Certifiable Trust in Autonomous Systems: Making the Intractable Tangible

Joseph B. Lyons, Matthew A. Clark, Alan R. Wagner, Matthew J. Schuelke

■ *This article discusses verification and validation (V&V) of autonomous systems, a concept that will prove to be difficult for systems that were designed to execute decision initiative. V&V of such systems should include evaluations of the trustworthiness of the system based on transparency inputs and scenario-based training. Transparency facets should be used to establish shared awareness and shared intent among the designer, tester, and user of the system. The transparency facets will allow the human to understand the goals, social intent, contextual awareness, task limitations, analytical underpinnings, and team-based orientation of the system in an attempt to verify its trustworthiness. Scenario-based training can then be used to validate that programming in a variety of situations that test the behavioral repertoire of the system. This novel method should be used to analyze behavioral adherence to a set of governing principles coded into the system.*

Advances in robotics may reshape the landscape of daily life, yet those in the military have been part of the robotics revolution for some time now. One cannot traverse far within military echelons nor listen to the popular press without hearing planning, discussion, and for some, a great deal of concern regarding the military's latest push toward autonomous systems. The military's use of drones (uninhabited aerial systems, or UASs) has been a ubiquitous topic of discussion/criticism within the popular media for several years since their highly publicized use in regions such as Pakistan, Yemen, and Afghanistan. Much of the chagrin surrounding these systems, despite the fact they are currently teleoperated with human oversight and command, has to do with whether or not we can or should trust them in a combat environment. Robotic systems within the military may be operated in hostile, complex situations and may, someday, be given the authority to execute lethal decisions within the battle space (Arkin 2009). However, future concept of operations (CONOPS) will likely inject greater autonomy into these systems that will ultimately increase the need

for understanding the trust dynamics that exist between humans and machines. As will be discussed in this article, the challenge of understanding these trust dynamics is more complicated than simply increasing the system's reliability.

Artificial intelligence methods need to be developed that allow robots to operate with a wide variety of different human users in complex social situations involving both conflict and cooperation; to learn from and personalize their decision making to the needs, training, culture, and norms of a particular person or group; and to reflectively consider how each possible action the system takes will affect not only the mission's goals but also the person's assessment of the robot's competency. Furthermore, AI designs that increase the level of transparency will not only engender operator trust during these scenarios but should additionally balance the level of trust between the operator, certifier, and designer of these systems. The recent United States Air Force (USAF) 30-year strategy announced future plans to invest in greater levels of autonomous systems, stating, "The accelerated development of artificial intelligence and like technologies will revolutionize the concept of autonomy. Whereas we view autonomous systems as those able to execute a set of pre-programmed functions, future systems will be better able to react to their environment and perform more situational-dependent tasks as well as synchronized and integrated functions with other autonomous systems. This will provide tremendous flexibility in highly-contested environments" (US Air Force 2014). One of the potential advantages of having systems that can learn and execute decision authority is that they may be more adaptable to dynamic situations relative to contemporary systems. In this future human-machine paradigm, the human should be considered a partner and not a governor or overseer of autonomy. He or she must be able to place trust in an autonomous partner and will not be able to take on the workload of the autonomy if a failure occurs. Furthermore, it may be necessary for the machine to recognize a person's mistakes and to challenge that person's decision making, especially if ethical guidelines may be violated (Arkin, Ulam, and Wagner 2012). Such topics dictate multidisciplinary approaches to autonomy to maximize the convergence of disciplines such as AI, computer science, engineering, and psychology. If a human-machine team is to, in fact, act as a partnership, how do we engender trust?

Military doctrine, notably within the United States Air Force has also acknowledged this issue as one of the most critical research challenges involving autonomous systems. "In the near to mid-term, developing methods for establishing 'certifiable trust in autonomous systems' is the single greatest technological barrier that must be overcome to obtain the capability advantages that are achievable by increasing use of autonomous systems" (Dahm 2010). The Defense Science Board report on autonomy stated:

> Test and certification techniques that are appropriate for autonomous systems may be dramatically different from those used for manned platforms: The projected exponential growth in Software Lines of Code (SLOC) and the nondeterministic nature of many algorithms will lead to prohibitive costs to test exhaustively. In lieu of this brute force approach, timely and efficient certification (and recertification) of intelligent and autonomous control systems will require analytical tools that work with realistic assumptions, including approaches to bound uncertainty caused by learning/adaptation or other complex nonlinearities that may make behavior difficult to predict. Test and certification will need to prove not just safety, but also level of competence at mission tasks. This will require clearly defined metrics for stability, robustness, performance, controllability, for example, and the development of new tools for software verifiability and certification. Over time, machine learning will become an important aspect to autonomous system performance and will pose extreme challenges to test and certification of systems (Defense Science Board 2012).

More recently, the National Academy of Sciences report, *Autonomy Research for Civil Aviation*, states regarding "Trust in adaptive / nondeterministic Increasingly Autonomous (IA) systems" that "Verification, validation, and certification are necessary but not sufficient to engender stakeholder trust in advanced adaptive/nondeterministic IA systems" (National Research Council 2014). In this spirit, the current article will examine the concept of trust in autonomous systems and it will offer a model to move this complex construct into tangible recommendations in an effort to reduce the current assurance burden.

For the purposes of this article, consider autonomy as "systems that have a set of intelligence based capabilities that allow it to respond to situations that were not pre-programmed or anticipated in the design (that is, decision-based responses). Autonomous systems have a degree of self-government and self-directed behavior (with the human's proxy for decisions)" (Masiello 2013). V&V for truly autonomous systems is inherently problematic for many reasons. For the purposes of this article: (1) the very nature of autonomy suggests that the system will evidence "decision initiative" wherein it behaves unpredictably; (2) the contexts in which autonomous systems may offer the greatest value are those characterized by high levels of uncertainty where humans and machines may work well together as part of a human-machine team; and (3) the assurance that an autonomous system will behave as intended, and the evidence required to create that assurance (or trust) in autonomous systems differs widely depending on one's perspective (for example, designers, testers, certifiers, or users; Department of Defense [2015]).

It is important to highlight that the term *autonomous system* is meant to refer to the whole system, including the platform, hardware, sensors, actu-

ators, software, and other aspects. This fact from a certification perspective, however, is problematic in essence. Contrary to the operator or user, the most difficult and challenging component within an autonomous system to trust (or gain assurance of) is the intelligent, learning, and adaptive software embedded within. This is primarily due to the certifier's responsibility to argue with confidence, supported by evidence that the system will always behave as intended. With that in mind, and due to the inability to test all possible combinations of failures, traditional verification and validation of critical software systems typically occur through a process of testing each software component in isolation, without regard to the interactions of each component, relying on future "integrated" tests to prove out the entire system and the complex interactions between the components. Within this environment, strict rules govern the design and operation of any software within the safety critical domain. Such restrictions inhibit all but the simplest form of automation. The more recent Joint Strike Fighter coding standards provide an example of some of these restrictions, that is, no recursion, no introduction of variables without immediately initializing them with meaningful values, and no initially "unreachable" code. (Lockheed Martin 2005). Additionally, the level or extent of testing required of all software is dependent on the level of risk that software will present to a system and its users.

