Human-Intelligent Systems Integration (2021) 3:79–97
https://doi.org/10.1007/s42454-020-00023-x

RESEARCH ARTICLE

IMPACTS: a trust model for human-autonomy teaming

Ming Hou[1] · Geoffrey Ho[1] · David Dunwoody[2]

Received: 16 July 2020 / Accepted: 4 December 2020 / Published online: 18 January 2021

Abstract
A trust model IMPACTS (intention, measurability, performance, adaptivity, communication, transparency, and security) has been conceptualized to build human trust in autonomous systems. A system must exhibit the seven critical characteristics to gain and maintain its human partner's trust towards an effective and collaborative team in achieving common goals. The IMPACTS model guided a design of an intelligent adaptive decision aid for dynamic target engagement processes in a human-autonomy interaction context. Positive feedback from subject matter experts who participated in a large-scale exercise controlling multiple unmanned assets indicated the decision aid's effectiveness. It also demonstrated the IMPACTS model's utility as a design principle for enabling trust between a human-autonomy team.

Keywords Trust · Human-autonomy teaming · Human-autonomy interaction · Human-agent interaction · Interaction-centered design · Intelligent adaptive system · Decision-making · Design guidance · Artificial intelligence · Unmanned system · Autonomous system

1 Introduction

Automation has been changing our lives dramatically over many decades, with both positive and negative impacts. To us, automation is any form of machine or software agent that performs a human task, and it is a general term that includes all machines, including autonomous systems and artificial intelligence (AI)-enabled machines. On the positive side, more intelligent automation technologies such as industry robots, personalized social robots, self-driving vehicles, and AI bring our day-to-day life and work more productivity, efficiency, and convenience. On the negative side, besides some job loss (Lamb 2016), one of the challenges when working with these intelligent machines is the risk that these agents become increasingly capable but occasionally fail (Desai et al. 2013; Onnasch et al. 2014; Sebok and Wickens 2017; Wickens et al. 2020). With more and more enabling AI technologies applied to enhance agent capabilities, these machines will become increasingly autonomous. Autonomy has been singled out as a key component of the third offset strategy for military applications, which intends to deliver leap-ahead battlefield technologies such as the robotic wingman concept that team up soldiers with autonomous systems to enhance military capability (Defense Science Board 2016).

A key challenge to successful and effective human-autonomy teaming is enabling "trust" between the human-machine team (Taylor and Reising 1995; Lee and See 2004). Trust is a foundation for effective teaming, without which, team members would spend unnecessary energy and time to re-inspect work and revalidate decisions. Lack of trust also results in less information sharing and an uneven distribution of workload (Salas et al. 2005; Sycara and Lewis 2004). To further detail the challenge, three "context constraints" technology, human capability and limitations, and system functionalities have been identified for building a trusted partnership between human and machine (Baber 2017; Hou et al. 2014). These constraints are further identified as the six barriers for human to trust autonomous systems: (1) machine's

✉ Ming Hou
Ming.Hou@forces.gc.ca

Geoffrey Ho
Geoffrey.Ho@forces.gc.ca

David Dunwoody
David.Dunwoody@forces.gc.ca

[1] Defence Research & Development Canada, Toronto, Canada

[2] Royal Canadian Air Force, Ottawa, Canada

lack of human sensing and thinking; (2) machine's inflexibility and adaptability to the changing situation; (3) machine's lack of sufficient transparency for predictable results or implication; (4) machine's lack of sufficient understanding human intent; (5) machine's ineffective interface to build human confidence; and (6) human's lack of understanding how to design machines that learn/adapt throughout their lifecycle (Defense Science Board 2016).

Trust is an abstract concept and is complex and multidimensional (Abbass et al. 2016; Lee and See 2004; Siau and Wang 2018). Trust can be attributed to a wide variety of entities, including humans, machines (hardware and software), organizations, institutions (e.g., trust in a legal system), and countries. Trust is not all or none but is continuous and can be attributed to an agent (e.g., a human or a machine) as a whole or to specific parts, capabilities, or functions of that agent. Moreover, it can be a situation or task dependent. Thus, this allows attributions of trust to vary even towards the same agent. For example, an individual may trust their spouse more than anyone else, yet they may not trust their spouse at all for specific tasks like cooking. This complexity highlights the need to be careful when examining trust as a construct that shapes (and is shaped by) behaviors of and interactions between humans and machines. As AI becomes more advanced and machines evolve into more complex and more autonomous systems, it is not surprising that how trust evolves with these systems will also change.

A number of definitions of trust exist from a wide variety of disciplines (e.g., Cho et al. 2015; Lee and See 2004). Typically trust is defined as an attitude that an agent(s) (the trustor) holds towards another agent(s) (the trustee), regarding the risk taken to depend on the trustee to achieve a goal or some positive outcome under uncertainty. Trust is commonly understood as a cognitive process and a relational mediator for interactions between humans, humans and organizations, and human-machine interactions. In the context of human interaction with intelligent technologies, trust has been regarded as a key element and a "fundamental enabler" in human interaction with autonomous systems (Hou et al. 2011, 2014).

A variety of models exist describing the development of trust in automation (e.g., Abbass et al. 2016; Hancock et al. 2011; Hoff and Bashir 2015; Lee and See 2004; Muir 1994; Schaefer et al. 2016; Sheridan 2019a). As a cognitive process, trust has a long-term tendency that is relatively static unless it is broken (Jarvenpaa et al. 1998; Mayer et al. 1995). This dispositional trust is shaped by one's personality, culture, and experiences and shapes one's general attitude towards automation regardless of the system (Hoff and Bashir 2015). Trust is also a dynamic cognitive state that evolves and changes during systems operation. This dynamic trust is determined by the situational context in which the interaction takes place (e.g., task complexity, workload, organizational setting) and by the learned trust through interactions with the system itself (e.g., reliability, false alarm rates, ease of use) (Hoff and Bashir 2015). In this regard, trust can be defined as a reactive and transient short-term mental state when interactions occur momentarily (McAllister 1995; Merritt and Ilgen 2008; Schaefer 2013). Therefore, trust can change and become enhanced or degraded over time (Desai et al. 2013; Schaefer 2013; Wilson et al. 2016), revealing its dynamic nature over repeated interactions or experiences.

Other models that look at trust in online social networks and decision support systems categorize the factors affecting trust into individual trust attributes and relational trust attributes (Cho et al. 2015; Lai et al. 2020a, b). Individual trust refers to constructs that are traced to the human's own characteristics, whereas relational attributes are derived from factors involving relationships. Individual trust attributes are further divided into logical trust and emotional trust. Logical trust is based on cognitive processes from interacting with the trustee (e.g., beliefs, confidence, risk, bias, experience, reliability). Emotional trust is influenced by a person's feelings (e.g., fear, hope, frustration). Relational trust factors include items like the similarity between people and the importance of individuals.

Similarly, Hancock et al. (2011) developed a model of trust in robotics that included human factors, robot factors, and environmental factors. Their model suggested that robot factors accounted for the most in the development of trust, followed by environmental factors, and almost no effect on the human factors. For the robotic factors, trust was most related to the robot's performance and less upon attributes such as its appearance. Subsequently, Schaefer et al. (2016) expanded this three-factor model to automation in general and developed a human-automation interaction (HAI) model. Again, this model stressed the importance of automation-related factors in trust development. It is found that human factors played a larger role in general automation trust, but that there was a paucity of research examining environmental factors (Schaefer et al. 2016). Robot and human factors of the HAI model mirror two of the three "context constraints" (technology, human, and system functionalities) identified by Hou et al. (2014) and further explained by Baber (2017) in his book review on the development of the human-autonomy collaborative partnership. The environmental factor is highly relevant to system functionalities based on application context (Section 3.2 provides an example of functional requirements based on an operational context, mission type, force structure, etc.). The relevance has also been discussed for designing more advanced autonomous systems such as intelligent adaptive systems (IASs) (Hou et al. 2014).

These models of trust in automation, while comprehensive, may not consider dimensions related to AI and IASs. AI technologies within IASs are able to adapt their support (to human) in a manner that is sensitive to both the external world and the internal (i.e., state of the human and machine) (Taylor and Reising 1995). Autonomous systems like IASs arguably differ from traditional automation because these systems have degrees of self-governance, learning, freedom of decision-making, and possibly free will in responding to the requests in extreme conditions within dynamic and indeterministic or uncertain contexts (Hou et al. 2014). Another way to think about traditional automation versus autonomy is in terms of Parasuraman et al. (2000) model of the levels of automation. This model suggests that automation replaces tasks at varying levels across the human information processing process, namely, information acquisition, analysis, decision-making, and actions. It improved the flexibility of Sheridan and Verplank's ten levels of automation (Sheridan and Verplank 1978) with the convenient decomposition of task performance, but still at a very coarse-grained conceptual level with a low resolution. Traditional automation has typically involved only using the machine to support humans in one of the four categories and commonly at lower to moderate levels of automation. However, autonomy allows the machine to perform tasks for all four of these constructs at high levels of automation.

The evolution of machines from relatively simple tools to automation to complete autonomy (i.e., static automation, flexible automation, intelligent adaptive automation, Hou et al. 2014) affects the trust relationship between the human and the machine (Yagoda 2011). For simple tools, trust is defined almost solely by the tool's performance. As machines incorporate greater intelligence with higher degree of self-governance and decision-making, the factors that affect the trust relationship become more complex. As machines evolve into highly autonomous systems (i.e., IAS) with greater AI, Sheridan (2019b) argues that the trust relationship will start to more closely mirror that of human to human trust. Siau and Wang (2018) suggest that the performance, process, and purpose of AI are different and more complex than previous technologies, and thus the factors that influence human trust in AI is also more complex. Hou et al. (2011, 2014) suggest that the relationship between human and AI agents should be built up based on aspects of human-human interactions to reflect dynamic and complex nature of human-autonomy teaming. It is evidenced by a study of the trust relationship with human team performance (McNeese et al. 2019). The study found that there was a lower level of trust in the autonomous agents in the low human performing teams than both medium and high performing teams, while there was a loss of trust in autonomous systems among human teams at all three performing levels over time. In fact, some have argued that true trust is only attained when the agent itself has the free will to enter the trust relationship (Abbass et al. 2016). It touches on the essential topic of human-machine interaction: roles, responsibilities, and authorities (Hou et al. 2014).

However, the free will for a machine to have greater self-governance and decision-making increasingly invites the uncertainty, vulnerability, and risk of the trust relationship. The risk is characterized by what Onnasch et al. (2014) call the lumberjack effect. That is, when a machine that operates at higher levels of automation fails, the consequences of the failure are more severe. The severity of consequences is categorized by Vicente (1990, 1999), Miller (2000), and Hou et al. (2014) for what they call the coherence and correspondence domain applications. For coherence domain applications such as a piece of document process software, the consequence of automation failure may not be severe. However, for correspondence domain applications such as safety and/or mission critical systems, the consequence can be catastrophic (e.g., recent Boeing 737 Max accidents) (Marks and Dahir 2020).

The rise of IASs and the employment of AI also raise a number of ethical concerns that have not been previously considered when examining trust in traditional automation (Awad et al. 2018). The most notable example is an autonomous vehicle in an imminent collision situation and must decide whether fatally injure its occupants or pedestrians (Awad et al. 2018). Thus, we must now consider the trustworthiness of a machine not only on its technological capabilities or task performance but also whether the outcomes are ethically sound. System designers need to consider specific policies of AI and autonomy and address associated complex legal, ethical, moral, social, and cultural issues (Hou et al. 2014). It also reminds all stakeholders that a set of global standards for AI and IAS systems are needed.

