acted with a human partner. Although participants rated the robot's social and experiential features as being significantly less prominent than those of a human, the prominence of these features was rated as being greater than those of a simple vending machine. Although Kahn et al. (2012) did not test this premise, we expect that participants would not have held a vending machine morally accountable for its malfunctions. Thus, the potential distinction between various non-human machines in terms of their perceived moral responsibility hints at the presence of a continuum. What remains unclear, however, are the factors that might influence the degree to which individuals or teams of people perceive various cognitive agents as being morally culpable.

Morality and Trust in Cognitive Agents

Tightly linked to the idea of morality in agents is the matter of trust. Trust is informed by how we expect an entity might act towards us in the future. In fact, perceptions of

moral character have been shown to directly impact levels of trust. Delgado, Frank, and Phelps (2005) found that participants who read written biographies describing a morally "bad" partner were less likely to rely on that partner in a subsequent economic trust game. In contrast, levels of trust for morally "good" partners were significantly elevated over the "bad" and "neutral" ones. Participants also shared more of their winnings with the good partner compared to the bad partner. This experiment clearly supports the idea that moral character influences trust and subsequent cooperation-related decision making.

Research has not yet examined how moral character influences the perception of humanness for non-human agents, and how these perceptions might affect trust and ratings of liking. Some research suggests that human-human interactions are directly comparable to human-machine interactions, and that differences in trusting are only quantitatively (rather than qualitatively) different (Nass et al., 1994). Consequently, this research proposes that automation can elicit social interactions that are similar in quality to the interactions occurring between humans. In support of this view, a recent study did not find any differences in trustworthiness of a medical robot tasked with taking a blood pressure reading when it had a silver face, a human face, or no face at all (Broadbent et al., 2013). In direct opposition to this media equation hypothesis, some researchers propose that trust between humans is *qualitatively* different than trust between humans and machines (de Visser et al., 2012; Dzindolet, Peterson, & Pomranky, 2003; Lee & See, 2004; Madhavan & Wiegmann, 2007). This automation theory has also found empirical support: when playing the ultimatum game, people are much more willing to accept unfair offers from a computer than they are from a human, perhaps due to a stronger emotional response towards the humans (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). It would therefore be useful to examine this issue further incorporating a theoretical framework such as McAllister's (1995) since his incorporates two different types of trust, cognitive and affective. That theoretical perspective would predict differential level of cognitive and affective trust toward the targets, human vs. technology.

Morality with Super Intelligent Agents

Researchers have proposed the possible advent of the ultra-intelligent machine (Good, 1965), a machine that can far surpass all the intellectual activities of any human, no matter how clever. If this machine could design even smarter machines, then this is the last invention man would have to make provided the machine is docile enough to tell us how to keep it under control (Good, 1965). Recent reports have speculated about when mankind may invent such a machine with dates ranging from early to late 21st century (Bostrom & Yudkowsky, 2014; Kurzweil, 2005). This also raises the possibility that we would attribute a mind to such a machine and that we would also imbue it with a sense of morality, perhaps even a superior morality to our own. It is doubtful people will accept the morality of a machine if it means that they have to give up control. Every movie about the future, at least in the West, plays with the theme of fear as a consequence of losing control to a machine. Current debates focusing on the use and policy of autonomous weapons raise a lot of emotions for this reason. Yet, it is not unthinkable that a machine, in some cases, could act with

superior morality compared to man. A lot of our personal morality is gut-based, whereas society's morality is rule and law based. There is an opportunity here to create machines that have human values encoded into them and act as moral agents of humanity. In this sense, we are in the unique position of defining our morality more precisely and having it more faithfully executed by machines. Design guidelines on how to implement these kinds of moral rules should be a primary focus of those in the human factors and broader scientific community. If successful guidelines are created, and healthy human-machine trust relationships are established, we can then fully apply artificial intelligence capabilities to enhance distributed cognitions and effectiveness among teams.

AI: OUR "COGNITIVE WINGMAN"

As outlined in the beginning of this chapter, the importance of cognitive-based trust (McAllister, 1995) and trustworthiness is simply indisputable (Coovert et al., 2017) and it is difficult to overstate the positive influence these constructs have on effectiveness of distributed teams. Although we consider trust and trustworthiness baseline requirements for performance, exploring elements beyond these constructs is vital to harnessing the full potential of each team. Given the current state of technology, there is a unique opportunity to embrace AI and the projected benefits may very well outweigh the risks. Leveraging AI as a cognitive "wingman" and maximizing trust among human-machine teammates can lead to numerous applied solutions relevant to distributed team cognition.