In addition to performance or operational requirements, the evaluation of safety requirements is intended to assess how critical or to what extent a failure could cause loss of life or significant cost. This criticality assessment is a key driver for evidence generated in both the verification and validation stages of a system design and is highly dependent on the intended operational environment and the requirements of the system. For example, some current military and civilian aviation system safety regulations, MIL-STD-882D and SAE ARP 4761, respectively, imply that safety-critical software must be deterministic or time invariant, meaning that for any given input, the software will produce the same output for all periods of assessment. Even for these systems, software failures due to programming bugs or unintended uses can produce catastrophic results. For example, adaptive flight-critical software (a sublevel of autonomous software) remains a significant challenge for the civil aviation community; proving to be "particular difficultly to certify because by definition [adaptive systems] change their software defined parameters whilst in operation in response to the experienced time varying operating environment" (Wilkinson, Lynch, and Bharadwaj 2013). Thus, for autonomous systems new AI development and test methodologies must be created that verify how these systems operate as well as characterize how and when they fail not only prior to deployment but continuously throughout the operational cycle of the system. This is an important challenge facing artificial intelligence researchers, robotics designers, testers, and certification authorities. Developing such tests may demand considerable out-of-the-box thinking, perhaps involving probabilistic models of performance, but specifically changing the current testing and evaluation paradigm to a more continuous, transparent approach as described in the Department of Defense's Test and Evaluation, Verification and Validation (TEVV) of Autonomy Working Group's strategy (Department of Defense 2015). Some of the novel approaches may include not only changes to the design process, but also changes to the testing process that incorporate scenario-based training as a means to evaluate the effectiveness of autonomy under varying levels of complexity/difficultly (for example, using scaffolding techniques).

These software constraints, and the verification activities performed to ensure that they are met, are imposed due to an implied argument of safety, an implied argument of risk, and ultimately trust: trust by certification authorities that a system is acceptably safe and secure within a particular context. Certification, or the comprehensive evaluation of a process, system, product, event, or skill typically measured against some existing norm or standard, presents an additional problem. Certification of airworthiness within civil aviation, for example, relies heavily on the remote human pilot to mitigate any risks that arise within untested, uncertain environments. This poses significant limitations in a human-machine team. "Aviation has been very successful with a human-centric paradigm, the idea that it is humans that save the day," (Warwick 2014). Within the automotive or ground autonomy domain, the National Highway Traffic Safety Administration (NHTSA) released a preliminary report concerning the development, testing, and licensure of driver-aided autonomous vehicles on national roadways. Even the highest level of autonomy within their horizon still employs the human operator as a failsafe mechanism, stating, "Several State automated vehicle laws consider the person who activates the automated vehicle system to be the 'driver' of the vehicle even if that person is not physically present in the vehicle. NHTSA, however, is not aware of any prototype automated vehicle systems that are capable of operating on public roads without the presence of a driver in the driver's seat who is ready to control the vehicle" (National Highway Traffic Safety Administration 2013). If the current certification paradigm relies on an implied argument of trust, an argument based heavily on the human to provide risk mitigation, the advantages of the autonomy will not be realized. New certification standards, design standards, requirements, and arguments of safety, must be developed to enable the next generation of autonomous capability.

This article attempts to address a set of requirements and design approaches to facilitate the engenderment of trust in autonomous systems. These requirements enforce transparency in an autonomous design based on a task-based perspective and using training/testing to examine the system's adherence to a set of principles in accordance with the trustworthiness of the system as a method for reducing future uncertainty regarding the system's behavior and intent. This article is not suggesting that current software verification, validation, and ultimately certification requirements no longer apply. However, it is important to understand that the linkage between operator trust and system certification drives a new paradigm in both how the system is designed and continuously verified. As stated in the recently published Defense Science Board *Summer Study on Autonomy*, the design cycle of an autonomous system is envisioned as a continuous process that "begins with experimentation and development of doctrine and CONOPs, followed by specification of operational requirements, and proceeds to system design, development, testing, training, operations, and maintenance. While all these functions are part of any normal process to field a new system, the distinction for an autonomous system is that the sequence is not linear, but a continuous process spanning the entire system lifecycle." (Defense Science Board 2016). Within the Department of Defense's Autonomy Community of Interest Working Group's TEVV strategy, it is proposed that a new, iterative method of self-verification be designed into an autonomous system such that during initial operation, the system can update or adapt its design while subsequently providing its own self-verification (Department of Defense 2015). The key concept that dramatically changes the landscape of traditional V&V is the ability for the autonomous agent to incrementally verify itself, arguing its own safety in a sense. This type of run-time verification elicits a capability to encode run-time contracts that become part of the autonomous agent's functional behavior. This enables the agent to incrementally check its behavior against these constraints, modify its response to conform to these constraints, and provide feedback to the operator about the state of its behavior. A recent study performed by the IDA Corporation investigated this approach to verification, referring to it as "Evidenced Based Licensure of Autonomous Systems." Although not completely published, this study focuses on changing the design process for autonomy to allow structured learning to be counted as evidence of certification. Using this new design for certification paradigm, it is proposed that a new set of design requirements specifically targeted to engender system transparency, trust, and trust calibration will augment and potentially reduce the burden on current verification, validation, and certification processes for future autonomous systems.

Trust, Trustworthiness, and Trust Calibration

Trust represents the willingness of individuals to be vulnerable to the actions of others with little ability to monitor the others (Mayer, Davis, and Schoorman 1995). In contrast to a V&V process that tests every aspect of software to create a lack of risk, trust represents the adoption of risk. A key element of Mayer and colleagues' (1995) model of trust is the separation of "trust" (a human psychological intention) from its antecedents (labeled trustworthiness). Antecedents of trust within the context of human-machine trust has been operationalized in terms of constructs that range from reliability (Wang, Jamieson, and Hollands 2009), system performance (Hancock et al. 2011), analytic transparency (Dzindolet et al. 2003), to the anthropomorphic features of the system (Pak et al. 2012) and social etiquette (Parasuraman and Miller 2004). From a V&V perspective it will be important for humans to be aware of the trustworthiness of the system, yet the salience of the specific facets of trustworthiness will likely differ based on the role of the human interacting with the system.

Prior research has shown that human reliance (that is, the behavioral manifestation of trust) on technology can be suboptimal (Lee and See 2004) and the introduction of technology in the form of automation may lead to unpredictable/unintended consequences (Parasuraman and Riley 1997). For example, a recent meta-analysis of the trust in automation literature found that systems characterized by the highest level of automation (action implementation) were associated with the highest degradation of performance when the system failed (Onnasch et al. 2014). Therefore, a critical facet of autonomous systems is promoting "appropriate reliance" (see Lee and See [2004] for a review).