One of the problems with AI models (e.g., deep learning and machine learning models) relates to the complexity and opacity of their outcomes such that AI decisions and actions are unexplainable and might appear illogical to humans (Rahwan et al. 2019). Hence, to maintain trust, there needs to be an increased focus on transparency, communication, and shared mental models of intent which are critical for building human trust in autonomous systems. To address the concern of AI opacity, the field of explainable AI (XAI) has emerged to develop techniques and AI models that allow for increased explainability, interpretability, and transparency of AI. Research under the Defense Advanced Research Projects Agency's (DARPA) XAI program (Arrieta et al. 2020; Gunning and Aha 2019), and projects under the Fairness, Accountability, and Transparency Machine Learning (FAT/ML) working group (fatml.org) continue to advance the knowledge in this area (Adadi and Berrada 2018).

The widespread adoption of AI and autonomy introduces new challenges that can impact trust. There now exists new threats to our digital security (e.g., speech synthesis for impersonation), physical security (e.g., using autonomous systems to attack), and political security (e.g., mass unwanted data collection) (Brundage et al. 2018). With the emergence of the Internet of Things (IoT) that promises various exciting applications from the power grid to smart cities and to networked autonomous vehicles, security has come to the forefront of the system design process. The risk of cyber-attacks is a major concern and a critical design challenge for autonomous systems when attackers penetrate the network from anywhere in the world (NIS Cooperation Group 2019). Moreover, when data-driven machine learning techniques (e.g., deep neural network) are used for object classification and threat assessment in sensor images, autonomous systems are vulnerable to a wide variety of attacks (Computing Community Consortium 2020; Szegedy et al. 2013; Vorobeychik and Kantarcioglu 2018).

Attacks to AI can take on many forms. For example, poisoning attacks introduce noise or very specific but small deviations into training data (such as changes in pixels) before a model is learned, thus affecting its subsequent performance by introducing bias into the system or causing the model to misclassify or make poor decisions (Vorobeychik and Kantarcioglu 2018; Zhang and Dafoe 2014). Evasion attacks attempt to avoid detection and try to mask malicious code to be classified as benign or safe. If a model is known, attackers may know how to manipulate it or "fool" the classifier. These so-called white-box attacks might focus on altering the state of the environment to influence the system to misinterpret a state (Vorobeychik and Kantarcioglu 2018).

Therefore, due to the dynamic nature of trust, adaptive behaviors of the autonomous system and measurable performance of the actions are required for predictability and reliability that will instill trust during IAS operations with different contexts. Sufficient levels of communication, transparency, and security are needed for trust development. Thus, in this paper, we provide not only a conceptual but also practical trust model for system designers to build trustworthy autonomous systems. This new trust model IMPACTS has been developed to address shared *intention*, performance *measurability*, predictable and reliable *performance*, context *adaptivity*, bidirectional *communication*, optimal *transparency*, and protective *security*. The next section explains what the IMPACTS model is, and followed by an example of the construct of an IAS as a decision aid guided by the IMPACTS model, and then followed with evaluation results in a recent field test of the entire system.

2 IMPACTS trust model

With their advanced capabilities and high degrees of self-governance, learning, and freedom of decision-making, AI and autonomous systems are becoming more and more capable and exhibiting leadership in more areas. Thus, human trust in these types of technologies should closely resemble that of human to human trust (Sheridan 2019b). For humans, strong leadership depends on trust that is a function of capability and integrity. Trust in technologies also cannot exist without any of these two variables. Technologically, there is no doubt that those autonomous systems with greater AI capabilities are more capable than their human partners in certain areas. The question is how to build up their human-like integrity to gain actual trust and demonstrate their true leadership? To exhibit integrity for autonomous systems, intention, measurability, performance, adaptivity, communication, transparency, and security are considered to be seven essential elements and building blocks of the IMPACTS model. The IMPACTS model is aimed to guide the construct of user's trust in autonomy when designing a human-agent partnership and addressing the three context constraints (i.e., human, technology, and environment) as well as six trust barriers discussed in the last section. These agents are intelligent and adaptive enough that they can "think for themselves" and can often conduct tasks on their own with their afforded authority (Hou et al. 2014). In some cases, they also decide which task is best suited to achieve the goal for themselves. Figure 1 illustrates the defining characteristics of the seven IMPACTS elements in this context.

2.1 Intention

Regardless of what you mean to do or hope to happen, your behaviors speak louder. It is critical that behaviors are aligned with one's intentions and that those intentions are properly communicated (Schaefer et al. 2017). By doing so, others can infer the motives of the agent based on the behavior and make accurate judgments of trust. A collaborative partnership must have the desire to support each other. The chosen desires are defined as intentions to which the agent commits resources for achieving the goals to help its human partner (Hou et al. 2014). Poor understanding of common goals is a potential risk for human-agent teaming. The recent Boeing 737 Max accidents demonstrate how disasters can happen when the human wants to take one action, but the machine agent is attempting to execute another (Endsley 2019). The way that the technology partner (e.g., adaptive intelligent agent (Hou et al. 2014)) is designed affects the performance of its human partner.

Understanding common goals and how to optimize the relationship between humans and technological systems is at the core of human-autonomy teaming. The agent must know what its human partner is trying to achieve so that it can pursue achieving its intention to help. Meanwhile, understanding the agent's supportive intention towards a common goal serves as a starting point for a human to trust his/her agent partner.

Furthermore, the ability of IAS to communicate intent ties into judgments of its ethical and legal behavior. Technology has matured to the point that it is no longer dependent on human involvement to perform activities such as driving cars or tracking and firing at enemy targets. IAS designers must now consider "should we?," rather than "can we?". The tension between what is possible and what is acceptable in terms of the range of capabilities and functionalities in IASs requires careful consideration of both ethical and legal issues; focusing

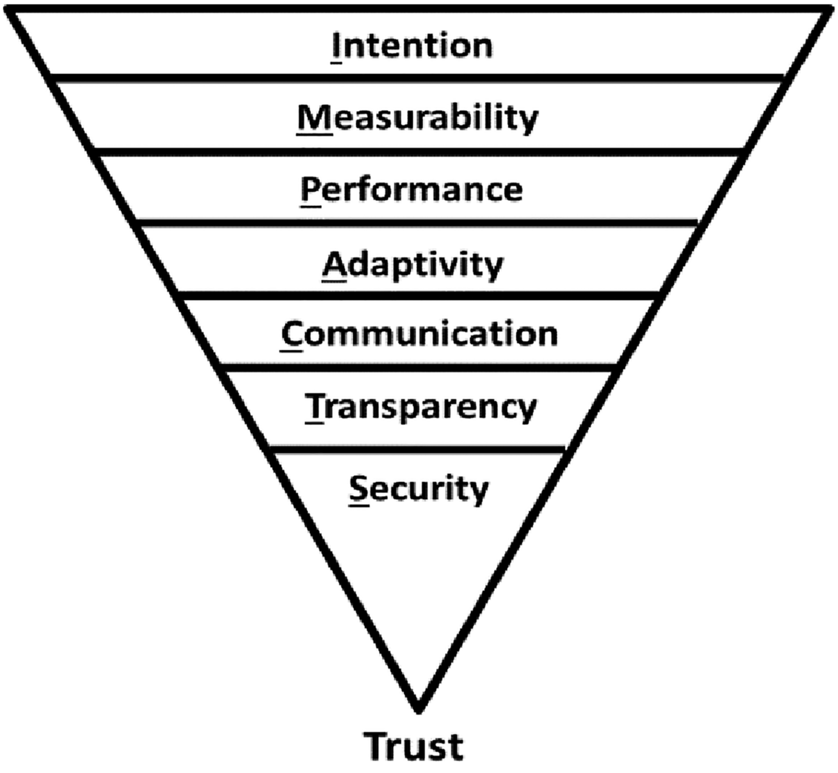


Fig. 1 IMPACTS: a conceptual and practical trust model

solely on technological and operational perspectives can lead to user rejection and/or severe consequences (Hou et al. 2014). For example, the ethical and legal ramifications of unmanned robotic systems can be extensive, and the British Broadcasting Corporation (Hughes 2013) cites a campaign calling for an international ban on "killer robots" due to concerns about the ethical and legal implications of fully autonomous drones. That is why Murphy and Woods (2009) proposed amending two of the three original Isaac Asimov's "laws" in his collection of short stories *I*, *Robot*, although they still reflect human understanding and expectation regarding the relationship of humans and autonomous machines. Given the legal and ethical responsibility of those who design and deploy autonomous machines and the complexity and dynamics of relationships between human and machine, Murphy and Woods (2009) propose two changes. First, *a human may not deploy a robot without the human-robot work system meeting the highest legal and professional standards of safety and ethics*. Second, *a robot must respond to humans as appropriate for their roles*.

The relationship between intent and ethics is commonly discussed as the principle of double effect. This principle states that an agent may cause or allow for a negative outcome as long as there is no malicious intent and that the negative outcome is proportional or less than to any anticipated positive outcome (Shaw 2006). Thus, understanding the intent of AI or autonomous systems will allow humans to judge its ethical and legal behavior and develop trust in their ethical behavior. It is also related to social relationships between autonomous systems and their human partners. Generally speaking, these IASs are not benign tools; they can engender social relationships. In fact, many researchers have deliberately engineered knowledge about social relationships into the design of intelligent systems so that they are sensitive to individual cultural, moral, social, and contextual differences and follow good etiquette based on social norms that apply to the user (Sheridan and Parasuraman 2006). Miller et al. (2007) also suggest that a wide range of social interactions could be modeled using quantitative computational tools.

2.2 Measurability

"We judge ourselves by our intentions and others by their behavior" (Covey 2008). Until we can read others' minds, we can only infer intentions. The inference may come from words, actions, and patterns. Hence, the development of trust is a process, and it has been said that it takes a lifetime to build trust, but only an instant to destroy it. Given the capabilities and limitations of technologies, trust should not be undermined by a single instant. You may not trust words or even question actions, but should not doubt patterns. Therefore, we can observe others' behaviors, measure their actions, and analyze their patterns. The same philosophy for human-human relations can also be applied to the human-agent relations where an agent may be a complex and opaque entity, especially when the AI is part of the entity. The agent system must be measurable so that agent's behaviors can be observed, its actions can be measured, and its patterns can be analyzed to gauge its intentions. Trust is something that is earned through observable behaviors or measurable actions or analyzable patterns. Then the results and/or implications can be judged to be helpful or not. The concept of trust can be separated into trust, the intention or attitude to be vulnerable to another agent, and trustworthiness, the perceived capability and/or integrity of the trustee agent. Trustworthiness largely depends on reliability (Hancock et al. 2011; Schaefer et al. 2016). If the machine does not work, how can a soldier, sailor, or aviator be expected to hand over a task to it that may mean life or death? It is essential to establish trustworthiness at the design time and provide adequate indicator capabilities within the agent system so that inevitable context-based variations in operational trustworthiness can be assessed and dealt with in real time (Defense Science Board 2016).

There are at least three ways that the measurement of trust can be applied to understand the interaction with agents. First, most commonly, trust towards the machine agent is measured to understand the degree of trust the human has towards the agent (e.g., Jian et al. 2000; Schaefer 2016; Yagoda 2011). Second, supported by a theoretical framework of autonomous systems underpinned by cognitive, intelligence, and systems sciences, some measures also try to capture the trustworthiness of a machine agent through mathematic models based on their performance and other objective qualities of the agent (e.g., communication and transparency) (Wang and Singh 2010; Wang et al. 2020a, b). Measures of trustworthiness may aid in the design and development of systems to optimize the machine itself. Third, real-time bi-directional measures of trust need to be explored so that adaptive trust can be allowed in a way that humans and IASs can dynamically calibrate their trust in one another (Awad et al. 2018; Sheridan 2019b). All these possible measures should be considered during the design process of the system for the dynamic process of human-autonomy interactions.