We do recognize that there is a plethora of dire, fear-mongering predictions regarding the use of AI, such as those proposed by Elon Musk and Stephen Hawking (Clifford, 2018). Although we agree that warnings against blindly trusting intelligent agents should not be ignored, strategies built on risk avoidance can prove costly and hinder efforts to maintain a competitive advantage. Unquestionably, there may be a danger in AI running wild. However, the dichotomous thinking that leaves the human teammate solely responsible for all critical and ethical decisions not only leads to unnecessary constraints, but also assumes the human teammate is always superior. Research suggests that 80% of the time that humans fail to take moral action, they actually recognize and understand the more appropriate action to take; these findings suggest a moral judgment-action gap regarding human intentions (Sweeney, Imboden, & Hannah, 2015) . While many humans are aware of their poor moral choices, sometimes they fail to even recognize the moral implications. Humans have limited attentional resources (Kahneman, 1973) and various endurance constraints. Thus, their inability to consistently devote time and awareness to ethical considerations is also a distinct concern; sometimes we simply miss the problem altogether. With so many examples of human teammates actively choosing wrong over right (e.g. Enron, Bill Clinton, David Petraeus), and the inherent limitations of human processing, it is important to challenge the notion of human superiority and consider how, or when, machines may prove optimal. Our overarching argument is that a hybrid human-AI team, when properly configured, will result in superior outcomes. Teams and individuals can achieve these outcomes by either acting independently or as homogeneous teams of each acting without the other.

AI Capabilities Impacting Distributed Team Cognition

Team Data Processing Capabilities

AI can increase the volume, velocity, and consistency of data processing to offer teams an unprecedented cognitive power that improves their information-to-decision capabilities. The main advantage to AI systems is their ability to access and process large volumes of diverse, digital information (Sycara & Lewis, 2004) and to make predictions at unprecedented speed. In recent displays of advanced AI capabilities, such as IBM's Project Debater (Teich, 2018) and Google's Automated Assistant (DeMers, 2018), the latest intelligent systems can sift through vast data repositories to swiftly detect pertinent information, rapidly manage knowledge, and transmit logical arguments and decisions commensurate to their human creators. The ever-increasing volume and velocity of information that is detected, stored, and manipulated through machine learning offers unbounded opportunities to optimize team functionality. When seated with the right communication links, the AI teammate can disperse these cognitive capabilities and optimize team performance from nearly anywhere in the world.

Team Communication Capabilities

Within the realm of distributed team cognition, AI can increase effectiveness by offering a broader, more comprehensive repertoire of cognitive skills. Combining these skills with advanced communication functions can help optimize critical team cognitions, including an increase in shared mental models (Mohammed & Dumville, 2001) and a more robust transactive memory (Wegner, 1987). Paired with human teammates, the rapid and expansive information processing capability of AI systems can help share vital, real-time information to the team while simultaneously serving as a gargantuan, on-demand repository of historical knowledge. Beyond knowledge sharing, AI systems can passively monitor and record team processes or actively push alerts and goal-relevant status updates. This type of information flow can help enrich cue strategies (Salas, Rosen, Burke, & Goodwin, 2009) across a distributed team and ensure a near real-time dissemination of mission information (Salas et al., 2009). These enhanced functions, coupled with the machine-learning features integral to AI systems, offer a powerful situational awareness capability that can be rapidly and consistently shared to all team members. Mitigating the cognitive processing limitations of their human teammates, the AI "cognitive wingman" can also detect and analyze vast amounts of available stimuli to maintain and optimize team shared awareness and simultaneously transmit relevant predictions to all teammates.

Team Situational Awareness Capabilities

AI systems are capable of gathering information at lower levels, while performing information-processing activities to support team coordination and team-level situational awareness (Cooke, Salas, Kiekel, & Bell, 2004). Situational awareness is really about possessing real-time knowledge of what is happening around you (Endsley, 2000; Parasuraman, Sheridan, & Wickens, 2008), or more precisely, "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their future status" (Endsley,

1988, p. 97). Endsley (1988) offered three levels of situational awareness: perception, comprehension, and prediction. When furnished with the right sensory capabilities, AI systems can be far more efficient than humans across all three levels to dramatically improve team behavior efficiencies. Specifically, AI can support rapid and thorough situational assessments, continuous situational monitoring, and alerting (Sycara & Lewis, 2004).

We do not expect all AI interpretations to be correct of course, as 100% reliability of autonomous agents cannot be guaranteed (de Visser & Parasuraman, 2011; de Visser et al., 2018). However, we can learn to counter the effects of imperfect performance of AI systems that, compared to human counterparts, have greater behavioral consistencies. Furthermore, we believe imperfections in machine learning are not the real concern. The actual issue is not whether autonomous agents are fallible, but whether or not human teammates are *less* fallible. The gaps in human-machine comparisons are perhaps most salient when evaluating capabilities to detect and interpret stimuli, retrieve and manipulate stored information, or project outcomes to aid in team decision making.

Additionally, researchers have noted that when human and machines work together, substantial performance gains often follow (de Visser & Parasuraman, 2011). Those that dogmatically argue in favor of the human may miss important opportunities to leverage our cognitive "wingman" to increase the consistency, speed, and volume of team situational awareness and improve information-to-decision capabilities. Since exploitation of AI systems can deliver important benefits to distributed team cognition, perhaps it is time to extend the repeated ideology of *keeping the man in the loop* by adding the notion of *keeping machines in the loop.*

In sum, we believe in leveraging AI capabilities to expand the volume of knowledge available to each teammate; improve the velocity of information transmitted throughout the team; and ensure consistency of team processes to vastly improve distributed team cognition. These advanced cognitive capabilities can improve a wide range of team processes, behaviors, and attitudes (see Table 9.1). We argue trustworthy AI systems are uniquely suited to deliver these capability requirements and improve the information-to-decision processes within teams. Specifically, we recommend developing and exploiting AI to deliver the following applied solutions: *enhance critical thinking and adaptive processes*; *improve ethical decision making within teams*; and ultimately, allow teams to *safely exercise disciplined initiative*.