As humans, we have a tendency to rely too much on technology (that is, misuse) or not to rely enough on technology (disuse; see Parasuraman and Riley [1997]). In fact, many high-profile aviation accidents have been blamed on overreliance on technology (for example, autopilot), which could lead to loss of situational awareness reducing the pilot's ability to react appropriately to situational demands (Geiselman, Johnson, and Buck 2013). The inverse, under reliance, is also a threat to autonomous systems within the military, as this would result in inefficiencies at the strategic, operational, and financial domains. The above examples highlight the need for a trust-calibration process during V&V within the military where a designer, tester, and user of a system analyze the trustworthiness of the system and make an informed decision to rely on that system, or not. Yet, the dynamic nature of human interaction with autonomous systems, complexities that arise from situational factors, and the unique information needs of the different perspectives (designer, tester, and

user) make this problem challenging. There are two approaches discussed herein that could improve trust calibration and ultimately form the basis of a method to certify trust of autonomous systems (that is, V&V): quantification/communication of trustworthiness of the system through transparency (ultimately to "verify" the trustworthiness of the system) and training/testing in a variety of scenarios to "validate" the operator's trust of the system under disparate situational constraints.

For artificial intelligence researchers, an examination of trust and intelligent systems encompasses both the creation of intelligent methods that allow people to accurately calibrate their trust and hence the risks they place in the system and also the creation of intelligent systems that evaluate the trustworthiness of the system's human partner. For near-term military applications the trust calibration of users is an immediate concern. Lack of trust has been shown to result in disuse; placing too much trust in a system tends to result in misuse (Parasuraman and Riley 1997). For true human-machine partnerships, the intelligent system will need to also evaluate the person's trustworthiness, autonomously assessing the risk associated with relying on a particular person's actions, information, or directions. In both cases it will be critical that the autonomous system be transparent and able to communicate with the human in a manner that takes into consideration the person's background, level of understanding, and predilections.

Verifying Trustworthiness Through Transparency

Researchers in the area of automation have demonstrated that one method to improve users' ability to calibrate their trust of an automated tool is to provide the user with information about the analytical underpinnings of the tool (for example, how the system works [Dzindolet et al. 2003] or the rationale for a recommendation [Lyons et al. 2016]) as well as by proving information about the awareness, understanding, and projected states of automated agents (Mercado et al. 2016). System interfaces and/or training manipulations that explain why a system may fail (Dzindolet et al. 2003, Kim and Hinds 2006) or that highlight the reliability levels of an automated tool (Wang, Jamieson, and Hollands 2009) have been shown to influence trust calibration. Recently, researchers have called for an expanded model of transparency to address the potential added complexity evident within autonomous systems and robotics. Transparency, in this sense, can be defined as the communication of system-centered factors and human-centered factors that promote shared awareness and shared intent within a human-machine team (see Lyons [2013]). This definition of transparency is much broader and intended to leverage the rich information taken for granted within human-human relationships. For instance, social cues, intentional cues, shared awareness of environmental conditions, and context-specific capability awareness could all play an important role in promoting better human-machine interactions and ultimately better trust calibration. This expanded model of transparency will be briefly discussed.

Lyons (2013) outlines various parameters of an expanded model of transparency that includes an intentional model, task model, analytic model, environment model, teamwork model, and human state model. The intentional model focuses on communicating to the operator the higher-level purpose of the technology, the method and style of interaction to be expected, the social/moral intentions of the technology (for example, is this technology designed to follow commands, to interrupt the task scenario when necessary, to reduce threats or risks at the expense of efficiency?), and some understanding of the technology's goal structure. This information should provide operators with some sense of general predictability with regard to how interactions with the technology might occur while also giving them a sense of the system's priorities. Researchers believe that the physical appearance of a robot can afford cues to the users pertaining to the robot's functionality (Fischer 2011; Goetz, Kiesler, and Powers 2003). Similarly, predictability of the system's social interaction and "moral" programming should help to foster appropriate trustworthiness. In this sense, the "how" should be defined in terms of broad categories of behavior perhaps akin to the laws of robotics coined by Isaac Asimov. Rather than solely having a task-driven model of behavior, the users should have an understanding of the robots moral, albeit, programmed, philosophy of interaction with humans.

The task model is much more context specific. The task model consists of the system's understanding of a task structure, information relating to the system's awareness of goals in relation to a task, information relating to the system's real-time progress in relation to those goals, and awareness of when the system makes a mistake (an inevitable aspect of any machine). A very useful example of the task model in practice is the Global Positioning System (GPS) tools that we use every day. The GPS highlights the destination, the route, and the position along the route for drivers in real time. When the GPS loses signal or when the driver misses a turn, the GPS will say "recalculating" or "lost satellite reception." Granted, many autonomous systems will be more complex than GPS systems; however, this relatively simple representation of the task model can go a long way in supporting operators who are interacting with the tool. This will provide useful information to the human regarding where the system is in terms of its task sequence and why it is performing a certain action or behavior. An important facet of the task model would be the

system's awareness of its capabilities in a given context. Understanding for example, that the reliability of the system is questionable under specific conditions would do a great deal to promote appropriate trust from the human users of the systems.

The transparency element that has thus far received the majority of attention has been the analytic model. The analytic model provides operators with an understanding of how the system works, what calculations and algorithms it uses, and why it might make an error. Research has shown that added information about the analytical properties of an automated system or robot can improve the operator's ability to accurately judge trust of the system (Dzindolet et al. 2003, Kim and Hinds 2006). However, this added information could have a negative impact on the operators if it led to information overload in a demanding task situation or if it confused novice operators by being too complex to understand. Therefore, the analytical model is probably best implemented in the training or socialization phase of the human-robot interaction. Further, great care should be taken to truly understand how much information is enough to promote optimal trust calibration while minimizing data overload. Research has shown that added transparency can be accomplished without having a negative workload effect (Mercado et al. 2016).

In contrast to the analytic model, the environment model should present operators with real-time information communicating the system's awareness of environmental conditions, constraints, and task-related limitations in relation to the environment. Returning to the GPS example, a system could inform operators of potential upcoming traffic jams, road conditions, or ongoing construction as a method to update the operator of the real-time constraints within the environment. Examples of how the environment model could be operationalized in military robotic air platforms could include understanding weather patterns or flight damage to a platform such as in NASA's Emergency Landing Planner (Meuleau et al. 2009), using digital terrain elevation data for collision avoidance (Koltai et al. 2014), awareness of threats in different geographical areas, or strength of satellite or network connectivity in different geographic regions. This awareness of environmental conditions will be particularly useful for distributed human-machine interactions (that is, where the operator and the system are not colocated). For instance, having shared awareness of environmental conditions is critical for autonomous systems for NASA, which may be monitoring systems on one planet and that operate on another (Stubbs, Wettergreen, and Hinds 2007). Shared awareness of environmental conditions will allow operators to anticipate action, understand anomalous behavior, and promote better adaptation to novel demands. Similarly, shared mental models have been shown to be useful in human-human teams by facilitating adaptability and performance (Marks et al. 2002).