2.3 Performance

Trust is never given. It must be built, earned, and maintained. The establishment of trust results from the reliable, consistent, and predictable performance of an agent over time to meet the goals of the human. In fact, performance was identified as the primary contributor for establishing trust for robots (Hancock et al. 2011) and significant for the development of trust for automation (Hoff and Bashir 2015; Schaefer et al. 2016). Performance entails a variety of attributes. For trust to be gained, the agent must demonstrate performance that is reliable (consistency over time), valid (the agent performs as intended), dependable (low frequency of errors), and predictable (meeting human expectations). Predictability is central to establishing trust. For humans, the greatest challenge an infant has is to determine whether or not he/she should trust her/his parents, and the greatest determinant of that is their predictability (Erikirk 1993). Thus, unanticipated actions from machine agents can often lead to rapid declines in trust.

Research on trusted performance has also shown that some nuanced agent behaviors can impact trust. For example, human's bias affects trust in decision-making (Lai et al. 2020a, b), and the suitability of cues and feedback has shown to affect trust (Schaefer et al. 2016). As well, alarm sensitivity affects trust. Too many alarms have been shown to negatively impact trust (e.g., the design of control panel used at the Three Miles Island nuclear station, Hou et al. 2014). The type of error (i.e., false alarms vs. misses) appears to affect trust differentially. False alarms are more salient but can instill greater trust than miss prone automation (Hoff and Bashir 2015). The key factors determining if trust degrades when agents commit errors are the complexity of the task and the consequences of the error (Schaefer et al. 2016). Thus, in the context of human-agent teaming, agents need to exhibit consistent, reliable, and predictable behavior and interactions with its human partner over time across different situations. Of course, predictable protocols and/or reliable service standards are needed to be established and communicated well with the common understanding of shared team goals or intents.

Further, system performance, including both agent and human performance should be considered at the same time when designing a human-machine system. An agent's behavior and performance should not be the cause of human-automation interaction issues but instead, provide aid to address issues like situation awareness (SA) loss, loss of skills, overtrust (complacency), or undertrust (skepticism) as discussed by Hou et al. (2014).

2.4 Adaptivity

In the context of human-agent teaming and due to the lack of human sensing, reasoning, and thinking for the agent, it is challenging to have its self-awareness, or perceive, understand, make decisions, and act on different contextual assumptions of the operational environment. Especially with dynamic changing environments and human mental states, the agent needs to have the capability to learn and understand its human partner's intentions, the changes in the environment, the system status, monitor the human cognitive workload and performance, and guard the human resources and time, and then change its course of action to help the human achieve the team's common goals. An agent that exhibits these adaptive and intelligent characteristics can then be a trusted partner to build a truly collaborative human-agent partnership. Thus, agent adaptivity needs to be demonstrated through its three defining characteristics: adaptation, autonomy, and cooperation in accomplishing tasks for its human partner (Jansen 1999). Such an adaptive intelligent agent (AIA) is then defined by Hou et al. (2011) as a personification of computer code and algorithms that mimics human behavior, perception, and cognition that can cooperate with other agents, that automatically act either autonomously or on behalf of its human partner, and that can adapt to changes in the human, system, or environment. An AIA is not a "normal" computer application. For example, a standard macro would not be classified as an AIA. Macros automate tasks for its user; however, they are usually dependent on the working environment and user inputs. Any deviation from the initial input or changes to the environment can cause the macro to fail. Therefore, agent adaptivity is defined as its ability to take actions either autonomously or on behalf of its human partner and adapt to the changes in the human, system, or environment (Hou et al. 2011, 2014).

One area related to adaptivity that has not received attention is the idea of adaptive trust from an agent (de Visser et al. 2018). In human relationships, when trust is broken, the offending partner can alter their behaviors to try to re-establish trust. AIAs may adapt and try to re-establish trust if it is broken by altering their behavior accordingly. This can be done in multiple ways. It could increase or decrease the frequency of communication or alter the form of feedback to provide additional details. Detection thresholds can also be dynamically changed to adjust for error rates. Moreover, an AIA could also apply more rigorous and computationally complex algorithms to improve accuracy.

Further, agent adaptivity should be measured through their performance based on the exhibition of descriptive, prescriptive, intelligent, adaptive, and cooperative characteristics (Hou et al. 2014). Agents should be able to inform operators about what is happening (i.e., being descriptive) and explain and specify what will or should be done next (i.e., being prescriptive); they should be able to learn from past experiences about human intentions; monitor the system, the environment, human workload, performance, and timeline to understand the situation (i.e., being intelligent); and, in response, change how they behave in any given situation (i.e., being adaptive); and they should enable themselves to communicate and cooperate with each other and act in accordance with the results of their communication (i.e., being cooperative). For example, an in-car GPS is able to learn from its satellite agent partner that the car is approaching a traffic jam (i.e., being cooperative); it can inform the driver of the situation (i.e., being descriptive) and advise the driver to take a different route (i.e., being intelligent and prescriptive). If the GPS had the ability to work with other in-car speed agents to detect that the car was traveling in a wrong direction (i.e., not consistent with previously learned driver intentions) or was going too fast for the driver to stop, it could then work with an agent controlling the brakes to stop the car automatically without asking the driver for authorization (i.e., being intelligent, adaptive, and cooperative). Overall, agent adaptivity is a key trust attribute and should be measurable through its performance. System designers should consider how to exhibit the adaptivity into their design and development process to facilitate human trust in their agent partner during their continuously complex and dynamic interactions.

2.5 Communication

Communication is a specialized type of behavior that uses ideas, words, sound, or even odor to convey intentions. Communication speaks to how teammates understand each other and how information is transferred in the team. However, communication issues are common in human-automation interaction. Human operators require feedback from automated systems so that they can understand what the automation will do (Olson and Sarter 2000), but automated systems do not always provide the types of feedback that operators would like. Automated systems perform better if they are told what humans intend to achieve, and not about the working environment (Harbers et al. 2012). A common issue demonstrated by automated systems is conversational inflexibility, which occurs when automated systems and agents are unable to react to the information of queries being offered by human participants. For instance, automated phone systems often require callers to listen to all options before he/she makes a decision. This is not practical and acceptable in the heat of battle for mission critical systems and will damage the trust of human in technologies regardless they are automated or autonomous systems. However, recent advances, such as Google Duplex, show promise in reducing conversational inflexibility in future systems.

Automation and autonomy are becoming increasingly prevalent as humanity continues to traverse the information age. As digitized systems become essential parts of military toolboxes and as machines slowly turn into independent entities that continue to replace and support humans in a variety of tasks, issues regarding human-machine interaction come to the forefront. Sheridan (2002) comments that as the frontiers between automation and operators blur, it becomes "increasingly critical" that automation designers realize they are building not only technology but also relationships. It is even more important for autonomous system designers to build a trusted relationship between human and autonomy through proper and effective communications.

Thus, it is critical to make good use of the communication tools, aligned with the intentions, to demonstrate system trustworthiness. An effective human-machine interface (HMI) should enable the autonomous agent to clearly, fairly, and directly explain and justify its intention, its actions, and its desired end states to its human partner on how it helps reach the common goals. These are critical characteristics that an agent needs to exhibit to build up human confidence and trust. HMI also needs to be flexible and offer the types of feedback its human partner would like so that effective communications happen at the right time, in the right format, through the right channel, and to the right recipient. Thus, autonomous system designers or developers need to understand that they are building not only technology but a partnership and HMI is the means between the two partners: human and autonomous system. If an HMI could facilitate effective communications, an appropriate trust partnership between human and autonomy would be enabled and maintained constantly.

2.6 Transparency

It is often noted that automation occurs in a black box. That is, it works in a fashion that the operator does not fully understand and has no way to validate. While this has remained acceptable when the system is reliable and is designed for simple tasks, the opacity of advanced agents is more problematic for trust. Autonomous systems with AI capabilities are expected to perform complex tasks, involving multifaceted decisions in dynamic and uncertain situations alongside human teammates with potentially vital consequences. Thus, in order to be trustworthy, autonomous systems need to be able to communicate and rationalize their actions so that the human can ascertain that their goals and methods for achieving those goals are aligned. This is particularly true if the human and the agent are working in different spaces, with different assumptions and mental models, or each has some unique information not directly available to the other.

Therefore, agent transparency is necessary to support trust (Chen and Barnes 2014; Verberne et al. 2012). Transparency is typically provided through the HMI, which communicates real-time information to its human partner about its intentions, goals, reasoning, decisions, actions, and expected outcomes. However, transparency can also be provided through indirect means, such as through observable behaviors and/or measurable actions; transparency can also be supported through effective communication, providing understandability of agent behaviors and predictability of its future actions.

Chen and her colleagues (Chen et al. 2016, 2018) provide a model under which transparency can work called the situation-awareness-based transparency (SAT) model. Following Endsley's three-stage model for SA (Endsley 1995), SAT focuses on providing the operator with information regarding (a) the current status of the agent, its actions, and plans; (b) the agent's reasoning process; and (c) the agent's predictions and uncertainty. Each stage of SAT is designed to provide the human operator with information to allow for each stage of SA, perception, comprehension, and projection.

Several studies demonstrate how added transparency facilitates trust in the system. For example, Helldin (2014) found improved trust when humans were provided with information regarding sensor accuracy and uncertainty, but at the expense of workload and decision time. Similarly, Chen et al. (2016) provided participants with greater transparency by applying the SAT model to the user interface of an autonomous system and found greater trust when the agent presented all three levels of SAT information.

Therefore, in order to build trust, the human partner needs to develop an appropriate mental model on agent intentions, reasoning, behaviors, and end states. Transparency helps achieve this goal by providing display different levels of information with reduced visual complexity (e.g., density, grouping, format, layout) while providing sufficient details like task complexity (e.g., number of paths, number of possible end states), conflicting interdependencies, and uncertainty in linkages. With the changing situational factors in the human, system, and environment, transparency is also needed to convey ongoing feedback to the human so that he/she maintains an accurate mental model of the events and system (Hou et al. 2014). Through transparent interactions, dynamic and appropriate levels of trust can be developed.

2.7 Security

Trust in AI and autonomous systems also depend on its security or its ability to remain protected from accidental events or deliberate threats. For example, military autonomous systems act as force multipliers, and AI has dramatically improved many defense capabilities, including cyber-security, intelligence, surveillance, and reconnaissance (ISR) and navigation systems (Mae Pedron and Jose de Arimateia 2020). Yet these systems are commonly networked, which invites opportunities for adversaries to attack the system itself, its sensors, or to manipulate their inputs or their models in the hopes of encumbering their performance. These inputs tend to come from a variety of sensors (e.g., radar, EO/IR) originating from multiple distributed locations, thus opening the door for multiple points of vulnerability. Therefore, strong and robust security is often a necessary condition that must be met for a trusted relationship between humans and autonomous systems. However, conveying trustworthy security can be difficult. Security tends to occur in the background, often occurs in a black box, and is typically not in the forefront of mission or task goals.

As discussed early in the Introduction section, there are a variety of AI-induced potential attacks, such as adversarial patterns, misleading examples, and data poisoning. It is a critical concern to special classes of weapons systems that use sensor suites and computer algorithms to identify and engage a target without a human in control (Mae Pedron and Jose de Arimateia 2020). A secure system must behave as designed and implemented following rules or laws even when under attack; otherwise, it cannot gain human trust.