Applied Solution 1: Exploit AI as a Cognitive Support System to Enhance Critical Thinking and Adaptive Processes in Teams

Benefits to Team Critical Thinking

Leveraging AI capabilities as a cognitive support system not only enhances situational awareness but can also help guide critical thinking in teams. It is well known that human thinking is often biased and distorted, yet our productivity and efficiency depend on the very quality of our thought processes (Elder & Paul, 2008). After building a list of common errors[1] that human teammates often make when constructing and evaluating information and arguments, Gerras (2008) argued that we need

TABLE 9.1
AI Capabilities and Potential to Improve Distributed Team Cognition

AI Enhanced Capabilities

Volume of Information:
- Knowledge repository
- Knowledge management
- Knowledge transmission

Velocity of Information processing:
- Detection (sensory)
- Manipulating
- Storing and classifying
- Retrieving

Awareness Levels (Endsley, 1988):
- Perception
- Comprehension
- Projection/prediction

Human Error Mitigation:
- Avoid endurance constraints
- Cognitive load
- Fatigue
- Mood
- Detect and mitigate bias
- Cultural insensitivity
- Cognitive/affective bias
- Moral intensity perceptions
- Resist social influence
- Groupthink
- Group polarization
- Conformity
- Unethical tendencies
- Machiavellianism
- Need for dominance
- Personality disorders

Improved Processes, Behaviors, and Attitudes

Temporal Processes (Marks, Mathieu, & Zaccaro, 2001):
- Transition phase
- Action phase
- Interpersonal phase

Common Behaviors (Salas et al., 2009):
- Communication (closed-loop, feedback)
- Coordination
- Cooperation (monitoring, backup)

Team Attitudes (Marks et al., 2001;Salas et al., 2009):
- Efficacy/potency
- Emotional regulation
- Trust
- Goal commitment

Team Cognition (Salas et al., 2009):
- Cue strategies
- Problem solving
- Mission information
- Shared mental models
- Transactive memory
- Team situational awareness

Team Adaptation:
- Adaptation phases (Burke, Coovert et al., 2006)
- Situational assessment
- Plan formulation
- Plan execution
- Team learning
- The *Four Rs* (Frick, Fletcher, Ramsay, & Bedwell, 2018)
  - Recognize
  - Reframe
  - Respond
  - Reflect

to think more critically about critical thinking itself. Human thought and decision making can suffer from limited cognitive load capacities and fatigue. These human endurance constraints can also be exacerbated through varying shifts in mood, leading to improper judgments or even a general unwillingness to engage in critical thought altogether. Even when the human processing systems are fresh and attentive, cognitive and affective bias can distort perceptions, judgments, and interpretations of environmental stimulus. Furthermore, teams are highly sensitive to social influence and are susceptible to the effects of groupthink (Janis, 1972), group polarization (Myers & Lamm, 1976), and pressures to conform (Milgram, 1974).

AI, in the form our "cognitive wingman." AI can help mitigate these human endurance constraints because they are not susceptible to fatigue or dramatic mood

swings. Cognitive activities, typically managed by human teammates, can be offloaded to AI support systems (Cooke et al., 2004) to relieve cognitive pressures across the team. These cognitive support systems are always ready and available to provide a full set of cognitive behaviors to back up, monitor, and reinforce their human teammates. When programmed correctly, they may prove less vulnerable to common biases that contaminate human thought, such as availability and representative heuristics; sample size and regression to the mean bias; overconfidence and arrogance; and finally, confirmation bias and fundamental attribution errors (Gerras, 2008). Although it is possible that human bias can contaminate the algorithms and data inputs necessary to run AI, these cognitive support systems are not directly influenced by social pressures. AI developers could also mitigate bias contamination by building systems that avoid the pitfalls of human cognition. With the right design, AI can guide teams to "think critically about critical thinking" (Gerras, 2008) by performing sensitivity analyses to account for the potential influence of perception bias, variations in cultural norms, and other sources for errors in thought.

Benefits to Adaptive Team Processes

Including AI in teams to improve situational awareness and enhance critical thinking can also help optimize team adaptation skills. Typically, critical thinking and situational awareness alone are not enough to deliver efficient team performance, as team adaptation skills are the crucial characteristics of effective teams (Frick et al., 2018). However, these two cognitive processes are important for improving a team's ability to navigate the team adaptation phases of situational assessment, plan formulation, plan execution, and team learning (Burke, Stagl, Salas, Pierce, & Kendall, 2006). Armed with the impressive volume of knowledge and velocity of information delivered by AI support systems, teams can now use advances in data processing and prediction to avoid the haphazard planning and execution often employed by maladaptive teams. When viewed through the lens of Frick et al.'s (2018) Four R heuristic of *Recognize*, *Reframe*, *Respond*, *and Reflect*, AI systems can provide the situational awareness and critical thinking support to help teams better engage in the following cognitive tasks:

- *Recognize*: Gain information and transmit team-relevant knowledge to ensure all teammates fully understand the current situation and share the same predictions of the future operating environment. Evaluate internal and external cues to identify the core problem that is driving the need for team adaptation.
- *Reframe*: Visualize the desired conditions and identify the capability shortfalls the team can resolve through adaptation. Propose multiple adaptation options, assess available resources, evaluate potential effectiveness, perform sensitivity analyses, and check for bias. Set and disseminate goals and new team roles to build shared mental models.
- *Respond*: Shape the environment and adjust the team to meet the new demands and execute team tasks as necessary to achieve the desired end state. Monitor performance, provide back-up behaviors, and communicate

effectively. Cooperate and coordinate with external teams to synchronize actions.

