ARL-TR-8636 ● FEB 2019

DEVCOM ARMY RESEARCH LABORATORY

Issues in Human–Agent Communication

by Michael J Barnes, Shan Lakhmani, Eric Holder, and JYC Chen

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8636 ● FEB 2019

Issues in Human–Agent Communication

by Michael J Barnes, Shan Lakhmani, Eric Holder, and JYC Chen
Human Research and Engineering Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
February 2019	Technical Report	September 2017–October 2018

4. TITLE AND SUBTITLE
Issues in Human–Agent Communication

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Michael J Barnes, Shan Lakhmani, Eric Holder, and JYC Chen

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Army Research Laboratory
(ATTN: RDRL-HRB-DE)
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-8636

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The report covers issues pertinent to the design and evaluation of the communication between humans and intelligent software agents necessary to enable a collaborative relationship. For human–agent interaction to be robust in a dynamic real-world situation, both software agents and humans must be able to communicate their overall intent in terms of mission objectives. Because of differences in their reasoning processes, capabilities, and knowledge bases, humans and agents are not an analog for human teams. We discuss the technical issues involved in effective communication including models of mutual transparency, natural language processing (NLP), artificial intelligence (AI), and explainable AI. Lacking a theory of mind, which enables humans to gain insight into their teammates' mental processes, agents have a difficult time anticipating human information needs and future actions. Research in collaborative planning involving multiple agents and research into synthetic shared mental models are used as exemplars of attempts to integrate humans and agents into a synergistic unit. However, we conclude that progress in NLP, explainable AI, and human science will be necessary before humans and agents communicate like human teams during complex, uncertain missions.

15. SUBJECT TERMS
machine learning, intelligent agent, human–machine communications, situation awareness-based agent transparency model, SAT model, human–agent teams, HATs

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Michael J Barnes
Unclassified	Unclassified	Unclassified	UU	29	19b. TELEPHONE NUMBER (Include area code) (520) 538-4702

Contents

List of Figures

1. Introduction

Both the promise and problems of autonomous systems will change the dynamics of future systems, not only in terms of the impact of the autonomous systems on society but also on their interactions with humans (Economist 2016; Schaefer et al. 2017). A partnership between humans and autonomous systems involves a blending of the artificial and the human into a cohesive system with all the advantages and limitations such a combination implies (Bradshaw et al. 2009; Chen and Barnes 2014). Autonomous systems can range from those that are standalone and only occasionally monitored by humans to human-directed systems that are closely supervised (Barnes et al. 2017). Software systems that are be able to act autonomously and update actions based on new information to achieve their objectives are identified as *intelligent agents* (IAs; Russell and Norvig 2009). In human–IA partnerships, a mixed initiative capability wherein humans and IAs share the decision space, but the human has ultimate authority, allows for flexibility while maintaining human responsibility in dangerous time-constrained situations (Chen and Barnes 2015; Barnes et al. 2017). In most cases, it would be impossible a priori to assign each to a specific role in a dynamic environment because their roles can change as the situation changes. For example, adaptive agents may take the decision initiative during high-workload mission segments without waiting for operator permission but return the decision initiative to the operator during normal operations (Chen and Barnes 2014). Some of the prescriptive rules pertaining to task allocation could be preset depending on the priorities of the mission. Other rules might change depending on the urgencies of the situation (e.g., autonomously shooting down an incoming missiles after a temporal deadline has expired [Barnes et al. 2017; Parasuraman et al. 2007]). However, in dynamic environments, communication, understanding of intent, and a common situation awareness (SA) are necessary for effective collaboration (Barnes et al. 2017; Evans et al. 2017; Holder 2018; Chen et al. 2018).