An essential facet of future autonomous systems will be their ability to effectively team with human counterparts (Chen and Barnes 2014, Lyons 2013). Fortunately, there is a wealth of knowledge from the literature on human-human teams that researchers can draw on in making recommendations for human-machine teams. Specifically, prior research has demonstrated the importance of understanding the roles, responsibilities, and duties of one's teammates (Marks et al. 2002; Volpe, Cannon-Bowers, and Salas 1996) as such information will allow individuals to anticipate the behavior of their teammates, provide backup behavior, and generally have a shared mental model of the team. The same information is important in human-machine teams, yet designers often neglect this fact. Therefore, the teamwork model within the expanded transparency theory suggests that adding information about the team dynamics between the human operator and an autonomous system(s) will improve the human-machine interaction because it will allow the operator (and to some extent the system) to anticipate the needs or actions of the teammate, ultimately reducing uncertainty. Parasuraman, Sheridan, and Wickens (2000) provide a useful framework for division of labor between humans and robots in their discussion of different stages of information processing and their discussion of levels of automation. They discuss information processing as consisting of information acquisition, information analysis, decision analysis and selection, and action implementation. Even such a high-level framework as this could be useful in terms of fostering a shared awareness between a human and a robot. Once the higher-level division of labor is shared and understood between both parties, it will be important for a set of norms to be defined to negotiate uncertainties and the dynamic nature of teamwork; this will be especially true if the human is executing supervisory control of multiple robots at one time. Examples of the teamwork model could include: (1) visualization of the division of labor in a given task context (this could change dynamically as a task scenario unfolds overtime), and (2) identification of roles and responsibilities both generally and within a particular task context.

Specific recommendations for fostering transparency from the perspective of designer, tester, and user can be found in table 1.

Developing Transparent Artificially Intelligent Systems

Artificial intelligence researchers have long recognized the value of creating transparent artificially intelligent systems (Minsky 1974). A transparent intelligent system is capable of communicating the reasons for its behavior to its human partner in a way

Transparency Factor	Perspective		
	Designer	Tester	User
Intention/social Intention	Design matches intended use Interaction style matches desired programming Goal-based behavior is traceable back to code Design does not violate expectations of the user	Design matches intended use Interaction style does not compromise safety or performance Testing scenarios incorporate a wide variety of contexts to foster awareness of the system's adherence to intentional programming	Design matches intended use Interface is intuitive and easy to use Interaction style facilitates engagement with the system Interaction style promotes dialogue/exchange, sense of psychological safety and competence
Environment	Algorithms/sensors are accurate at a representative sample of ranges of specified capability Algorithms/sensors are sensitive enough to changes in the environment to be able to inform the user	Algorithms/sensors are accurate at a representative sample of ranges of specified capability Changes in sensor data are traceable to changes in the environment	Interface and or training fosters awareness of how the system acquires and analyzes information from the environment Interface and training foster awareness of how the system's capabilities change in different environmental conditions
Task	Behavior is traceable to code System performs effectively	System meets performance standards in diverse scenarios	Interface is familiar to the user Interface and training promotes situational awareness of task-based attention of the system Allows user to anticipate the system's future action
Analytic	Logic driving behavior is traceable to code	Observed behaviors align to the moral programming of the system in a variety of representative domains	Interface is intuitive to the user Logic driving behavior is clear to the user Logic driving behavior is traceable by the user Rationale for errors and potential errors is clear to the user
Teamwork	Transition between human and system-driven requirements is traceable to sensor inputs from the system	Transition between human and system occurs at logical critical decision points in diverse scenarios	System communicates awareness of division of labor and adjusts as situations change System informs user when it is safe the transfer authority to the user unless doing so would compromise safety
Human State	Not covered in this report	Not covered in this report	Not covered in this report

Table 1. Transparency Recommendations for Verifying System Trustworthiness for Designers, Testers, and Users of Autonomous Systems.

that will be understood. Depending on the application, the presentation of these reasons may be verbal, written, auditory, individualized, based on intimate experiences with the person, or guided by the rule of law or by the laws of war. The establishment, maintenance, and calibration of trust may demand that an intelligent system's behavior be transparent. Moreover, V&V best practices would be well served by introspective mechanisms allowing testers to peer into an autonomous system's reasoning.

The possibility of creating transparent, artificially intelligent systems has been examined, occasionally under different monikers, as part of several of artificial intelligence's subdomains. In data mining and machine learning, for example, transparency-related research has explored methods for characterizing data in terms of human interpretable models (Ruping 2006). An interpretable model is a model that is efficient, accurate, and understandable. At a more foundational level, Halpern and Pearl (2005) examine formal methods focused on resolving causation and using these uncovered causes to generate automated explanations. Another perspective has been provided by the human factors and user design community. Chen and Barnes (2015) define transparency "as the descriptive quality of an interface pertaining to its ability to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process." As part of their SAT model, they present this information in three different levels: the first level presents the agent's desires and intentions; the second level offers the agents beliefs as well as the environmental constraints that may be causing these beliefs; the third and final level displays the agent's predictions for the future (Barnes, Barber, and Procci, 2016).

Unfortunately, these approaches do not individualize their communications with the user or consider the person's knowledge, background, or the timing of the transparency messages. One challenging aspect of creating a transparent intelligent system is knowing when to be transparent. A system that explains its behavior too often may increase the cognitive load placed on the person. Moreover, improperly timed explanations can affect trust as extensively as an error itself, perhaps even appearing disingenuous or deceptive. Robinette, Howard, and Wagner (2015) presented human subjects with a guidance robot that led them to a meeting room in a virtual world. While in the room, an emergency was generated and the robot offered to guide the person to the nearest emergency exit. Previous studies had shown that if the robot made navigation errors on the way to the meeting room, subjects tended not to trust the robot and elected not to use it to find the exit (Robinette, Howard, and Wagner 2015). Yet, when the robot explained why it had failed or apologized for the failure, the subject's trust would be repaired only if the timing of the explanation or apology was right. Explanations generated immediately after the mistake had no impact on trust but explanations made just prior to the person's decision to use the robot during the evacuation resulted in complete trust repair even though the wording of the messages was identical. Achieving transparency may therefore be significantly more challenging than simply providing the user with reasons for the system's behavior.

As discussed, transparency can potentially be arrived at using a series of interconnected models that present specific types of information at the right time. It is worth considering how the underlying computational representations that foment a robot's behaviors might be used to generate transparent statements that lend understanding to the system's operation and reasoning. As has been well noted, some representations (such as decision trees) lend themselves naturally to communication with people. Unfortunately, many representations that currently show the most promise (for example, neural networks, graphical models) are the least naturally understandable. Yet some recent work is demonstrating that even these types of computational representations may afford interesting methods for demonstrating transparency (Yosinski et al. 2015). Additional research is needed to provide insight related to the design of transparent intelligent systems. However, it will be imperative that multidisciplinary approaches and teams be formed to address these challenges. AI researchers are adept at computational methods and representations, yet they may be joined by social scientists to test out methods for garnering understanding and acceptance of these computational representations outside of AI communities. Otherwise, we run the risk of developing elegant algorithms and mathematical formalizations that the majority of individuals cannot understand, limiting their ability and motivation to appropriately calibrate their trust of such systems.

Scenario-Based Training

Once the design parameters are implemented to verify the trustworthiness of the system as described, human-machine teams should engage in a series of exercises to validate the trustworthiness of the system based on operator perceptions using various scenarios that manipulate factors that shape trust. The scenarios used during the human-machine training should test the envelope both in terms of performance expectations but also uncertainty. Testers will want scenarios that create morally contentious situations for the autonomy to see how it will react to ambiguous stimuli. Users will want to train with systems during ideal and degraded communications links so that they can get an understanding of the range of potential situations that may exist.