- *Reflect*: Evaluate the effectiveness of the team adaptation and solidify or refine the new team processes as appropriate.

Applied Solution 2: Apply AI as an Ethical Cognitive Support[2] Tool (Mathieson, 2007)

An Ethical Cognitive Support Tool

Given human temptations to indulge in amoral pursuits (Ludwig & Longenecker, 1993) and our vulnerability to subconscious cognitive bias, perhaps we should use caution when depending on the human as the sole ethical decision maker. Specifically, we should seek to determine which teammate (human or machine) is *more* dangerous and how we might leverage AI to reduce the risk or error involved with human judgment. Applying AI as an ethical cognitive support tool (Mathieson, 2007) has the potential to vastly improve the ethical conduct and effectiveness of teams. Although many might question the ethical implications of AI utilization, there is less discussion about how AI could actually improve organizational ethics within the realms of leadership, climates and cultures, and decision making.

Ethical leadership is believed to produce a trickle-down effect in organizations, as an ethical (or unethical) culture can originate with leaders and cascade directly onto followers within a team (Schaubroeck et al., 2012). These leaders can also indirectly influence the ethical culture of teams across various hierarchical levels in an organization (Schaubroeck et al., 2012). Brown, Treviño, and Harrison (2005) describe ethical leadership as normatively appropriate behavior which can be promoted to followers through two-way communication, reinforcement, and decision making. AI can serve as an important partner in promoting this normatively appropriate behavior for teammates to emulate, especially through gained efficiencies in the communication of ethical codes and the selection of ethical decisions.

Leaders and members of distributed teams often occupy boundary-spanning positions (Drescher, Korsgaard, Welpe, Picot, & Wigand, 2014; Druskat & Wheeler, 2003), and therefore are more likely to encounter ethical dilemmas and ambiguity (Brown et al., 2005) as they coordinate with external teams and encounter unique environmental stimulus. AI support systems could help guide teams through ethical ambiguity by generating various viewpoints, recognizing and transmitting value differences, interpreting ethical codes from different cultures, and predicting unintended consequences of ethical decisions. In fact, most research shows a code of ethics, along with positive ethical cultures and climates, improves ethical decision making (O'Fallon & Butterfield, 2005). Thus, AI support systems can help transmit an organization's code of ethics across distributed teams, while promulgating and reinforcing the normatively appropriate behavior within each teammate's unique contextual environment. This constant awareness of culture and norms can lead to greater ethical team outcomes, as our "cognitive wingman" can ensure teams have the body of knowledge necessary to make informed ethical decisions. If developed

TABLE 9.2
Factors Influencing Perceptions of Moral Intensity

Six Moral Intensity Factors	Description
(1) Magnitude of consequences	Agreement on importance
(2) Social consensus	Event probability × likely effect
(3) Probability of effect	Time between the decision and its consequences
(4) Temporal immediacy	Nearness among decision makers and those affected
(5) Proximity	Intensity of impact by group size
(6) Concentration of effect	Amount of benefit (harm)

Source: Jones, 1991

properly, AI systems could even provide greater insight into implicit perceptions regarding the moral intensity of sensitive issues.

Calculating Moral Intensity Perceptions

When designing and implementing an AI ethical cognitive support system, it is important to recognize and accommodate the six factors (see Table 9.2) that can influence human perceptions regarding moral intensity (Jones, 1991) and whether a decision actually has ethical importance (Mathieson, 2007). Research shows strong support for the idea that perceptions of moral intensity can influence ethical decision making (O'Fallon & Butterfield, 2005). Thus, building AI systems to predict moral intensity perceptions could help guide teams through ethical decision-making processes. By producing and sharing the confidence intervals of these predictions, AI can arm teams with the moral intensity risk associated with various ethical decision-making options. Specifically, our "cognitive wingman" could calculate the probability of effects of various decisions and predict consequences regarding the amount of harm; the degree of social consensus on importance; the amount of time and proximity for an anticipated effect; and even the group size impacted. Given this new depth of information, teams can better navigate basic components of moral decision making (Rest, 1986), which includes the moral nature of issues, moral judgments, moral intent, and moral actions. If AI can help teams take a more deliberate approach in making informed ethical decisions and create a shared awareness, perhaps it is also time build these systems to exercise some basic initiative on behalf of the team.