To counter attacks, many traditional techniques can be applied, and new ones are being explored to provide for stronger security. For instance, to guard against poisoning attacks, methods such as data sub-sampling, outlier removal, and trimmed optimization can better ensure model and data integrity (Vorobeychik and Kantarcioglu 2018). More recently, federated learning, whereby machine learning models are optimized through multiple decentralized devices, protects privacy by avoiding explicit sharing of data (Yang et al. 2019). Another approach that is gaining recognition is the use of blockchain technology to handle data storage and exchanges (Pilkington 2016). Blockchains, first developed to ensure the safe and private transactions of Bitcoin, use a distributed recording method and immutable records of every transaction. The other advantage of blockchain is that it offers transparency since the transactions are distributed and open to everyone, thus establishing a means to provide accountability and traceability. Blockchain technology has been suggested for other types of transactions, such as sharing of medical records by ensuring the integrity of privacy logs (Sutton and Samavi 2018). It has also been suggested as a way to provide explainable AI since it offers transparency, openness, and traceability (Nassar et al. 2019).

However, building trust in security is more nuanced than simply providing strong security. Peiters (2011) differentiates between explanations-for-confidence and explanations-for-trust. Confidence refers to self-assurance without considering the risks or alternative options. In these situations, there is no choice in the matter, but people still need to have confidence that the system will work as intended (e.g., confidence in electricity supply or the rules of traffic). In contrast, trust involves self-assurance based on a decision to rely on an agent, human, machine, or organization.

Peiters (2011) argues that for information security, systems need to provide explanations-for-confidence, as opposed to explanations-for-trust. In the case of security, explanations of confidence do not need to provide the inner workings of the black box but instead should focus on how the user is protected from adversarial threats. In contrast, explanations-for-trust delve into the inner workings of the black box and are better for explaining AI decisions. In fact, he suggests that providing too much detail in explanations for security may be counterproductive to the goals of the explanation and may actually be detrimental to trust. He also suggests that security explanations do not have to reside in the system itself but *is the role of the designers* or the business strategy of an organization to provide explanations-for-confidence.

Similar to Hou et al.'s W5+ for IAS design principles (2011, 2014), Vigano and Magazenni (2018) build upon Peiters (2011) work and provide a framework for explainable security based on W5+ of security: who, what, where, when, why, and how. The framework suggests explanations for security consider not only who the messages are directed to, but what information is provided, where and when they should receive it, why it is necessary, and the medium of the message.

With the growth of AI and autonomous system technologies, adversaries and agents with malicious intent will continually look for new methods to manipulate the performance of these technologies. Strong security measures are required to establish trust so that people can benefit from these systems. However, robust security measures are not enough to gain confidence from all stakeholders. Designers and organizations need to build confidence for autonomous systems and AI technologies by providing goal-directed explanations of how security measures are in place (i.e., at the right level of detail) to protect and ensure the performance of the system. Therefore, security becomes an essential character of IMPACTS model to guide designers for enabling trust in autonomous systems such as IASs.

Trust is the vertex and essential ingredient in effective relationships including human-autonomy partnerships. Trust is the careful balance upon which healthy relationships grow between the partners when considering physical, intellectual, emotional, relational, and even spiritual aspects of human-human relationships. To truly be trustworthy, to be consistent, predictable, reliable, and demonstrate human-like integrity with shared intentions accurately through its adaptive behavior and measurable performance, transparent communications, and secured protection are indeed IMPACTS that only the human can make with the agent partner. Engineers, researchers, and technologists developing autonomous systems must carefully design them to inspire confidence and build trust, and the IMPACTS model is a conceptually practical tool to guide the design and development of effective human-autonomy teaming.

3 Authority pathway for weapon engagement—an example of IMPACTS-based design

A target engagement process is often complex, lengthy, and error-prone. It is even more challenging if a combat team does not have sufficient doctrinal knowledge on rules of engagement (ROEs) and international laws of armed conflict (LOAC) when operating in a foreign country. These issues caused the loss of situation awareness (SA) and consequent mission failure when an unmanned aircraft system (UAS) crew engaged a target during a few trials conducted at Defense Research and Development Canada (DRDC) (McColl et al. 2016).

To assist the UAS crew in doctrine training and provide them decision support with shared SA to succeed their missions, an intelligent adaptive decision aid called Authority Pathway for Weapon Engagement (APWE) was developed to ensure UAS crew to make critical target engagement decisions, maintain mission SA, and follow ROEs, LOAC, and standard operating procedures (SOPs). APWE is a good example of a human-autonomy teaming to resolve a military operational issue. The development of APWE followed the design principles based on the IMPACTS model to demonstrate its characteristics of intention, measurability, performance, adaptivity, communication, transparency, and security.

3.1 APWE concept

The goal of APWE is to support the UAS crew by guiding them through the required steps and permissions needed to conduct a lawful and successful engagement of a target following a positive identification (PID). Specifically, the APWE supports a use case of a pattern of life (POL) operation during which a person of interest (POI) is positively identified and then engaged following the necessary permissions being granted from a tasking authority (TA). This process represents the most complex use case of the APWE insofar as it might take many minutes or hours to complete and involves significant and complex decision-making with respect to the ROEs in effect for the mission and applicable military laws.

Figure 2 depicts the logic and flow of the ten steps required to undertake a weapon engagement on a target. The process essentially commences once a POI has been detected, localized, and retained as a potential target. The subsequent steps involve completing a PID checklist, notifying the TA and requesting authorization for the use of a kinetic response; performing a collateral damage estimate (CDE) while simultaneously completing the weapon engagement planning; notifying the TA of the UAS readiness to engage and requesting weapon release authorization; lazing the target (as required); and releasing the weapon. Maintaining continuous eyes on the target is a requisite parallel activity that must occur throughout the target engagement process. Battle damage assessment (BDA), not shown in the figure, is performed subsequent to the weapon release.

During a sortie, the UAS may take on different roles. For instance, in the case of a call for direct fire, the UAS may be responsible for the targeting and weapons release, whereas in the case of a call for indirect fire, the UAS crew may only be providing a target lazing capability for another manned or unmanned asset. The APWE seeks to address two critical functional areas of the target engagement process: (1) maintaining a shared mission SA among all UAS crew members and (2) ensuring that all applicable, relevant ROE and LOAC are considered in a timely manner and that the SOPs are followed (i.e., built-in capabilities and integrity). One of the underlying premises of the APWE concept is that the judicious use of IAS technologies (intelligent adaptive automation and intelligent adaptive interface, Hou et al. 2014) applied to the UAS ground control station (GCS) HMI will support the individual and collaborative decision-making during the target engagement process. The goal is to facilitate effective interactions and positive relationship between the UAS crew and the embedded IAS agents thus foster trust.

3.2 Design challenges and requirements

The lack of a common mission SA was identified as one of the key issues contributing to the mission failure during an analysis of a few human factors experiments conducted on various UAS missions previously (McColl et al. 2016). To alleviate this issue, a common/shared target engagement SA display should be provided to the crew. Additionally, target engagement procedures can vary as a function of the operational context, the type of mission, and the force structure. The APWE needs to operate in an environment in which operational and emulated command and control (C2) systems will

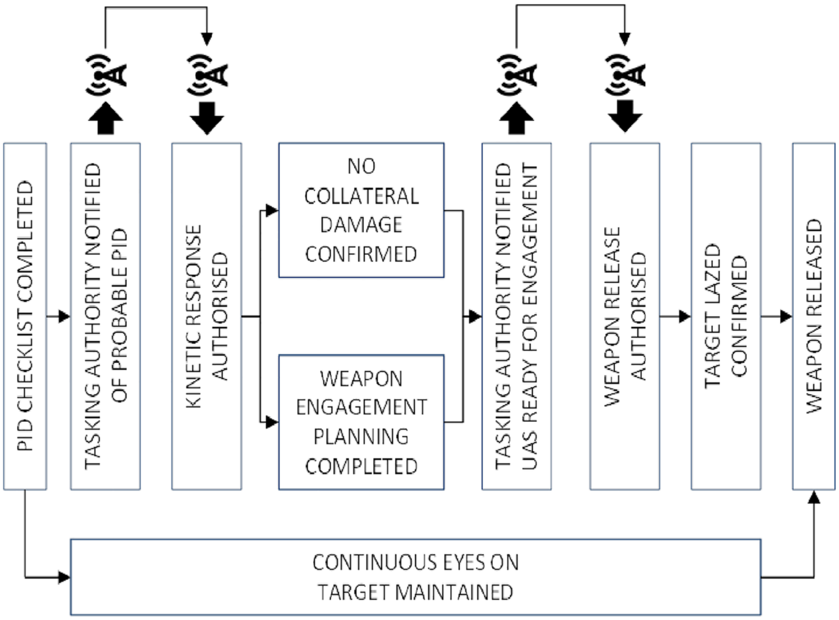

PID CHECKLIST COMPLETED
TASKING AUTHORITY NOTIFIED OF PROBABLE PID
KINETIC RESPONSE AUTHORISED
NO COLLATERAL DAMAGE CONFIRMED
WEAPON ENGAGEMENT PLANNING COMPLETED
TASKING AUTHORITY NOTIFIED UAS READY FOR ENGAGEMENT
WEAPON RELEASE AUTHORISED
TARGET LAZED CONFIRMED
WEAPON RELEASED
CONTINUOUS EYES ON TARGET MAINTAINED

Fig. 2 Ten steps of Authority Pathway for Weapon Engagement required releasing a weapon based on a positive identification (steps requiring external authorization from the tasking authority are depicted by a "transmitter" icon) (Original by the authors)

be integrated. It needs to support a variety of missions that will result in a specific instantiation based on mission templates. The fog of war, the unpredictability of military operations, and the dynamic nature of UAS missions are all reasons why these missions are subject to change, sometimes referred to as dynamic re-tasking, i.e., a change in mission during a sortie. During dynamic re-tasking, the system is required to reconfigure itself to meet the evolving mission requirements. Basically, the APWE needs to have C2 agility, which is the capability of a system to successfully affect, cope with, and/or exploit changes in circumstances. Agility enables entities to effectively and efficiently employ the resources they have in a timely manner (NATO STO SAS 085 2013).

The so-called agility enablers are responsiveness, versatility, flexibility, resilience, adaptiveness, and innovativeness (NATO STO SAS 085 2013). The APWE needs to have C2 agility requirements since the target engagement requires many of these enablers. In particular, the responsiveness and adaptivity of the system will impact the efficiency with which requests to the TA can be communicated and processed. Current target engagement systems are limited by relying too heavily on radio communications and chat for UAS operations. For example, these enablers are important in the case when a troops-in-contact (TIC) mission pre-empts an ongoing POL mission. The system must be adaptive and react to this change by presenting a modified instance of the APWE that reflects the relevant set of applicable ROE and is consistent with the procedures and time-critical operational context. The APWE also needs to automatically detect this change in mission by interpreting UAS system data (e.g., C2 systems, sensor), or the APWE ROE could be changed manually by the UAS crew. These changes in ROE would result in a new or modified APWE display to inform the crew of the changes.

Therefore, the design of APWE needs to address at least the following challenges based on IAS framework (Hou et al. 2014):

- The system must adapt to the mission type and UAS role, both of which may vary over the course of the sortie.
- The system must provide user-specific views (e.g., UAS pilot, payload operator, intelligence analysts, external authorities, other users).
- The system must allow for the evolution of the level of automation consistent with future UAS requirements for additional semi-automated crew functions related to target engagement.