Experts perform better than novices do at estimating the trustworthiness of autonomous systems

because these systems are more transparent to experts (Chen, Barnes, and Harper-Sciarini 2010; Desai et al. 2009; Lee and See 2004). The more users understand how a system functions and develop the skills to identify when the system is approaching the bounds of acceptable operation, the more accurate their estimates of trustworthiness should be (Dzindolet et al. 2003; Wang, Jamieson, and Hollands 2009). Unfortunately, the circumstances that necessitate autonomous systems are often very complex, and expertise among the population of users during verification and validation is typically low because it may be a new system (Hancock, Billings, and Schaefer 2011). Therefore, when attempting to assess the trustworthiness of a system during verification and validation accurately, a fundamental goal is to develop expertise as rapidly as possible in relevant, highly-complex domains.

The application of instructional scaffolding (IS) to human-machine scenario-based training could enable more accurate estimates of system trustworthiness during verification and validation. IS fosters learning in complex situations where immediate high-level performance is not feasible by providing learners with additional support to extract meaning during training scenarios (Hogan and Pressley 1997). The distinguishing feature of IS is the reciprocal relationship between trainee proficiency and training difficulty (Puntambekar and Hubscher 2005). As learners become more proficient, the training becomes more complex and provides diminishing levels of support (Pea 2004). For example, when training with an autonomous system, scaffolding could slow down desired response time to better explain a complex relationship, highlight important information, provide potential explanations for observed uncertainty, guide moral decision making, or generate deeper learning by asking trainees to explain their decision-making processes. Moreover, scaffolding offers the possibility of multitiered learning by demonstration involving hierarchical representations (Garland and Lesh 2003) and planning (Hoang, Lee-Urban, and Muñoz-Avila 2005).

The goal of accurately assessing system trustworthiness by using training to foster human-machine trust relies on complementary instructional objectives across a wide variety of cognitions, behaviors, and attitudes in a highly complex and novel setting. Because expert teams are able to effectively execute combinations of taskwork and teamwork called for by the environment, task, and work situation, team training often focuses on the development of task and team competencies (Cannon-Bowers et al. 1995). Therefore, when using training to foster trust in human-machine teams, one group of instructional objectives should focus on developing task competencies that allow the human to fulfill his or her individual role assignment. A second group of instructional objectives should focus on developing team competencies that allow the human to fulfill his or her team role assignments (Salas et al. 2008).

Partially overlapping with the first two groups of instructional objectives, but more pertinent to the current effort, a third group of instructional objectives should focus on fostering trust through the application of transparency factors in training content. The shared goal of these trust-oriented instructional objectives is to increase the expertise of the trainee with regard to system functioning. In other words, the trainee should fully understand the intent, environment, task, analytic, and teamwork models of the autonomous system (Lyons 2013; Stubb, Wettergreen, and Hinds 2007). Experiencing similarities and differences of behavior across a wide variety of situations promotes deeper learning and enhances adaptable performance, so the training should present transparency factors in as many different scenarios as feasible (Burke et al. 2006; Chen, Thomas, and Wallace 2005; Gorman, Cooke, and Amazeen 2010; Han and Williams 2008). Such situational variance could trigger machine learning in AI and provide affordances for their subsequent representation, particularly during debriefing scenarios where the human operator might interrogate the rationale for the system's actions and decision logic. Thus, the training should present basic content about the transparency factors to trainees, and then the trainee and AI should engage in varied scenarios that increase in complexity as the trainee becomes more proficient at not only his or her taskwork and teamwork but also his or her understanding of how the AI reacts to situational constraints (Holzinger et al. 2009; Lateef 2010; Salas, Wildman, and Piccolo 2009).

The intentional model is the first factor of transparency and consists of the purpose, moralities, and goals of the machine. Intention is a primal element in the development of trust because intention sets the general bounds for expectation and operates as a source for meaning during uncertainty (Desai et al. 2009, Lee and See 2004, Lyons 2013). Therefore, training should introduce the intentional model as early as possible but also provide cross-references throughout. Intention is only as good as the situation allows (Gollwitzer and Sheeran 2006). Therefore, the boundary conditions of the intentional model will be most apparent if the trainee observes the machine attempting to stay true to its intentional model under difficult circumstances such as conflicting goals, ethical dilemmas, or morally contentious situations. In each scenario scaffolding could remind the trainee of the original intentions, explain why the present situation might prevent the system from staying true to a given intent, or ask the trainee to generate his or her own reasons why original intention and actual outcomes differed.

The environmental model is another factor of transparency consisting of the awareness of the machine to its environment. An understanding of

how the machine collects and uses environmental information allows a user to anticipate machine actions, derive meaning from anomalous output, and adapt to novel environments (Hancock, Billings, and Schaefer 2011; Lee and See 2004; Lyons 2013). Similar to the intent model, knowledge of the environmental model is foundational to establishing accurate trust, but the very nature of the environmental model also makes it highly situation specific. Therefore, the environmental model must also be addressed both early, to deliver the overarching principles, and throughout training, to provide nuanced examples. Environmental modeling is only as good as the collection and processing of situational information allows. Therefore, the more a trainee observes how incomplete, conflicting, or incorrect situational input affects the machine state, the faster identification of environmental model boundary conditions will occur. In each scenario, scaffolding could highlight important but overlooked environmental information, explain suspected conflicts among stimuli, or offer advice about how to recover from a loss of situational awareness.

The task factor of transparency consists of how the machine accomplishes its responsibilities. Knowledge of how the system executes assigned tasks is important for setting realistic expectations and accurately predicting future behavior because interpositional knowledge provides an overall framework for understanding the purpose of the team and how each member contributes (Baker et al.1992; Blickensderfer, Cannon-Bowers, Salas 1998; Cannon-Bowers and Salas 1998; Cannon-Bowers et al. 1998; Volpe, et al. 1996). The task model relates to perceptions of trustworthiness by helping to define the reliability and robustness of the automated system, which leads to trust, perceived utility, and reliance (Chen, Barnes, and Harper-Sciarini 2011; Hancock, Billings, and Schaefer 2011; Lee and See 2004; Sanchez et al. 2011). In order to train the task model, scenarios could showcase a variety of tasks assigned to the system — some of which it performs well or adequately and some poorly or less than adequately. All the while, instructional scaffolding can help foster mastery of the taskwork model. For example, slowing scenarios down could provide the time needed to adequately explain the component, coordinative, and dynamic relationships among task products. Additionally, the true reliability of the system could be tracked and summary information presented to the trainee continuously, at important decision points, or at the end of scenarios. Regardless of approach, the goal is to ensure the trainee can differentiate between conditions that allow or do not allow the machine to be trusted in completing its taskwork.

The teamwork model focuses on the interaction among team members independent of taskwork. Potential skills in the model include adaptability, feedback or communication, and coordination (Cannon-Bowers et al. 1995; Salas et al. 1992). Knowledge of the teamwork model is important in the development of accurate trustworthiness perceptions because teamwork defines interaction among members and reinforces trust (Desai et al. 2009; Sheng, Tian, and Chen 2010; Hancock, Billings, and Schaefer 2011). Participating in progressively more difficult scenarios alongside the autonomous agent will help to develop teamwork skills because the trainee will become more proficient at interacting and coordinating with the system and collectively adapting to new situations. Instructional scaffolding can help foster mastery of the teamwork model through actions such as offering suggestions on how best to cooperate or coordinate with the system, directing trainee attention to missed requests for assistance from the system, or providing immediate and specific feedback after the execution of teamwork behaviors.