Applied Solution 3: Deliberately Leveraging AI to Safely Exercise Disciplined Initiative

The Importance of Disciplined Initiative

Deliberately leveraging AI to safely exercise disciplined initiative and accept prudent risk is a critical consideration, especially to the functionality and effectiveness of distributed and crisis action teams. Within a military context, decentralized execution, or what is now often referred to as mission command doctrine

(Department of the Army, 2019), is an important concept for meeting the demands of modern warfare. The mission command doctrine is the US Army's answer to directing and controlling teams, while appropriately empowering decision making and decentralized execution at lower levels. Perhaps now more than ever, war is extremely complex, rapidly changing, and always uncertain (McMaster, 2015). Therefore, a dogmatic approach of centralized control will likely remain too inflexible to capitalize on unforeseen opportunities or to allow teams to quickly adapt during moments of ambiguity. Ever since the German army first began exploring this concept in the late 1800s, the ideas of decentralized leadership and empowering subordinates to take initiative (Matzenbacher, 2018) have repeatedly been proven beneficial to team performance on the battlefield. Now more than ever, modern warfare and the rapid pace of conflict does not allow subordinates to wait for updated orders. Predicated on mutual trust, commanders on the battlefield provide intent and then grant subordinates the flexibility to take critical and decisive action to meet that intent.

Interestingly, the benefits and need for decentralized control is not unique to military organizations. Distributed teams and remote collaboration are becoming increasingly common in our modern work environment. Many teams now work together across vast distances, but often at the same time over a shared visual workspace (Gutwin & Greenberg, 2004). These remote team designs must share a common workplace awareness about the environment and how each physical workspace might change over time. Similar to a military context, these professional work groups will require near real-time awareness of various teammates and how they are interacting with and continuously adapting to their distinct physical workspace. Unfortunately, it is extremely difficult for distributed teams to maintain this necessary awareness (Gutwin & Greenberg, 2004) and to effectively execute team coordination. These limitations lead to process loss and decrements to productivity (Steiner, 1972), along with shortfalls in implicit and explicit team communication (Fiore & Salas, 2004). These gaps can produce detrimental misunderstandings that wreak havoc on distributed team cognition, leaving remote and crisis action teams with a limited ability to adapt and execute decentralized operations or tasks.

Enabling Decentralized Execution

Keeping our cognitive teammate *in the loop* can vastly improve the shared awareness needed for both military teams and professional work groups. Since AI can rapidly transmit information across the team and avoid human endurance constraints, it can continuously monitor and disseminate real-time information relevant to the physical and digital environment. This can increase awareness levels and mitigate the communication and coordination gaps that prevent effective and decentralized execution among distributed teams. Organizations properly enabled with AI can have faster access to mission orders (or goal-based tasks) and an enhanced ability to accept measurable and prudent risk. With greater team situational awareness and rapid information-to-decision capability, teams can develop the cohesion and collective efficacy needed to build mutual trust. AI can help create a shared understanding, disseminate clear intent and evolving goals throughout the team, and calculate

how to safely practice decentralized execution (Department of the Army, 2019) of important tasks or missions.

CONCLUSIONS

The goal of our chapter has been to contemplate issues related to the adoption of truly helpful and powerful technologies. These technologies will likely be empowered by AI. The history of technological development is littered with failed products that were produced and, had they been adopted, would certainly have helped society. There are many reasons technologies are not embraced—poor human factors design is likely at the front of that list. Yet, as sociotechnical systems theory so clearly demonstrates, we need to consider human and social issues in addition to technological ones. As such, our chapter has focused not on algorithms and hardware, but rather on issues associated with the human side of the equation. We examined trust and specific issues related to trust in advanced technologies. Following trust, we interwove humanness and the role it plays with the anthropomorphism of software and embodied technologies. Embodied humanness predicts a linear to nonlinear function describing the Uncanny Valley, where increasing degrees of preference for humanness are suddenly replaced by dislike, when the technology becomes *too* human. The third section of our chapter poses our hybrid perspective on when technology will become supremely useful and most adopted. Colloquially stated, our premise is one where we understand the strength and limitations of human cognitions, as well as the strengths and limitations of advanced AI technologies; so, let us combine the two in such a fashion the technology acts as a "cognitive wingman." By providing this type of support, AI can compensate for the bounds on and biases of human cognitive processing.

In developing a premise or theory as to why technology is adopted we need to specify constructs linked in a nomological network encompassing features of the technology, human, and task environment in which the technology is to be deployed. With constructs, such as trust, identified we must then focus on measurement issues to operationalize the premise/theory and begin to examine and test linkages among and between constructs. With a measurement system in place, such as that provided by latent change scores, we have a solid approach to measure trust and how it changes in response to specific aspects of the technology. For example, how much does transparency lead to increased levels of trust? We can explicitly examine the changing levels of trust in the human as we change levels and types of transparency in the technology. Concomitantly, when trust is violated as in the case of false alarms or when the system simply errs, latent change scores will reflect the decrement in trust as a function of those alarms or errors. Finally, we can use the same methodology to evaluate the effectiveness of alternative trust repair strategies.

The issue of appearance on a dimension of humanness for cognitive agents is an important one. We have, essentially, a balancing act where we want greater similarity to humans than that which is presented by the glowing red light of HAL in *2001: A Space Odyssey*; yet many find human-appearing robots such as Erica (Collins, September 13, 2018) creepy and well into the Uncanny Valley. Finding this sweet spot is an important issue for the acceptance of technologies. It is perhaps even more

complicated a problem than it may appear on the surface as much research in psychology has demonstrated appearance is correlated with attributions of other characteristics as well. For example, Arbogast (2018) recently demonstrated a significant relationship between appearance of facial features (in both men and women) and an attribution of toxic leadership. In addition to these main effects, individual differences also likely play a role in preferences. As such, it may be most effective to allow individuals to select or even configure the appearance of their personal cognitive wingman in order to maximize acceptance.