The need for C2 agility of the APWE is further compounded by the requirement for UAS collaboration scenarios wherein one UAS asset, for instance, may provide a target lazing capability for another manned or unmanned asset that is providing a weapons fire. Collaborations among assets, whose crews are not co-located, present specific coordination and communication challenges related to maintaining mission SA while executing critical decision-making. For example, timely communication and coordination are vital to ensuring the quick response times required in the case of time-sensitive targeting (TST), in both deliberate and dynamic targeting scenarios while considering CDE, ROE, airspace, and other restrictions during the targeting process.

To address these challenges and meet those high level system requirements, a cognitive task analysis was conducted, resulting in a decision ladder for the UAS target engagement tasks. The results were used to identify areas where IAS technologies could be applied for both automation and HMI supports (McColl et al. 2017). The decision ladder identified as pertinent to the APWE included two possible types of automation requirements: (1) intelligent adaptive automation (IAA) requirements and (2) intelligent adaptive interface (IAI) requirements (Hou et al. 2014; McColl et al. 2016; McColl et al. 2017). IAA requirements are related to behind the scenes calculations and the process performed on behalf of (or instead of) a UAS crew member. IAI requirements involve visual and aural cues, checklists, and indications that consolidate or accentuate information directed to the crew member. The two main areas addressed by the APWE are sharing SA and ensuring that procedures are clearly identified in the decision ladder and followed. For the ten steps comprising the target engagement process shown in Fig. 2, nineteen decision ladder items were identified from the CTA (McColl et al. 2017).

With respect to the automation of crew functions, some functions are obvious candidates for automation, while the automation of other functions may be contrary to doctrine and/or legal considerations. For example, certain aspects of the CDE activity can be greatly facilitated through the use of AI-based agent technologies. Automatically signaling potential fratricide or civilian casualties and other relevant information to the crew early on in the target engagement process could facilitate crew decision-making, leading to remedial measures and alternate course of action. Overall, a hierarchal team of collaborating software agents (i.e., AIAs (Hou et al. 2011)) has been identified as automation aids included in the APWE prototype. These agents were responsible for tasks such as monitoring mission conditions and operator states; detecting anomalies, threats, or other events of interest; generating notification messages and other information sharing; and presenting options to the operator. Additional automation logic was implemented using a hierarchal state machine that captured the mission logic and provided a link between the AIAs and the operators (McColl et al. 2017).

3.3 IMPACTS-based APWE design

With the understanding of the requirements mentioned above and the objective of APWE is to support a variety of operational roles and contexts, the IMPACTS trust model was applied in conceiving the structure, the functionalities, and the interaction mechanisms of APWE along with the interaction-centered design (ICD) principles (Hou et al. 2014; McColl et al. 2017) for APWE prototype development. Since different users require different functionalities from their APWE views, a client server software architecture was used for the design so that each view is a separate client application or configuration. This approach allows all stakeholders to have a common understanding of shared mission goals and SA as well as their own required views. There were basically five different views developed for different users within a typical APWE prototype for a UAS target engagement mission.

1. The Shared SA View is for sharing the current status of the target engagement activity with all users and accepts no user input. As illustrated in Fig. 3, it displays the engagement status in the form of a graphic that indicates completed APWE steps and uncompleted steps with different color coding. The Shared SA View can be integrated into other applications, such as an overall mission SA display, as shown in Fig. 4.

Figure 4 depicts APWE re-configurable HMI components which display (1) the engagement procedure and the crew's current status in the state board (top left corner); (2) the requirements to progress to the next step of the procedure; (3) a map display focusing on the targeted location; (4) relevant sensor feed; (5) a list of ROEs in effect; and (6) a summary of pertinent crew text chat.

2. The UAS Crew View is intended for use by the UAS crew members not performing the pilot or mission commander roles (e.g., payload operator or intelligence analysts). In addition to the Shared SA View, this view provides information panels with a summary of the various steps (see Fig. 5). The information panels also allow a crew member to request actions from other users or set reminder alarms. When clicking on a specific information item, this view offers a drill-down capability to access increasing levels of detail. The information panels are related to another important feature of the APWE checklists that crew members must complete and submit as part of the authorization request process.
3. The UAS Crew Commander View is intended for the pilot, generally acting as a crew commander, i.e., the person responsible for the mission and weapon engagements. In addition to the functionality of the UAS Crew View, this view allows the crew commander to validate checklists and submit kinetic response and weapon release authorization requests to the TA.
4. The Tasking Authority/White Cell View is for use by the TA but also for a role-player acting as the TA during an experimental trial (referred as the White Cell for a synthetic experimental setting). In addition to the read-only access to the APWE status (Shared SA View) and information panels, this view allows the dedicated user to grant

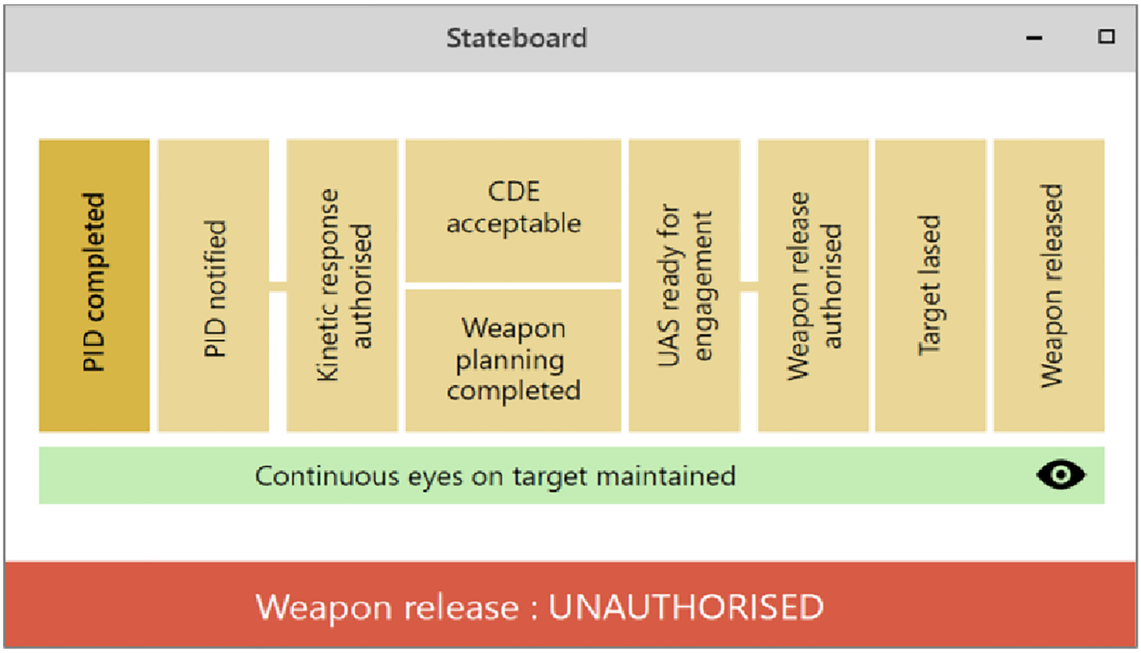
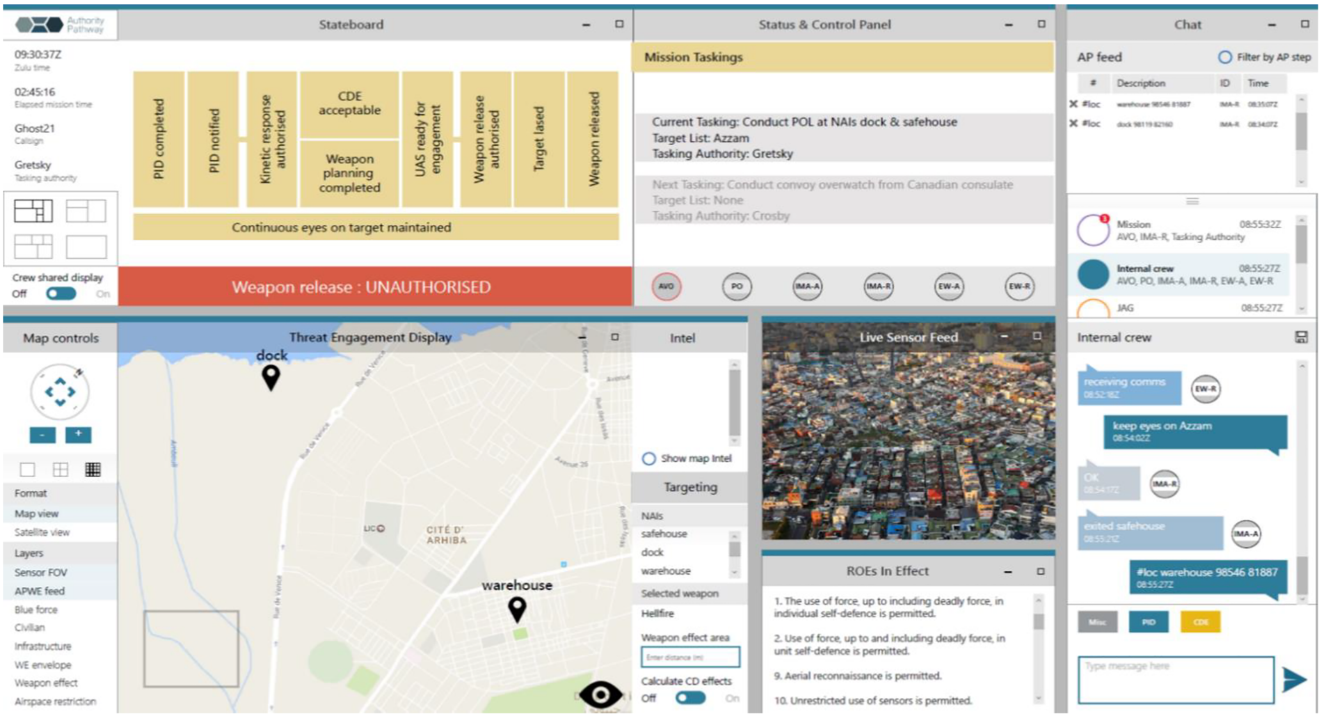

Stateboard
PID completed
PID notified
Kinetic response authorised
CDE acceptable
Weapon planning completed
UAS ready for engagement
Weapon release authorised
Target lased
Weapon released
Continuous eyes on target maintained
Weapon release : UNAUTHORISED

Fig. 3 APWE Shared SA View (an AI state board for increased process transparency, predictability, and communication (original by the authors))



Current Tasking: Conduct POL at NAIs dock & safehouse
Target List: Azzam
Tasking Authority: Gretsky

Next Tasking: Conduct convoy overwatch from Canadian consulate
Target List: None
Tasking Authority: Crosby

	#	Description	ID	Time
✕	#loc	warehouse 98546 81887	IMA-R	08:35:07Z
✕	#loc	dock 98119 82160	IMA-R	08:34:07Z

Mission 08:55:32Z
AVO, IMA-R, Tasking Authority

Internal crew 08:55:27Z
AVO, PO, IMA-A, IMA-R, EW-A, EW-R

JAG 08:55:27Z

NAIs: safehouse, dock, warehouse

Selected weapon: Hellfire

Weapon effect area: Enter distance (m)

Calculate CD effects: Off / On

1. The use of force, up to including deadly force, in individual self-defence is permitted.

2. Use of force, up to and including deadly force, in unit self-defence is permitted.

9. Aerial reconnaissance is permitted.

10. Unrestricted use of sensors is permitted.

EW-R: receiving comms 08:52:18Z

keep eyes on Azzam 08:54:02Z

IMA-R: OK 08:54:17Z

IMA-A: exited safehouse 08:55:21Z

#loc warehouse 98546 81887 08:55:27Z

Type message here

Fig. 4 An APWE reconfigurable state board within a UAS GCS HMI (top-left corner) (original by the authors)

or deny authorizations from the pilot. This view allows the user to review a checklist and to flag items that are incomplete, missing, or otherwise inadequate. The user can reject a request and provide instructions for rectifying it for subsequent resubmission.