The analytic facet of transparency is akin to how the machine processes information and makes decisions. As such, the analytic model has direct ties to each of the other transparency factors: intention and environment serve as input to the analytic model while taskwork and teamwork are the processes and observable outputs (Lee and See 2004, Lyons 2013). Of all the transparency factors, the analytic model provides the most complete understanding of the machine and is important to trust because a holistic understanding provides for realistic expectations, promotes more accurate predictions, and allows users to evaluate critically any errors that might occur (Cuevas et al. 2007; Hancock et al. 2011; Uggirila et al. 2004). When using IS scenario-based training to foster trust through understanding of the analytic model, each scenario should highlight select strengths or weaknesses of the analytic processes. As each scenario gets more complicated, the trainee can observe when the analytic model works well, when it only performs adequately, and when it fails. Additionally, scaffolding can provide feedback, present hints, offer explanations, and question trainees to develop a deep understanding of this transparency factor. From these experiences, the trainee will better understand when it is and is not appropriate to trust the machine.

Conclusion

Verification and validation of truly autonomous systems presents a challenge to contemporary methods based on brute force, labor-intensive analysis of code. When considering future autonomous systems, such methods are as outdated as they are intractable. The research and development community is in need of novel V&V methods that allow unpredictable, learning-based systems to be tested. These approaches necessitate collaboration between the AI community and social scientists both to develop novel representations and mathematical formalizations of

autonomous systems and also to test out methods to ensure that the complexity of AI can be condensed into meaningful chunks for individuals who exist outside of the AI community. Levels of human-machine trust will be more accurate if transparency is considered (Lyons 2013). Training based on testing the intentional, task-based, analytical, teamwork, and environment aspects of transparency can also be used to establish predictability based on the system's behavior in a series of scenarios using techniques like IS. The present article concludes with a call on the AI community (that is, designers) to consider the importance of ensuring that future AI systems be not only trustworthy, but that they be developed and tested with the appropriate affordances to promote appropriate trust among users and testers.

References

Arkin, R. C. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, FL: CRC Press. doi.org/10.1201/9781420085952

Arkin, R. C.; Ulam, P.; and Wagner, A. R. 2012. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proceedings of the IEEE* 100(3): 571–589. doi.org/10.1109/JPROC.2011.2173265

Baker, C. V.; Salas, E.; Cannon-Bowers, J. A.; and Spector, P. 1992. The Effects of Interpositional Uncertainty and Workload on Team Coordination Skills and Task Performance. Paper presented at the annual meeting of the Society for Industrial Organizational Psychology, Montreal.

Blickensderfer, E. L.; Cannon-Bowers, J. A.; and Salas, E. 1998. Cross Training and Team Performance. In *Making Decisions Under Stress: Implications for Individual and Team Training,* ed. J. A. Cannon-Bowers and E. Salas, 299–312. Washington, DC: American Psychological Association. doi.org/10.1037/10278-01

Burke, C. S.; Stagl, K. C,; Salas, E.; Pierce, L.; and Kendall, D. 2006. Understanding Team Adaptation: A Conceptual Analysis and Model. *Journal of Applied Psychology* 91(6): 1189–1207. doi.org/10.1037/0021-9010.91.6.1189

Cannon-Bowers, J. A., and Salas, E. 1998. Team Performance and Training in Complex Environments: Recent Findings from Applied Research. *Current Directions in Psychological Science* 7(3): 83–87. doi.org/10.1111/1467-8721.ep10773005

Cannon-Bowers, J. A.; Salas, E.; Blickensderfer, E.; and Bowers, C. A. 1998. The Impact of Cross-Training and Workload on Team Functioning: A Replication and Extension of Initial Findings. *Human Factors* 40(1): 92–101. doi.org/10.1518/001872098779480550

Cannon-Bowers, J. A.; Tannenbaum, S. I.; Salas, E.; and Volpe, C. E. 1995. Defining Competencies and Establishing Team Training Requirements. In *Team Effectiveness and Decision Making in Organizations*, ed. R. A. Guzzo and E. Salas, 333–380. San Francisco: Jossey-Bass.

Chen, G.; Thomas, B.; and Wallace, J. C. 2005. A Multilevel Examination of the Relationships Among Training Outcomes, Mediating Regulatory Processes, and Adaptive Performance. *Journal of Applied Psychology* 90(5): 827–841. doi.org/10.1037/0021-9010.90.5.827

Chen, J. Y., and Barnes, M. J. 2015. Agent Transparency for Human-Agent Teaming Effectiveness. In *Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC),* 1381–1385. Piscataway, NJ: Institute for Electrical and Electronics Engineers. doi.org/10.1109/SMC.2015.245

Chen, J. Y. C.; Barnes, M. J.; Harper-Sciarini, M. 2011. Supervisory Control of Multiple Robots: Human-Performance Issues and User-Interface Design. *IEEE Transactions on Systems, Man, and Cybernetics,* Part C: Applications and Reviews, 41(4): 435–454. doi.org/10.1109/TSMCC.2010.2056682

Cuevas, H. M.; Fiore, S. M.; Caldwell, B. S.; and 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.

Dahm, W. J. A. 2010. *Technology Horizons: A Vision for Air Force Science and Technology During 2010–2030*, 42. Arlington, VA: United States Air Force Headquarters.

Defense Science Board. 2016. *Defense Science Board (DSB) Summer Study on Autonomy*, 36. Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics.

Defense Science Board. 2012. *Defense Science Board (DSB) Task Force on the Role of Autonomy in DoD Systems*, 100. Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics.

Department of Defense. 2015. *Autonomy Community of Interest (COI) Test and Evaluation, Verification and Validation (TEVV) Working Group: Technology Investment Strategy 2015–2018*, 4, 9, 14. Washington, DC: Office of the Assistant Secretary of Defense for Research and Engineering. Link: defenseinnovationmarketplace.mil/resources/OSD_ATEVV_STRAT_DIST_A_SIGNED.pdf

Desai, M.; Stubbs, K.; Steinfeld, A.; and Yanco, H. 2009. Creating Trustworthy Robots: Lessons and Inspirations from Automated Systems. Paper presented at the AISB Convention: New Frontiers in Human-Robot Interaction, 8–9 April, Edinburgh, Scotland.

Dzindolet, M. T.; Peterson, S. A.; Pomranky, R. A.; Pierce, L. G.; and Beck, H. P. 2003. The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies* 58(6): 697–718. doi.org/10.1016/S1071-5819(03)00038-7

Fischer, K. 2011. How People Talk with Robots: Designing Dialogue to Reduce User Uncertainty. *AI Magazine* 32(4): 31–38. doi.org/10.1609/aimag.v32i4.2377

Garland, A., and Lesh, N. 2003. Learning Hierarchical Task Models by Demonstration. Technical Report, Mitsubishi Electric Research Laboratory (MERL), USA — (January 2002). Cambridge, MA: Mitsubishi Electric Research Laboratory.