As stated above, we believe the cognitive wingman offers the best of both the human world and the technological/AI world. Humans are limited by bounds on their perceptual and cognitive processing systems. We are further limited by faulty heuristics and biases. If we construct AI technologies in such a fashion to augment cognition and reduce cognitive and perceptual limitations while simultaneously overcome biases, the combined human-AI technology team could be quite potent and effective. Even in the relatively benign realm of food processing, providing robots the ability to see has led to a 100% increase in the amount of food processed and other areas have seen a million-fold improvement (Kahn, 2018). To keep these advances coming we must further develop our theories of why technologies are adopted. We have made solid strides in this area by keeping trust as a core construct and examining factors that influence it. For example, minimizing false alarm rates, transparency (of various types), ability to configure critical aspects of the system according to individual difference preferences, and effective strategies for trust repair are key aspects that must be kept in mind. We also need to attend to additional ways to signal trust. This can be done via interface development and the oh so important training programs that can be used to make explicit those factors discussed in this chapter.

What is some of the low-hanging fruit we might consider in the human-cognitive wingman approach? The world continues to change rapidly and we are prone to change blindness and loss of situational awareness. A cognitive wingman could be most helpful in these areas by bringing the problem back into focus for the human. As described previously, critical thinking is one of the most important aspects of cognition done well. A cognitive wingman-human team can provide results superior than can be achieved by either alone. Of course, whenever teamwork is involved it is critical to monitor and maximize team functioning to ensure process gain over process loss. A well-constructed cognitive wingman, in addition to providing task-specific knowledge and expertise, can also monitor the process between members of the team, including the AI members and, when appropriate, act as facilitators ensuring effective team process.

In closing, let is point out the essential importance of ethical AI and how it may ensure ethical behavior among humans as well. The essence of ethical behavior in our society is not always clearly defined; there are many gray areas. Yet in order to construct an ethical AI system we need to really grapple with what it means to be an ethical human. Doing this hard work of defining the criterion space of the construct, providing examples for training so the system can have codified knowledge, and developing a learning system which itself understands ethical behavior, are minimally three components that would serve as the kernel to such an ethical AI system. Grappling with the development of these three components would force us

to profoundly consider what it means to be ethical in our society today. This process will help us manage the current ethical problems faced by humans and will ensure ethical behavior from AI in the role it plays as our cognitive wingman.

NOTES

1. The nine most common errors reflecting weak critical thinking are (Gerras, 2008): arguments against the person, false dichotomies, appeals to unqualified authorities, false causes, appeals to fear, appeals to the masses, slippery slope arguments, weak analogies, and red herrings.
2. Early discussions of AI envisioned its applicability for decision support. Our perspective is the utility of AI is broader, encompassing many diverse cognitive activities. As such we employ the term "cognitive support."

REFERENCES

Alarcon, G. M., et al. (2017). A descriptive model of computer code trustworthiness. *Journal of Cognitive Engineering and Decision Making*, *11*, 107–121.

Arbogast, M. S. (2018). *Egos gone wild: Threat detection and the domains indicative of toxic leadership*. Unpublished dissertation. Tampa, FL: University of South Florida Department of Psychology.

Army, U. S. (2019). *Army Doctrine Publication (ADP) No. 6–0, mission command: Command and control of army forces*. Washington, DC: US Department of Defense.

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, *29*(1), 94–103.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence*, *1*, 316–334.

Broadbent, E., Kumar, V., Li, X., Sollers, J., Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. *PLoS One*, *8*(8), e72589. https://doi.org/10.1371/journal.pone.0072589

Brown, M. E., Treviño, L. K., & Harrison, D. A. (2005). Ethical leadership: A social learning perspective for construct development and testing. *Organizational Behavior and Human Decision Processes*, *97*(2), 117–134.

Burke, C. S., Stagl, K. C., Salas, E., Pierce, L., & Kendall, D. (2006). Understanding team adaptation: A conceptual analysis and model. *Journal of Applied Psychology*, *91*(6), 1189–1207.

Burke, J., Coovert, M. D., Murphy, R., Riley, J., & Rodgers, E. (2006). Human-robot factors: Robots in the workplace. *Proceedings of the annual meeting of the Human Factors and Ergonomics Society* (pp. 870–874). San Francisco, CA: Sage.

Byrne, E. A., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological psychology*, *42*(3), 249–268.

Clifford, C. (2018). *Steve Wozniak explains why he used to agree with Elon Musk, Stephen Hawking on A.I.—but now he doesn't*. Retrieved from www.cnbc.com/2018/02/23/steve-wozniak-doesnt-agree-with-elon-musk-stephen-hawking-on-a-i.html

Collins, P. (2018, September 13). Creepy Japanese robot "with a soul" will replace a human news anchor. *The Daily Good*. Retrieved from www.good.is/articles/robot-soul-anchor.

Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. In E. Salas & S. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 83–107). Washington, DC: APA Books.

Coovert, M. D., Lee, T., Shindev, I., & Sun, Yu. (2014). Spatial augmented reality as a method for a mobile robot to communicate intended movement. *Computers in Human Behavior*, *34*, 241–248.