5. The Experimenter View is to support experimentation purpose. The APWE prototype has a record and playback capability for after-action review (AAR) that includes observations from experimenters. The Experimenter View allows the user to generate events and comments for subsequent review as part of the AAR. For example, the user can identify specific instances during which Eyes on Target were lost and then regained. Other Experimenter View functions may include the capability to record observations concerning the operator's state.

The design of all these views allows APWE to automatically present and update the status of each of the steps required to release a weapon, based on the inputs from the UAS crew, external TA, and AIAs when engaging with a target. APWE provides not only the shared SA about the past and current status of the target engagement process but also its intention of the next step. APWE also provides a variety of means to users to share their intentions so that the UAS crew and their TA effectively and efficiently progress through the target engagement procedure. Thus this capability exhibits the *intention* attribute of the IMPACTS model.

With this configuration, APWE behaviors are measurable against other components in the HMI. For example, with an understanding of the UAS crew's intention for target engagement and decision-making, APWE includes mechanisms for communicating requests for TA's authorization along with access to the completed checklists. The TA can authorize or refuse a request based on his/her knowledge of the ROEs and LOAC for the current mission situation. In the case of a refused request, the checklist items that were incomplete, missing, or otherwise inadequate will be flagged and returned as

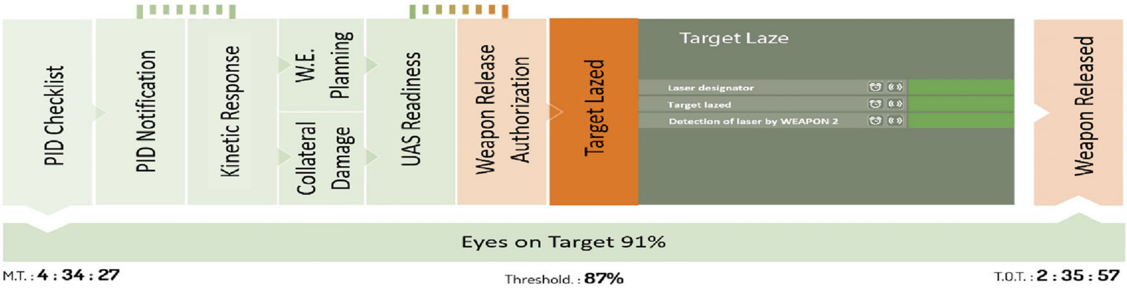


Fig. 5 APWE information panel (Original by the authors)

part of the response for an authorization request. In the case of an approved request, additional information or guidance may be provided as part of the response. Among other means, UAS Crew View (Fig. 5), UAS Crew Commander View, and TA View all allow the various users to have redundant information to compare or as a backup if anything missing. Different UAS crew members can also choose the components needed for their tasks. These functionalities enable APWE to demonstrate the *measurability* characteristic of the IMPACTS model.

APWE provides an intelligent, adaptive "state board" interface which is used to visualize the entire weapon engagement sequence, as shown in Figs. 3 and 4. The state board provides SA and feedback to all the UAS crew members and external users (e.g., the TA), consistently outlining the current step of the engagement, allowing them to easily and efficiently determine target eligibility for weapon engagement. This IAI technology (Hou et al. 2011, 2014) is not only a communication tool but also provides transparency and predictability for the next steps on the checklist. This capability aids the UAS crew with decision-making by supporting interactions with the TA and estimating collateral damage during the target engagement process. This support helps reliably prevent the crew from inadvertently engaging a target before the TA has granted proper permission to do so. AIAs within APME exhibit consistent, predictable, and reliable behavior and interactions with their human partner over time across different situations. Thus, the demonstrated consistency, predictability, and reliability are what the *performance* property of the IMPACTS model requests.

APWE adapts its interface (e.g., different views) to the UAS crew and other external users based on different ROE and LOAC information and communication requirements. Within APWE, target engagement procedures can vary as a function of the operational context, the type of mission, and the force structure. The steps involved in a TIC situation are not the same as those required for engaging a target during a POL mission. Also, it can be time-consuming for the UAS crew to identify all of the applicable and relevant ROE for a given situation. APWE address these concerns by being adaptive to the specific situation and generating appropriate checklists to guide the target engagement process, including facilitating the assessment of relevant ROE and other items related to the LOAC. Thus, APWE has the *adaptivity* (to three "context constraints" (Hou et al. 2014)) trait of the IMPACTS model to (1) adapt to the mission type and UAS role, both of which may vary over the course of the operation; (2) provide user-specific views (e.g., different UAS crew roles and external TA); and (3) allow for the evolution of AI capability consistent with future UAS requirements for additional semi-automated crew functions related to target engagement.

APWE provides bi-directional communication capability in a variety of ways (e.g., features provided by UAS Crew View, UAS Crew Commander View, and TA View, etc.). For example, authorization requests are made by the UAS crew to the TA at two steps of the target engagement process: (1) after the PID checklist has been completed and confirmed, at which point the UAS crew requests authorization for a kinetic response, and (2) once the weapon (engagement) planning and CDE steps have been completed and confirmed at which time the UAS crew requests an authorization for weapon release (see Figs. 5 and 6.). The crew status displays are equipped with a drill-down feature that allows the operators to access increasing levels of detail concerning information items of a given checklist or response to a request for authorization of weapon release. All these features enable bi-directional communications between the AIAs and their human partners. The automated alerts, notifications, and warnings also contribute to increasing the responsiveness and therefore facilitating the communications during a target engagement. All these features demonstrate the *communication* requirement of the IMPACTS model.

APWE exhibits *transparency* of the IMPACTS model in different formats. The most prominent feature of APWE is the intelligent adaptive target engagement state board, as shown in Figs. 3 and 4. This IAI provides an unambiguous and instantaneous view of the target engagement process status. This view serves two major purposes: (1) it contributes to the overall mission SA so that the UAS crew is aware of the current step in the process; and (2) it facilitates any necessary remedial actions by clearly indicating why a given step is stalled or why a request for authorization has been denied. Basically, this view provides information with a summary of the various steps and allows a crew member to request actions from other users or set reminder alarms. When clicking on a specific information item (e.g., on the APWE information panel, as shown in Fig. 5), this view offers a drill-down capability to access increasing levels of detail. The state board is also related to another important feature of APWE—checklists that UAS crew members must complete and submit as part of the authorization request process.

APWE demonstrates the *security* property of the IMPACTS model by providing explanation-for-confidence and thus trust. Due to the complex and dynamic nature of the APWE process in the operational tempo, AI agents were designed and developed to assist decision-making, and the operator does not need to know the detailed inner workings of the black box (i.e., the target engagement process). Upon receiving an order to launch a lethal weapon, it automatically gathers information from all sources about the target area including assets being considered for engagement of the target and their capabilities (for potential civilian casualties, etc.), checks the status of the weapon planning process, and finds related ROEs, LOAC, and SOPs preprogrammed in the system. Then the CDE process is conducted to check whether the potential strike follows the appropriate ROEs, LOAC, and SOPs (e.g., lethal vs. non-lethal). If not, it

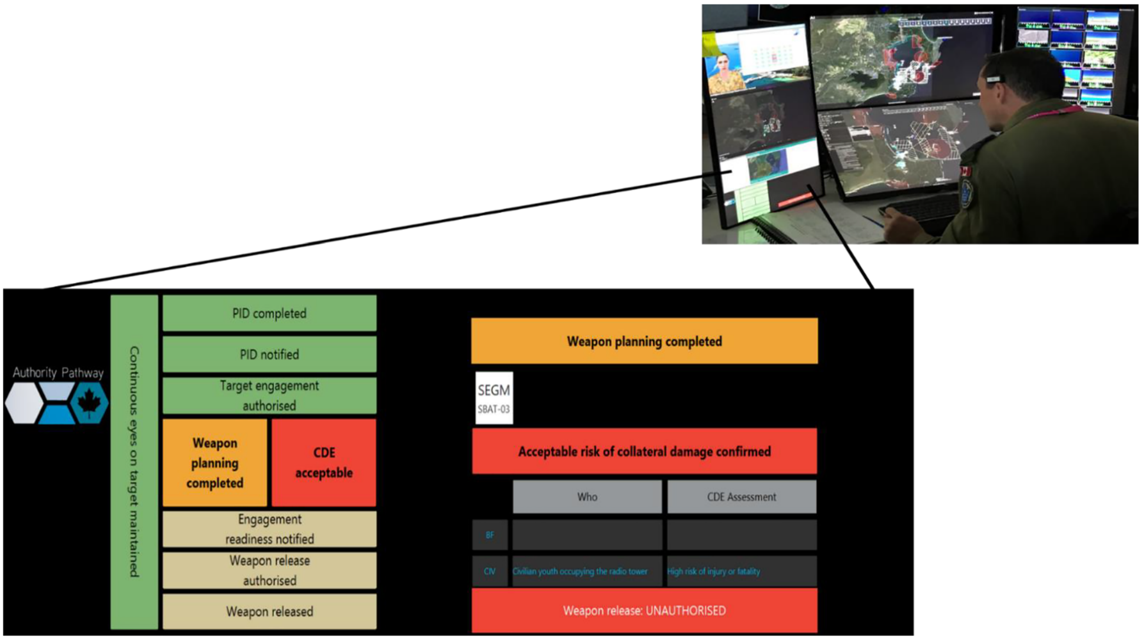

Authority Pathway
Continuous eyes on target maintained
PID completed
PID notified
Target engagement authorised
Weapon planning completed
CDE acceptable
Engagement readiness notified
Weapon release authorised
Weapon released
Weapon planning completed
SEGM
SBAT-03
Acceptable risk of collateral damage confirmed
Who
CDE Assessment
BF
CIV
Civilian youth occupying the radio tower
High risk of injury or fatality
Weapon release: UNAUTHORISED

Fig. 6 An UAS operator working with APWE in the AIM System during TTCP "Autonomous Warrior 2018" Joint Exercise (Original by the authors)

will refuse the launch of the lethal weapon with a suggested solution (e.g., non-lethal) to the TA so that the TA can alter or stop the launch. This explanation-for-confidence capability provides APWE secured protection against accidental or malicious use of the system or even deliberate cyber-attacks to facilitate trust. Figure 6 illustrates "the explanation of AI decisions" as a step of APWE security process after successful completion of weapon planning and CDE process based on built-in ROEs, LOAC, and SOPs.

4 APWE field test

To assess its military utility and effectiveness, investigate interoperability among multiple military systems, and increase the expanded (multinational) crew's SA concerning a specific weapon engagement authorization process, APWE was modified and integrated into a joint C2 system which has a variety of highly disruptive technologies from three of The Technical Cooperation Program (TTCP) nations: Australia, the UK, and the USA. One of the C2 technologies is also called IMPACT (but has no relationship with current model), which is a C2 system for controlling multiple unmanned systems (Draper et al. 2017). This allied IMPACT C2 system was then referred to as AIM. The AIM was successfully evaluated through a large joint service TTCP Autonomy Strategic Challenge milestone demonstration and evaluation exercise "Autonomous Warrior 2018" (AW2018) in Australia in November 2018 (Frost et al. 2019; Bartik et al. 2020). During the AW2018 exercise, APWE automatically conducted CDE by monitoring the data feeds from the AIM to keep track of vehicle capabilities and monitoring automatic asset planning to understand which vehicles are being considered for engagement of the target. During the asset planning, restrictions are checked to determine if the desired effect is being triggered (lethal vs. non-lethal). Figure 6 shows an AIM operator working with APWE which was integrated within the AIM during the AW2018 exercise.