Geiselman, E. E.; Johnson, C. M.; and Buck, D. R. 2013. Flight Deck Automation: Invaluable Collaborator or Insidious Enabler? *Ergonomics in Design: The Quarterly of Human Factors Applications* 21(3): 22–26. doi.org/10.1177/1064804613491268

Goetz, J.; Kiesler, S.; and Powers, A. 2003. Matching Robot Appearance and Behavior to Tasks to Improve Human-Robot Cooperation. *Proceedings of the 2003 IEEE International Workshop on Robot and Human Interactive Communication,* 55–60. Piscataway, NJ: Institute for Electrical and Electronics Engineers. doi.org/10.1109/roman.2003.1251796

Gollwitzer, P. M., and Sheeran, P. 2006. Implementing Intentions and Goal Achievement: A Meta-Analysis of Effects and Processes. *Advances in Experimental Social Psychology* 38: 69–119. doi.org/10.1016/S0065-2601(06)38002-

Gorman, J. C.; Cooke, N. J.; and Amazeen, P. G. 2010. Training Adaptive Teams. *Human Factors* 52(2): 295–307. doi.org/10.1177/0018720810371689

Halpern, J. Y., and Pearl, J. 2005. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British journal for the Philosophy of Science* 56 (4): 843–887. doi.org/10.1093/bjps/axi147

Han, T. Y., and Williams, K. J. 2008. Multilevel Investigation of Adaptive Performance: Individual- and Team-Level Relationships. *Group Organization Management* 33(6): 657–684. doi.org/10.1177/1059601108326799

Hancock, P. A.; Billings, D. R.; and Schaefer, K. E. 2011. Can You Trust Your Robot? *Ergonomics in Design* 19(3): 24–29. doi.org/10.1177/1064804611415045

Hancock, P. A.; Billings, D. R.; Schaefer, K. E.; Chen, J. Y. C.; de Visser, E. J.; and Parasuraman, R. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors* 53(5): 517–527. doi.org/10.1177/0018720811417254

Hoang, H.; Lee-Urban, S.; and Muñoz-Avila, H. 2005. Hierarchical Plan Representations for Encoding Strategic Game AI. In *Proceedings of the First Artificial Intelligence and Interactive Digital Entertainment Conference,* 63–68. Menlo Park, CA: AAAI Press.

Hogan, K., and Pressley, M. 1997. Scaffolding Student Learning: Instructional Approaches and Issues. *Advances in Learning and Teaching*. Cambridge, MA: Brookline Books.

Holzinger, A.; Kickmeier-Rust, M. D.; Wassertheurer, S.; and Hessinger, M. 2009. Learning Performance with Interactive Simulations in Medical Education: Lessons Learned from Results of Learning Complex Physiological Models with the HAEMOdynamics SIMulator. *Computers and Education* 52(2): 292–301. doi.org/10.1016/j.compedu.2008.08.008

Kim, T.; and Hinds, P. 2006. Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interactions. In *Proceedings of the 15th International Symposium on Robot and Human Interactive Communication* (RO-MAN06), 80–85. Piscataway, NJ: Institute for Electrical and Electronics Engineers. oi.org/10.1109/roman.2006.314398

Koltai, K.; Ho., N.; Masequesmay, G.; Niedober, D.; Skoog, M.; Cancanindin, A.; Johnson, W.; and Lyons, J. 2014. Influence of Cultural, Organizational, and Automation Capability on Human Automation Trust: A Case Study of Auto-GCAS Experimental Test Pilots. Paper presented at the International Conference on Human-Computer Interaction Aerospace. Santa Clara, CA, July 30–August 1.

Lateef, F. 2010. Simulation-Based Learning: Just Like the Real Thing. *Journal of Emergencies, Trauma and Shock* 3(4): 348–352. doi.org/10.4103/0974-2700.70743

Lee, J. D., and See, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46(1): 50–80. doi.org/10.1518/hfes.46.1.50.30392

Lockheed Martin. 2005. *Joint Strike Fighter Air Vehicle C++ Coding Standards*, 42. Document Number 2RDU00001 Rev C. Bethesda MD: Lockheed Martin Corporation.

Lyons, J. B. 2013. Being Transparent About Transparency: A Model for Human-Robot Interaction. In *Trust and Autonomous Systems: Papers from the AAAI Spring Symposium* (Technical Report SS-13-07), ed. D. Sofge, G. J. Kruijff, and W. F. Lawless. Menlo Park, CA: AAAI Press. doi.org/10.1177/1064804615611274

Lyons, J. B.; Koltai, K. S.; Ho, N. T.; Johnson, W. B.; Smith, D. E.; and Shively, J. R. 2016. Engineering Trust in Complex Automated Systems. *Ergonomics in Design* 24(1): 9–12.

Marks, M. A.; Sabella, M. J.; Burke, C. S.; and Zaccaro, S. J. 2002. The Impact of Cross-Training on Team Effectiveness. *Journal of Applied Psychology* 87(1): 3–13. doi.org/10.1037/0021-9010.87.1.3

Masiello, T. J. 2013. Air Force Research Laboratory Autonomy Science and Technology Strategy, 3. Strategy Report, Air Force Research Laboratory, Wright-Patterson AFB, Ohio. Online: www.defenseinnovationmarketplace.mil/resources/AFRL_Autonomy_Strategy_DistroA.PDF.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An Integrated Model of Organizational Trust. *Academy of Management Review* 20(3): 709–734.

Mercado, J. E.; Rupp.; M. A.; Chen, J. Y. C.; Barnes, M. J.; Barber, D.; and Procci, K. 2016. Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors* 58(3): 401–415. doi.org/10.1177/0018720815621206

Meuleau, N.; Plaunt, C.; Smith, D.; and Smith, T. 2009. An Emergency Landing Planner for Damaged Aircraft. In *Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference*. Palo Alto, CA: AAAI Press.

Minsky, M. 1974. *A Framework for Representing Knowledge*. Cambridge, MA: Massachusetts Institute of Technology.

National Highway Traffic Safety Administration. 2013. *Preliminary Statement of Policy Concerning Automated Vehicles*. Washington, DC: Government Printing Office.

National Research Council. 2014. *Autonomy Research for Civil Aviation: Toward a New Era of Flight*, 4. Washington, DC: The National Academies Press.

Onnasch, L.; Wickens, C. D.; Li, H.; and Manzey, D. 2014. Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors* 56(3): 476–488. doi.org/10.1177/0018720813501549

Pak, R.; Fink, N.; Price, M.; Bass, B.; and Sturre, L. 2012. Decision Support Aids with Anthropomorphic Characteristics Influence Trust and Performance in Younger and Older Adults. *Ergonomics* 55(9): 1–14. doi.org/10.1080/00140139.2012.691554

Parasuraman, R., and Miller, C. 2004. Trust and Etiquette in High-Criticality Automated Systems. *Communications of the ACM* 47(4): 51–55. doi.org/10.1145/975817.975844

Parasuraman, R., and Riley, V. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39(2): 230–253.

Parasuraman, R.; Sheridan, T. B.; and Wickens, C. D. 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(5): 573–583.