Coovert, M. D., Miller, E. E. P., & Bennett, W. (2017). Assessing trust and effectiveness in virtual teams: Latent growth curve and latent change score models. *Social Sciences*, *6*(3), 87.

Delgado, M., Frank, R., & Phelps, E. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*, 1611–1618.

DeMers, J. (2018). Google assistant is getting better: Here's what that means for marketers. *Forbes.* Retrieved from www.forbes.com/sites/jaysondemers/2018/05/29/google-assistant-is-getting-better-heres-what-that-means-for-marketers/#20547fe8616b

de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012). The world is not enough: Trust in cognitive agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*, 263–267. https://doi.org/10.1177/1071181312561062

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From "automation" to "autonomy": The importance of trust repair in human–machine interactions. *Ergonomcs*, *61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725.

de Visser, E. J., & Parasuraman, R. (2011). Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209–231.

Drescher, M. A., Korsgaard, M. A., Welpe, I. M., Picot, A., & Wigand, R. T. (2014). The dynamics of shared leadership: Building trust and enhancing performance. *Journal of Applied Psychology*, *99*(5), 771.

Druskat, V. U., & Wheeler, J. V. (2003). Managing from the boundary: The effective leadership of self-managing work teams. *Academy of Management Journal*, *46*(4), 435–457.

Dzindolet, M., Peterson, S., & Pomranky, R. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

Elder, L., & Paul, R. (2008). Critical thinking: The nuts and bolts of education. *Optometric Education*, *33*(3).

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting* (Vol. 32, No. 2, pp. 97–101). Los Angeles, CA: SAGE Publications.

Endsley, M. R. (2000). Theoretical underpinnings of situation awareness. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness and measurement* (pp. 3–32). Boca Raton, FL: CRC Press.

Epley, N., Waytz, A., & Cacioppo, J. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*, 864–886.

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors*, *54*(6), 1008–1024.

Fiore, S. M., & Salas, E. E. (2004). Why we need team cognition. In E. E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 235–248). Washington, DC: American Psychological Association.

Fraser, N. M., Hipel, K. W., Kilgour, D. M., McNeese, M. D., & Snyder, D. E. (1989). An architecture for integrating expert systems. *Decision Support Systems*, *5*(3), 263–276.

Frick, S. E., Fletcher, K. A., Ramsay, P. S., & Bedwell, W. L. (2018). Understanding team maladaptation through the lens of the four R's of adaptation. *Human Resource Management Review*, *28*(4), 411–422.

Galford, R. M., & Drapeau, A. S. (2003). The enemies of trust. *Harvard Business Review*, *81*(2), 85–95.

Gees-Blair, K., Rice, S., & Schwark, J. (2013). Using system-wide trust theory to reveal the contagion effects of automation false alarms and misses on compliance and reliance in a simulated aviation task. *International Journal of Aviation Psychology*, *23*, 245–266.

Gerras, S. J. (2008). Thinking critically about critical thinking: A fundamental guide for strategic leaders. *Carlisle, Pennsylvania: US Army War College*, *9*.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. *Advances in Computers*, *6*(99), 31–83.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Gutwin, C., & Greenberg, S. (2004). The importance of awareness for team cognition in distributed collaboration. In E. E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 177–201). Washington, DC: American Psychological Association.

Ho, C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0747563210001536

Ho, C.-C., MacDorman, K. F., & Pramono, Z. A. D. D. (2008). Human emotion and the uncanny valley. In *Proceedings of the 3rd international conference on Human robot interaction—HRI '08* (p. 169). New York: ACM Press. https://doi.org/10.1145/1349822.1349845

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*. Boston, MA: Houghton Mifflin.

Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of Management Review*, *16*(2), 366–395.

Kaber, D. B. (2018a). A conceptual framework of autonomous and automated agents. *Theoretical Issues in Ergonomics Science*, *19*(4), 406–430.

Kaber, D. B. (2018b). Issues in human—automation interaction modeling: Presumptive aspects of frameworks of types and levels of automation. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 7–24.

Kaber, D. B., & Endsley, M. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113–153.

Kahn, N. (2018, September 12). Seefood: Why robots are producing more of what you eat. *Wall Street Journal*. Retrieved from www.wsj.com/994b4773-d72c-47e3-9091-d440428ba204.

Kahn, P. H., Severson, R. L., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., . . . Freier, N. G. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction—HRI '12* (p. 33). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2157689.2157696

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Englewood Cliffs, NJ: Prentice-Hall.

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Academy of Management Review*, *34*(3), 401–422.

Kim, T., & Hinds, P. (2006, September). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication* (pp. 80–85). Piscataway, NJ: IEEE.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., . . . Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences*, *104*(50), 20084–20089.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. New York: The Viking Press.

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

Lewicki, R. J., & Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 287–313. https://doi.org/10.1146/annurev-orgpsych-032516-113147

Ludwig, D. C., & Longenecker, C. O. (1993). The Bathsheba syndrome: The ethical failure of successful leaders. *Journal of Business Ethics*, *12*(4), 265–273.

Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *Trust and autonomous systems: Papers from the 2013 AAAI Spring symposium* (pp. 48–53). Boston, MA: AAAI Press.