AW2018 exercise included both live and synthetic trials, both run in parallel over the trial period, and both involving extensive data collections. In the live trials, a team of seven trained military subject matter experts (SMEs) observed an experienced AIM operator tasking both live and simulated assets over three different well-designed use cases. SMEs captured their observations during the scenario using a real-time data logging system. The evaluation team then used the after-action report to facilitate discussion of key events in the use case scenario with the SME group, extracting key observations for later analysis.

In the synthetic trials, evaluation team members observed as these same SMEs took turns to serve as the AIM operator, directly tasking simulated assets in a single scenario based on a use case involved an infrastructure protection against unlawful entry and attack by hostile actors. The evaluation team also collected extensive data on the overall mission and task effectiveness, workload, SA, trust, system usability, and human-autonomy teaming performance. The trustworthiness of the AIM system was assessed by the SMEs via a questionnaire twice over the course of the exercise, once after acting as the operator in the synthetic trial and once at its completion. SMEs rated the AIM system on a scale from 1 to 7 (1 = low, 7 = high) across seven known trust dimensions and provided some clarifying comments associated with their scores (Bartik et al. 2020). The trust dimensions are:

- Reliability, the ability of the system to operate on missions, perform tasks, and deliver effects as specified
- Dependability, the ability of the system to be relied upon to operate on missions, perform tasks, and deliver effects as specified
- Predictability, the ability of the system to respond to events and to operate, perform, and deliver effects consistently and reliably as planned and anticipated
- Availability, the ability of the system to operate on missions, perform tasks, and deliver effects when requested
- Resilience, the ability of the system to transform, renew, and recover in timely response to events
- Safety, the ability of the system to operate without harmful states
- Security, the ability of the system to remain protected against accidental or deliberate attacks.

The AW2018 was a large-scale exercise and extremely complex. It involved multiple prototype C2 technologies from four different TTCP nations, who integrated, demonstrated, and evaluated the technologies over the course of three weeks. As a result, individual technologies could not be assessed separately against each of the individual experimental measures including those trust dimensions. However, empirical results relating directly to the evaluation of APWE, such as SME responses to questionnaires completed immediately after the conduct of live and synthetic use cases, and information collected during debriefing sessions with the SMEs were still analyzed to inform the utility, effectiveness, and interoperability of the technology.

First, regarding trust dimensions, APWE directly contributed to system reliability and security because "it doesn't let you inadvertently engage before you have permission" and "prevented casualties" by conducting CDE in support of weapon engagements. SMEs reported that it "helps prompt what needs to be done," "is very easy to use," "is a good model; logical and assuring," and "love[d] the Authority Pathway function and reliability."

Second, it provided transparency about the status of the target engagement process and predictability of the current and next steps on its state board based on lawful ROEs. A SME reported that APWE was "helpful in prompting the operator" and commented himself as "massive fan (of APWE), logical and great visibility of engagement status."

Third, APWE was identified as a contributor to calibrated trust, information presentation, and adaptability factors to support human-autonomy teaming due to its communication capability to make the status of target engagement readily available when needed and stop engagement whenever CDE failed.

Overall, APWE was reported to be a significant factor behind the success of the AIM system. SMEs were enthusiastic about the APWE application, with four out of the seven reporting that APWE was one of the top three strengths of the AIM system. The implementation of APWE within AIM was considered "exemplary, with major enhancements" because it "could seriously benefit future operations." In particular, APWE "takes a lot of stress away from the operator," and most importantly it was "the most trustworthy of the whole thing (the AIM system) because the increased SA and reduced workload and potential human error…".

As an intelligent decision aid, APWE automatically streamlines engagement processes and displays the engagement status dynamically and intuitively, thereby reducing engagement times and errors while enhancing the operators' SA. The evaluation results and feedback from SMEs at the AW2018 exercise clearly demonstrated the utility, effectiveness, and interoperability of APWE concept and technology to support human-autonomy teaming. Indirectly, it provided empirical evidence of the logic (i.e., built-in capabilities and integrity) behind the design of the system, which is the concept and model of the IMPACTS for enabling the trust between human and autonomous systems.

5 Conclusion

How to build a trusted relationship between human and their intelligent machine partners (i.e., IAS, Hou et al. 2014) is becoming increasingly challenging to system designers when technologies like AI and autonomous systems advance at an ever-rapid pace. Trust can be defined as an attitude or belief that an operator is confident and willing to act based on the recommendations, actions, and decisions of an autonomous system or a group of agents. Given its importance to the design and implementation of human-autonomy teaming technologies, a review of the state-of-the-art of theoretical, mathematical, and performance-based models and measures of trust and how they might inform the deployment of suitable trust metrics during human-in-the-loop (HTIL) trials has been conducted in detail. The review identified the main theoretical and practical approaches to studying human-autonomy trust (e.g., human factors, psychology, cognitive science, computer science, human-computer interaction), as well as the linkages among them. Although there are many trust models from different fields and their efficacy has been investigated in a laboratory-based environment, little has been done to assess their effectiveness for the operational environment. They are helpful to understand trust factors, but few are practically useful or are too complex for system designers to build up an assured trust relationship between human and autonomous systems. However, the review of these trust models and metrics was used to identify seven critical and applicable characteristics of a conceptual and practical trust model: IMPACTS (intention, measurability, performance, adaptivity, communication, transparency, and security).

The IMPACTS model focuses on the most critical principles for practically building human trust in autonomous systems when teaming up these two partners together towards achieving the team's common goals. The model was used to guide the design and development of an intelligent adaptive decision aid for assisting a UAS crew in decision-making during complex, lengthy, and error-prone target engagement processes. The APWE decision aid exhibits the seven characteristic elements of the IMPACTS model: intention (shared mental model of team intents), measurability (observable behaviors and measurable actions to demonstrate the understanding of common goals), performance (consistent and reliable behaviors with predictable outcomes to support the shared intention), adaptivity (agile to changing context and environment as well as human needs), communication (bidirectional interactions to facilitate joint decision-making), transparency (status of various agent tasks), and security (protection against accidental or deliberate attacks). To assess the utility of the IMPACTS model in design, an APWE prototype was then designed, developed, integrated, demonstrated, and evaluated during an international autonomy and AI technology demonstration and evaluation exercise. Positive feedback on APWE's usability, utility, effectiveness, and interoperability to support human-autonomy teaming indicated the potential for the IMPACTS model to guide the system design to inspire users' confidence and instill trust. If designed well, based on the IMPACTS principles, a trusted and collaborative partnership between human and autonomous systems can be enabled and assured.

Since trust is an abstract construct and a dynamic psychological state with multiple affecting factors from human limitations, technology capabilities, and environment constraints, the critical trust properties are still a context-dependent issue. Although many trust studies have been conducted in multiple research fields over many decades, still very little is empirically known about trust within the context of human-autonomy teaming especially in an operational environment. This reality limits the scope of this research that focuses on meaningful and practical guidance for the designers of IASs. However, with the rapid advancement and applications of AI and autonomous systems in our society, it demands a comprehensive set of global standards for all stakeholders to follow. Thus, only the seven characteristics are chosen to formulate the essential properties of IMPACTS model here for the system design process to enable trust between the user and the autonomous system. These seven characteristics need to be empirically validated further for their efficacy and effectiveness to provide a theoretical foundation of the IMPACTS model.

Given that the validity of IMPACTS model was only indirectly assessed through APWE experiments during the AW2018 exercise, the model needs to be further validated directly and/or modified/refined accordingly through human-autonomy teaming HTIL trials. More detailed instructions for system designers also need to be developed to guide their design activities so that the IMPACTS model can be applied broadly for effective human-technology interactions in addressing trust, accountability, legal, ethical, policy, regulation issues, etc.

Acknowledgments The authors would like to thank Emily Herbert and Jayshal Sood for their editorial support to this paper. The Royal Canadian Air Force's support to the APWE and AW2018 evaluation trials is also appreciated.

References

Abbass A, Petraki E, Merrick K, Harvey J, Barlow M (2016) Trusted autonomy and cognitive cyber symbiosis: open challenges. Cogn Comput 8:385–408

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, ..., Chatila R (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58:82–115

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I (2018) The moral machine experiment. Nature 563:59–64

Baber C (2017) BOOK REVIEW Intelligent adaptive systems: an interaction-centered design perspective. Ergonomics 60(10):1458–1459

Bartik J, Rowe A, Draper M, Frost E, Buchanan A, Evans D, Gustafson E, Lucero C, Omelko V, McDermott P, Wark S, Skinner M, Vince J, Shanahan C, Nowina-Krowicki M, Moy G, Marsh L, Williams D, Pongracic H, Thorpe A, Keirl H, Hou M, Banbury S (2020) Autonomy strategic challenge (ASC) allied IMPACT final report. TTCP TR-ASC-01-2020

Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B et al. (2018) The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Future of Humanity Institute, University of Oxford

Chen JYC, Barnes MJ (2014) Human–agent teaming for multi-robot control: a review of human factors issues. IEEE Transactions on Human–Machine Systems 44:13–29

Chen JYC, Barnes M, Selkowitz AR, Stowers K (2016) Effects of agent transparency on human-autonomy teaming effectiveness in Proc. IEEE International Conference on Systems, Man, and Cybernetics SMC 2016, October 9–12, Budapest, Hungary

Chen JYC, Lakhmani SG, Stowers K, Selkowitz AR, Wright JL, Barnes M (2018) Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theor Issues Ergon Sci 19(3):259–282

Cho JH, Chan K, Adali S (2015) A survey on trust modeling. ACM Computing Surveys 48(2):28

Computing Community Consortium (2020) Assured autonomy: path toward living with autonomous systems we can trust. Computing Community Consortium, Washington, DC. Retrieved from https://cra.org/ccc/wp-content/uploads/sites/2/2020/10/Assured-Autonomy-Workshop-Report-Final.pdf. Accessed 29 Oct 2020

Covey SMR (2008) The speed of trust: the one thing that changes everything. Free Press, New York, NY de Visser EJ, Pak R, Shaw TH (2018) From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. Ergonomics 61(10):1409–1427

Defense Science Board (2016) Summer study on autonomy. Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, Washington, D.C, pp 20301–23140

Desai M, Kaniarasu P, Medvedev M, Steinfeld A, Yanco H (2013) Impact of robot failures and feedback on real-time trust. 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Tokyo, pp. 251-258, doi: https://doi.org/10.1109/HRI.2013.6483596

Draper M, Calhoun G, Hansen M, Douglass S, Spriggs S, Patzek M, Rowe A, Evans D, Ruff H, Behymer K, Howard M, Bearden G, Frost E (2017) Intelligent multi-unmanned vehicle planner with adaptive collaborative control technologies (IMPACT). 19th International Symposium of Aviation Psychology 226–231

Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. Hum Factors 37(1):32–64

Endsley MR (2019) Human factors & aviation safety: testimony to the United States House of Representatives hearing on Boeing 737-Max 8 crashes. Human Factors and Ergonomics Society, December 11. Retrieved from https://www.hfes.org/Portals/0/Documents/Human_Factors_and_the_Boeing_737-Max8-FINAL.pdf?ver=2020-08-28-163636-570. Accessed 16 June 2020

Erikirk E (1993) Childhood and society: the landmark work on the social significance of childhood. W. W. Norton & Company, New York

Frost E, Calhoun G, Ruff H, Bartik J, Behymer K, Springs S, Buchanan A (2019) Collaboration interface supporting human-autonomy teaming for unmanned vehicle management. In Proceeding of the 20th International Symposium on Aviation Psychology, pp. 151-156