As IA complexity increases, so too does the necessity for effective communication. Cooke (2015) argues that an efficient teaming relationship depends more on effective interactions than it does on having an extensive shared knowledge base. Besides having a common linguistic framework, each team member must know when to push information to their partner and when to ask for information. Thus, it is important for both the human and the IA not only to have SA of the tasking environment but also to have SA of each other's roles in order to respond to their partner's requirements without overt communications (Scerri et al. 2003; Chen et al. 2018). We discuss three main themes. The first topic is a description of a human–agent architecture and why it is different than human–human teams, stressing the

importance of mutual transparency. Next, we discuss the technical issues involved with a human communicating with artificially intelligent (AI) systems including multimodal interfaces, linguistic constraints, types of AI, and the importance of explainable AI (XAI) to ensure mutual understanding. Finally, we discuss the importance of shared intent to foster a natural rhythm of push and pull of information between operators and IAs.

2. Human–Agent Collaboration Architecture

It would be a mistake to consider human teams as anything but a metaphor for human–agent interactions; humans and agents differ both in their capacities and their roles, especially in military environments. As indicated by Fig. 1, architectures of human processing and machine processing entail different representations of the world (Chakraborty et al. 2017). The agent's world model depends on its formal knowledge representations such as production systems (if–then rules), probabilistic modeling, optimization algorithms, and so on (Chen and Barnes 2014; Pynadath et al. 2018). In contrast, human decision making depends on heuristics, emotion, and imagery as well as formal logic. Agent decisions tend to be exact (or stated in terms of probabilities), and their accuracy depends on the appropriateness of the formalism they are based on and the limited knowledge they have of the real world. Humans make decisions that are broader and more flexible while sometimes using heuristics that are error prone even when they have the correct information to solve a problem (Kahneman 2011).

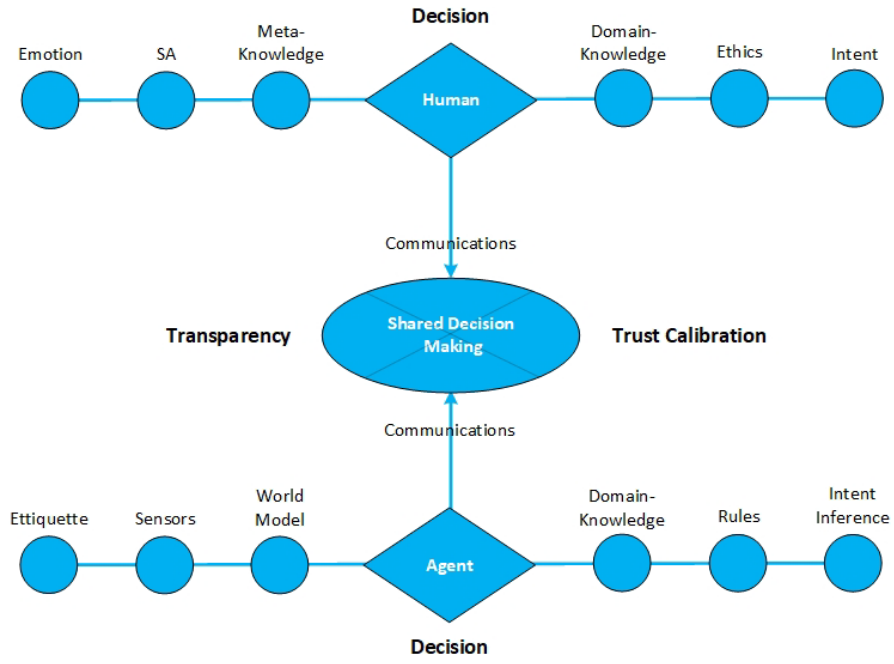

Fig. 1 Shared decision space between humans and agents (Barnes et al. 2017)