Pea, R. D. 2004. The Social and Technological Dimensions of Scaffolding and Related Theoretical Concepts for Learning, Education, and Human Activity. *Journal of the Learning Sciences* 13(3): 423–451. doi.org/10.1207/s15327809jls1303_6

Puntambekar, S., and Hubscher, R. 2005. Tools for Scaffolding Students in a Complex Learning Environment: What Have We Gained and What Have We Missed? *Educational Psychologist* 40(1): 1–12. doi.org/10.1207/s15326985ep4001_1

Robinette, P.; Howard, A.; and Wagner, A. R. 2015. Timing Is Key for Robot Trust Repair. In *Seventh International Conference on Social Robotics* (ICSR 2015), Lecture Notes in Com-

puter Science 9388, 574–583. Berlin: Springer. doi.org/10.1007/978-3-319-25554-5_57

Ruping, S. 2006. Learning Interpretable Models. Ph.D. dissertation, Department of Informatics, Dortmund University of Technology, Dortmund, Germany (d-nb.info/997491736).

Salas, E.; DiazGranados, D.; Klein, C.; Burke, C. S.; Stagl, K. C.; Goodwin, G. F.; and Halpin, S. M. 2008. Does Team Training Improve Team Performance? A Meta-Analysis. *Human Factors* 50(6): 903–933. doi.org/10.1518/001872008X375009

Salas, E.; Dickinson, T.; Converse, S. A.; and Tannenbaum, S. I. 1992. Toward an Understanding of Team Performance and Training. In *Teams: Their Training and Performance,* ed. R. W. Swezey and E. Salas, 3–29. Norwood, NJ: ABLEX.

Salas, E.; Wildman, J. L.; and Piccolo, R. F. 2009. Using Simulation-Based Training to Enhance Management Education. *Academy of Management Learning and Education* 8(4): 559–573. doi.org/10.5465/AMLE.2009.47785474

Sanchez, J.; Rogers, W. A.; Fisk, A. D.; and Rovira, E. 2011. Understanding Reliance on Automation: Effects of Error Type, Error Distribution, Age and Experience. *Theoretical Issues in Ergonomics Science* 15(2): 134–160. doi.org/10.1080/1463922X.2011.611269

Sheng, C.; Tian, Y.; and Chen, M. 2010. Relationships Among Teamwork Behavior, Trust, Perceived Team Support, and Team Commitment. *Social Behavior and Personality* 38(10): 1297–1305. doi.org/10.2224/sbp.2010.38.10.1297

Stubbs, K.; Wettergreen, D.; and Hinds, P. J. 2007. Autonomy and Common Ground in Human-Robot Interaction: A Field Study. *IEEE Intelligent Systems* 22(2): 42–50. doi.org/10.1109/MIS.2007.21

Uggirala, A.; Gramopadhye, A. K.; Melloy, B. J.; and Toler, J. E. 2004. Measurement of Trust in Complex and Dynamic Systems Using a Quantitative Approach. *International Journal of Industrial Ergonomics* 34(3): 175–186. doi.org/10.1016/j.ergon.2004.03.005

US Air Force. 2014. *America's Air Force: A Call to the Future*. Washington, DC: United States Air Force. Online: airman.dodlive.mil/files/2014/07/AF_30_Year_Strategy_2.pdf

Volpe, C. E.; Cannon-Bowers, J. A.; Salas, E.; and Spector, P. E. 1996. The Impact of Cross-Training on Team Functioning: An Empirical Investigation. *Human Factors* 38(1): 87–100. doi.org/10.1518/001872096778940741

Wang, L.; Jamieson, G. A.; and Hollands, J. G. 2009. Trust and Reliance on an Automated Combat Identification System. *Human Factors* 51(3): 281–291. doi.org/10.1177/0018720809338842

Warwick, G. 2014. Certifiable Trust Required to Take Autonomous Systems Past Unmanned. *Aviation Week and Space Technology,* August 11. Online: aviationweek.com/commercial-aviation/certifiable-trust-required-take-autonomous-systems-past-unmanned

Wilkinson, C.; Lynch, J.; and Bharadwaj, R. 2013. Final Report — Regulatory Considerations for Adaptive Systems. NASA/CR–2013-218010. Washington, DC: National Aeronautics and Space Administration.

Yosinski, J.; Clune, J.; Nguyen, A.; Fuchs, T.; and Lipson, H. 2015. Understanding Neural Networks through Deep Visualization. Unpublished Paper. arXiv preprint arXiv:1506.06579. Ithaca, NY: Cornell University Library.

Joseph B. Lyons is the technical advisor for the Human Trust and Interaction Branch within the 711 Human Performance Wing at Wright-Patterson AFB, OH. Lyons received his Ph.D. in industrial and organizational psychology from Wright State University in Dayton, OH, in 2005. Some of Lyons's research interests include human-machine trust, interpersonal trust, leadership, and social influence. Lyons has worked for the Air Force Research Laboratory as a civilian researcher since 2005, and between 2011–2013 he served as the program officer at the Air Force Office of Scientific Research where he created a basic research portfolio to study both interpersonal and human-machine trust. Lyons has published in a variety of peer-reviewed journals such as *Human Factors, Journal of Applied Social Psychology, Journal of Psychology, The Leadership Quarterly, Stress and Health, Anxiety Stress and Coping, Journal of Change Management, International Journal of Industrial Ergonomics, Personality and Individual Differences, Team Performance Management,* and *Military Psychology*.

Matthew A Clark is the branch chief for the Autonomous Controls Branch, AFRL/RQQA. Clark started his career in the Air Force Research Lab in 1998 supporting large-scale aircraft component thermal, acoustic, and static combined environment structural testing. In 2000 and 2010, respectively, he received his bachelor's and master's degrees in electrical engineering from Wright State University in Dayton, OH. In 2010, Clark served at the Air Force Material Command headquarters providing support for the test and evaluation infrastructure, strategic planning, and operational cyber security, receiving the Exemplary Civilian Service Award. In 2011 he returned to the Air Force Research Laboratory to work on the verification and validation of autonomous control systems and applications. His research interests include verifiable intelligent control systems and run-time assurance of intelligent systems.

Alan Wagner is a senior research scientist at Georgia Institute of Technology's Research Institute and is a member of the Institute of Robotics and Intelligent Machines. Wagner's research has won several awards including being selected by the Air Force Young Investigator Program. His research on deception has gained significant notoriety in the media, resulting in articles in the *Wall Street Journal, New Scientist Magaz*ine, and the journal *Science,* and is described as the 13th most important invention of 2010 by *Time Magazine*. His research has also won awards within the human-robot interaction community, such as the best paper award at RO-MAN 2007. He received his Ph.D. in computer science from Georgia Institute of Technology. He also holds a master's degree in computer science from Boston University and a bachelor's degree in psychology from Northwestern University.

Matthew J. Schuelke is a senior human factors engineer for SRA International working under contract with the Human Trust and Interaction Branch within the 711 Human Performance Wing at Wright-Patterson AFB, OH. Schuelke received his Ph.D. in industrial and organizational psychology from the University of Oklahoma in 2010 where he performed training and leadership research funded from sources such as the Defense Advanced Research Projects Agency, the National Science Foundation, and the Army Research Institute. Schuelke is published in journals such as the *Journal of Applied Psychology, Leadership Quarterly*, and *Multivariate Behavioral Research*.