Gunning D, Aha DW (2019) DARPA's explainable artificial intelligence program. AI Mag 40(2):44–58

Hancock PA, Billings DR, Oleson KE, Chen JYC, de Visser E, Parasuraman R (2011) A meta-analysis of factors impacting trust in human-robot interaction. Hum Factors 53:517–527

Harbers M, Jonker C, van Reimsdijk B (2012) Enhancing team performance through effective communications. Paper presented at The Annual Human-Agent-Robot Teamwork (HART) Workshop. Boston, MA

Helldin T (2014) Transparency for future semi-automated systems. PhD dissertation, Örebro Univ, Örebro, Sweden

Hoff KA, Bashir M (2015) Trust in automation integrating empirical evidence on factors that influence trust. Hum Factors 57:407–434

Hou M, Zhu H, Zhou MC, Arrabito R (2011) Optimizing operator-agent interaction in intelligent adaptive interface design. IEEE Transaction Systems, Man, and Cybernetics Part C: Applications and Reviews 41(2):161–178

Hou M, Banbury S, Burns C (2014) Intelligent adaptive systems: an interaction-centered design perspective, 1st edn. CRC Press, Boca Raton

Hughes S (2013) Campaigners call for international ban on 'killer robots'. BBC News. Retrieved from http://www.bbc.co.uk. Accessed 8 Jul 2020

Jansen BJ (1999) A software agent for performance enhancement of an information retrieval engine (doctoral dissertation). A & M University, Texas: UMI Dissertation Services

Jarvenpaa S, Knoll K, Leidner D (1998) Is anybody out here? Antecedents of trust in global virtual teams. Journal of Management Information Systems 14(4):29–64

Jian JY, Bisantz AM, Drury CG (2000) Foundations for an empirically determined scale of trust in automated systems. Int J Cogn Ergon 4(1):53–71

Lai K, Oliveira H, Hou M, Yanushkevich SN, Shmerko V (2020a) (In press) Assessing risks of biases in cognitive decision support systems. European Signal Processing Conference

Lai K, Yanushkevicha SN, Shmerkoa V, Hou M (2020b) Risk, trust, and bias: causal regulators of biometric-enabled decision support. Special Selection on Intelligent Biometric Systems for Secure Societies, IEEE Access 8:148779–148792

Lamb C (2016) The talented Mr. robot: the impact of automation on Canada's workforce. Brookfield Institute for Innovation and Entrepreneurship, Toronto, Canada, June. Retrieved from https://brookfieldinstitute.ca/wp-content/uploads/TalentedMrRobot_BIIE-1.pdf. Accessed 18 Jul 2020

Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors 46(1):50–80

Mae Pedron S, Jose de Arimateia DC (2020) The future of wars: artificial intelligence (AI) and lethal autonomous weapon systems (LAWS). International Journal of Security Studies, 2(1). Article 2

Marks S, Dahir AL (2020) Ethiopian report on 737 max crash blames Boeing. The New York Times. Retrieved from https://www.nytimes.com/2020/03/09/world/africa/ethiopia-crash-boeing.html. Accessed 29 Oct 2020

Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. Acad Manag Rev 20:709–734

McAllister DJ (1995) Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. Acad Manag J 38:24–59

McColl D, Banbury S, Hou M (2016) Testbed for integrated ground control station experimentation and rehearsal: crew performance and authority pathway concept development. In: Lackey S, Shumaker S (eds) Virtual, Augmented and mixed reality. LNCS, vol 9740. Springer, Heidelberg, pp 433–445

McColl D, Heffner K, Banbury S, Charron M, Arrabito R, Hou, M (2017) Authority pathway: intelligent adaptive automation for a UAS ground control station. In Proceedings of HCI International Conference, Vancouver, July

McNeese N, Demir M, Chiou E, Cooke, N (2019) Understanding the role of trust in human-autonomy teaming. In Proceedings of the 52nd Hawaii International Conference on System Science, pp 254–263

Merritt SM, Ilgen DR (2008) Not all trust is created equal: dispositional and history-based trust in human–automation interactions. Hum Factors 50:194–210

Miller C (2000) Intelligent user interfaces for correspondence domains: moving IUI's "off the desktop". In Proceedings of the 5th International Conference on Intelligent User Interfaces, pp. 181–186. New York, NY: ACM Press

Miller CA, Wu P, Funk H (2007) A computational approach to etiquette and politeness: validation experiments. In D. Nau, & J. Wilkenfeld (Eds.), Proceedings of the First International Conference on Computational Cultural Dynamics, pp.57–65. August 27-28, Menlo Park, CA: AAAI press

Muir BM (1994) Trust in automation: part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 37(11):1905–1922

Murphy RR, Woods DD (2009) Beyond Asimov: the three laws of responsible robotics. IEEE Intell Syst 24(4):14–20

Nassar M, Salah K, Rehman MH, Svetinovic D (2019) Blockchain for explainable and trustworthy artificial intelligence. Data Mining Knowledge Discovery 10(1). https://doi.org/10.1002/widm.1340

NATO STO SAS 085 (2013) C2 agility – task group SAS-085 final report (STO technical report STO-TR-SAS-085). NATO Science and Technology Organization, Brussels

NIS Cooperation Group (2019) EU coordinated risk assessment of the cybersecurity of 5G networks. NIS cooperation group. Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62132. Accessed 29 Oct 2020

Olson WA, Sarter NB (2000) Automation management strategies: pilot preferences and operational experiences. Int J Aviat Psychol 10(4):327–341

Onnasch L, Wickens CD, Li H, Manzey D (2014) Human performance consequences of stages and levels of qutomation: an integrated meta-analysis. Hum Factors 56(3):476–488

Parasuraman R, Sheridan TB, Wickens CD (2000) A model of types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics – Part A 30:286–297

Peiters W (2011) Explanation and trust: what to tell the user in security and AI? Ethics in Information Technology 13:53–64

Pilkington M (2016) Blockchain technology: principles and applications. Research handbook on digital transformations. Edward Elgar Publishing, pp. 225. https://doi.org/10.4337/9781784717766.00019

Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon JF, Breazeal C, Crandall JW, Christakis NA, Couzin ID, Jackson MO, Jennings NR, Kamar E, Kloumann IM, Larochelle H, Lazer D, McElreath R, Mislove A, Parkes DC, Pentland AS, Roberts ME, Shariff A, Tenenbaum JB, Wellman M (2019) Machine behavior. Nature 568:477–486

Salas E, Sims DE, Burke CS (2005) Is there a 'big five' in teamwork? Small Group Res 36(5):555–599

Schaefer KE (2013) The perception and measurement of human–robot trust (doctoral dissertation). University of Central Florida, Orlando

Schaefer KE (2016) Measuring trust in human robot interactions: development of the 'trust perception scale-HRI'. In: Mittu R, Sofge D, Wagner A, Lawless W (eds) Robust intelligence and trust in autonomous systems. Springer, Boston

Schaefer KE, Chen JYC, Szalma JL, Hancock PA (2016) A meta-analysis of factors influencing the development of trust in automation. Human Factors 58(3):377–400

Schaefer KE, Straubb ER, Chen JYC, Putney J, Evans AW III (2017) Communicating intent to develop shared situation awareness and engender trust in human-agent teams. Cogn Syst Res 46:26–39

Sebok A, Wickens CD (2017) Implementing lumberjacks and black swans into model-based tools to support human-automation interaction. Hum Factors 59:189–202

Shaw J (2006) Intention in ethics. Canadian J of Philosophy 36(2):187–224

Sheridan TB (2002) Humans and automation: system design and research issues. Wiley-Interscience, Santa Monica

Sheridan TB (2019a) Extending three existing models to analysis of trust in automation: signal detection, statistical parameter estimation, and model-based control. Hum Factors 61(7):1162–1170

Sheridan TB (2019b) Individual differences in attributes of trust in automation: measurement and application to system design. Front Psychol 10:1117

Sheridan TB, Parasuraman R (2006) Human-automation interaction. In: Nickerson RS (ed) Reviews of human factors and ergonomics, vol. 1. Santa Monica, HFES

Sheridan TB, Verplank WL (1978) Human and computer control of undersea teleoperators (report no. N00014-77-C-0256). MIT Cambridge Man Machine Systems Laboratory, Cambridge, MA

Siau K, Wang W (2018) Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technology Journal 31(2):47–53

Sutton A, Samavi R (2018) Tamper-proof privacy auditing for artificial intelligence systems. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18). AAAI Press, pp. 5374–5378. Retrieved from https://www.ijcai.org/Proceedings/2018/0756.pdf. Accessed 16 June 2020

Sycara K, Lewis M (2004) Integrating intelligent agents into human teams. In: Salas E, Fiore S (eds) Team cognition: process and performance at the inter and intra-individual level. American Psychological Association, Washington

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. Retrieved from https://arxiv.org/abs/1312.6199. Accessed 16 June 2020

Taylor RM, Reising J (eds) (1995) The human-electronic crew: can we trust the team? (report no. WL-TR-96-3039). Paper presented at Third International Workshop on Human-Electronic Crew Teamwork, Cambridge, United Kingdom. Dayton, OH: Wright Air Force Research Laboratory

Verberne FM, Ham J, Midden CJH (2012) Trust in smart systems: sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. Hum Factors 54:799–810

Vicente KJ (1990) Coherence- and correspondence-driven work domains: implications for systems design. Behaviour and Information Technology 9(6):493–502

Vicente KJ (1999) Cognitive work analysis. Lawrence Erlbaum Associates, Mahwah

Vigano L, Magazenni D (2018) Explainable security. Retrieved from https://arxiv.org/abs/1807.04178

Vorobeychik Y, Kantarcioglu M (2018) Adversarial machine learning, 1st edn. Morgan & Claypool

Wang Y, Singh MP (2010) Evidence-based trust: a mathematical model geared for multiagent systems. ACM Trans. Autonomous and Adaptive Systems 5(4):14

Wang Y, Hou M, Plataniotis K, Kwong S, Leung H, Tunstel E, Rudas I, Trajkovic L (2020a) Towards a theoretical framework of autonomous systems underpinned by intelligence and systems sciences. IEEE/CAA Journal of Automatica Sinica. https://doi.org/10.1109/JAS.2020.1003432

Wang Y, Yanushkevich S, Hou M, Plataniotis K, Coates M, Gavrilova M, Hu Y, Karray F, Leung H, Mohammadi A, Kwong S, Tunstel E, Trajkovic L, Rudas IJ, Kacprzyk J (2020b) A tripartite framework of trustworthiness of autonomous systems. In Proceedings of the 2020 IEEE Systems, Man, and Cybernetic International Conference, Toronto, Canada, Oct., W15.1.1–6.

Wickens CD, Onnasch L, Sebok A, Manzey D (2020) Absence of DOA effect but no proper test of the lumberjack effect: a reply to Jamieson and Skraaning (2019). Hum Factors 62(4):530–534

Wilson JM, Straus SG, McEvily B (2016) All in due time: the development of trust in computer-mediated and face-to-face teams. Organ Behav Hum Decis Process 99:16–33

Yagoda RE (2011) WHAT! You want me to trust a robot? The development of a human robot interaction (HRI) trust scale. M.S. Thesis, Dept. of Psychology, N.Carolina State Univ., Raleigh, NC

Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning: concepts and applications. ACM Transactions on Intelligent Systems and Technology 10(2):1–19. https://doi.org/10.1145/3298981

Zhang B and Dafoe A (2019) Artificial intelligence: American attitudes and trends. Future of Humanity Institute, University of Oxford. Retrieved from https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/us_public_opinion_report_jan_2019.pdf. Accessed 27 Oct 2020

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.