Lyons, J. B., Ho, N. T., Koltai, K. S., Masequesmay, G., Skoog, M., Cacanindin, A., & Johnson, W. W. (2016). Trust-based analysis of an Air Force collision avoidance system. *Ergonomics in Design*, 9–12. https://doi.org/10.1177/106484615611274

Lyons, J. B., Koltai, K. S., Ho, N. T., Johnson, W. B., Smith, D. E., & Shively, R. J. (2016). Engineering trust in complex automated systems. In *Ergonomics in design, January, 13–17*. Boston, MA: Sage.

MacDorman, K. F. (2005). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In *ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science*. Retrieved from www.macdorman.com/kfm/writings/pubs/MacDorman2006SubjectiveRatings.pdf

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, *7*(3), 297–337. https://doi.org/10.1075/is.7.3.03mac

MacDorman, K. F., & Kageki, N. (2012). The uncanny valley: The original essay by Masahiro Mori. In *IEEE spectrum*. New York, NY: IEEE Press.

Madhavan, P., & Wiegmann, D. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*, 356–376.

Mathieson, K. (2007). Towards a design science of ethical decision support. *Journal of Business Ethics*, *76*(3), 269–292.

Matzenbacher, M. B. (2018, March–April). The US Army and mission command. *Military Review*, 2018, pp. 61–71.

May, J. F., & Baldwin, C. L. (2009). Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(3), 218–224. https://doi.org/10.1016/j.trf.2008.11.005.

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, *38*(1), 24–59.

McMaster, L. G. H. (2015, March–April). The army operating concept and clear thinking about future war. *Military Review,* 2015, pp. 6–21.

McNeese, M. D. (1986). Humane intelligence: A human factors perspective for developing intelligent cockpits. *IEEE Aerospace and Electronic Systems*, *1*(9), 6–12.

Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper & Row.

Mohammed, S., & Dumville, B. C. (2001). Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *22*(2), 89–106.

Mori, M. (1970). The uncanny valley. *Energy*, *7*(4), 33–35. Tokyo: Nihon Enerugi Gakkaishi/ Journal of the Japan Institute of Energy.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*(4), 602.

Nass, C., Moon, Y., Fogg, B., Reeves, B., & Dryer, D. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, *43*(2), 223–239.

Nass, C., Steuer, J., & Tauber, E. (1994). Computers are social actors. In *CHI '94 Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 73–78). Boston, MA: ACM.

O'Fallon, M. J., & Butterfield, K. D. (2005). A review of the empirical ethical decision-making literature: 1996–2003. *Journal of Business Ethics*, *59*(4), 375–413.

Parasuraman, R., Sheridan, T., & Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *30*, 286–297.

Parasuraman, R., Sheridan, T., & Wickens, C. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, *2*, 140–160.

Rest, J. R. (1986). Moral development: Advances in research and theory. Retrieved from http://hdl.handle.net/10822/811393

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*, 393–404.

Salas, E., Rosen, M. A., Burke, C. S., & Goodwin, G. F. (2009). The wisdom of collectives in organizations: An update of the teamwork competencies. In *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 39–79). New York, NY: Psychology Press.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*(5626), 1755–1758.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2011). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, *7*(4), 413–422. https://doi.org/10.1093/scan/nsr025.

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. *Automation and Human Performance: Theory and Applications*, 37–63.

Scerbo, M. W. (2008). Adaptive automation. *Neuroergonomics: The Brain at Work*, *3*, 239.

Schaubroeck, J. M., Hannah, S. T., Avolio, B. J., Kozlowski, S. W., Lord, R. G., Treviño, L. K., . . . Peng, A. C. (2012). Embedding ethical leadership within and across organization levels. *Academy of Management Journal*, *55*(5), 1053–1078.

Seyama, J., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and Virtual Environments*, *16*(4), 337–351. Retrieved from www.mitpressjournals.org/doi/abs/10.1162/pres.16.4.337

Snyder, D. E., & McNeese, M. D. (1987). *Conflict resolution in cooperative systems* (No. AAMRL-TR-87-066). Wright-Patterson AFB, OH: Harry G Armstrong Aerospace Medical Research Lab.

Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(43), 18362–18366. https://doi.org/10.1073/pnas.0910063106

Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.

Sweeney, P. J., Imboden, M. W., & Hannah, S. T. (2015). Building moral strength: Bridging the moral judgment-action gap. *New Directions for Student Leadership*, *146*, 17–33. https://doi.org/10.1068/p6900

Sycara, K., & Lewis, M. (2004). Integrating intelligent agents into human teams. In E. E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 203–231). Washington, DC: American Psychological Association.

Teich, D. A. (2018). IBM research project debater: Closer to passing the turing test. *Forbes*. Retrieved from www.forbes.com/sites/davidteich/2018/06/26/ibm-research-project-debater-closer-to-passing-the-turing-test/

Thompson, J. C., Trafton, J. G., & McKnight, P. (2011). The perception of humanness from the movements of synthetic agents. *Perception*, *40*, 695–704. https://doi.org/10.1068/p6900

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior* (pp. 185–208). New York: Springer.

Wellens, A. R. (1993). Group situation awareness and distributed decision making: From military to civilian applications. In J. Castellan (Ed.), *Individual and group decision making: Current Issues* (pp. 267–291). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zak, P. J. (2017). The neuroscience of trust. *Harvard Business Review*, *95*(1), 84–90.