The challenge to designing an effective human–agent team is to combine the narrow exactness of a model-driven approach with the flexibility (and broader meta- knowledge) of the human. In this regard, a number of shared mental model (SMM) algorithmic approaches are being developed to enable humans and agents to predict the information needs, behaviors, and individual roles necessary for their mutual understanding of the task environment (Scerri et al. 2003; Pynadath and Marsella 2005; Wang et al. 2016). An SMM is the intersecting knowledge of; the other's role that humans and agents require to collaborate effectively, but SMMs do not preclude each partner having their own specialized functionality (Yena et al. 2006; Chen and Barnes 2014). To communicate, each agent must be able to interpret the intent, environmental cues, and symbolic referents of its partner (Lyons 2013; Chen et al. 2018). Thus, a communications architecture consists both of distinct processing units and interfaces that enable mutual interpretability. For example, deep learning approaches, which are notoriously opaque, may require an extra layer of processing to make their results transparent (Chakraborty et al. 2017; Pynadath et al. 2018).

The interface depicted in Fig. 1 requires a linguistic framework, mutual transparency, and calibrated trust (Lee and See 2004) in their respective roles for each of the partners. In summary, human–agent teams need to communicate in a similar fashion as humans, although their underlying processing and capabilities are quite different.

3. Mutual Transparency

Chen and colleagues (Chen et al. 2014, 2018; Chen and Barnes 2015) define transparency in terms of understanding the internal underpinnings of the agent's courses of action (COAs). The SA-based Agent Transparency (SAT) model defines the agent's suggested COAs as comprising three transparency levels (L): the agent's perception of its plan (L1), its logic (L2), and its predicted outcomes and their perceived likelihood (L3). SAT is similar to Endsley's (2015) original SA model but derived from the IA's perspective. The SAT model enables the operator to gain insight (SA) into the agent's world model and allows human operators to compare that information with their own SA of the ongoing real-world situation. The SAT model was tested in three diverse military paradigms, showing improved calibration and performance (reduced misuse and disuse of autonomously generated COAs) for an agent that conducted parameter defense (Mercado et al. 2016), infantry support (Selkowitz et al. 2016), and convoy in-route planning (Wright et al. 2016). Subjective trust was either not affected or actually improved as SAT levels increased, showing that operators trusted agents that reported information indicating the agent's misalignments with the real world (i.e., SAT

helped reduce misuse). Thus, an imperfect agent that was transparent was considered useful because it provided enough information for the operator to know the agent's limitations and uncertainties (Mercado et al. 2016; Stowers et al. 2016). One limitation of the SAT research is that it measured performance for static decisions. Although the trials themselves were quite different, they did not depend on previous trials (each was a frozen slice of time). Collaborations in a dynamic world will require two-way transparency; the agent as well as the human must be able to understand the mission's objectives and their respective roles as the real-world environment changes continuously (Chen et al. 2018).

To reflect a dynamic environment, Fig. 2 depicts a continuously changing world depicted by feed-forward changes to SAT parameters based on previous feedback and a feedback mechanism changing the SAT model's inputs based on the changing environment (Chen et al. 2018). In most situations, the transition between feed forward and feedback is orderly and only minor changes are necessary. However, rapid changes may be necessary when unanticipated events occur. The dynamic quality of military environments will require continuous communications between humans and agents to ensure mutual understanding of the changing situation. An important implication of Fig. 2 is that is the human has a privileged loop, that is, the human can change either the weights of the parameters that the agent is using for its current method of computing a COA or simply change the COA (cf., Marathe et al. 2018). For example, in a real-time planning paradigm, the agent might choose a ground robot to surveil a bridge but the robot falls behind schedule and the urgency for surveillance has increased. The human can change the agent's objective by increasing the importance of timeliness; and assess the agent's revised COA or suggest a COA (e.g., redirect an unmanned aerial system [UAS]) and let the agent assess its implications (Calhoun et al. 2018). In both cases, the decision making and required communications are bidirectional with the human being the senior partner. The discourse between the operator and the agent needs to be iterative, requiring mutual SA of the mission objectives, reasoning, and expected outcomes reflecting the dynamic nature of unfolding real-world events (Chen et al. 2018). To explore the need to communicate in real time, the technical underpinnings of the human–agent communication structural components are discussed in the next section to explicate the issues involved in bidirectional collaboration between humans and AI software.

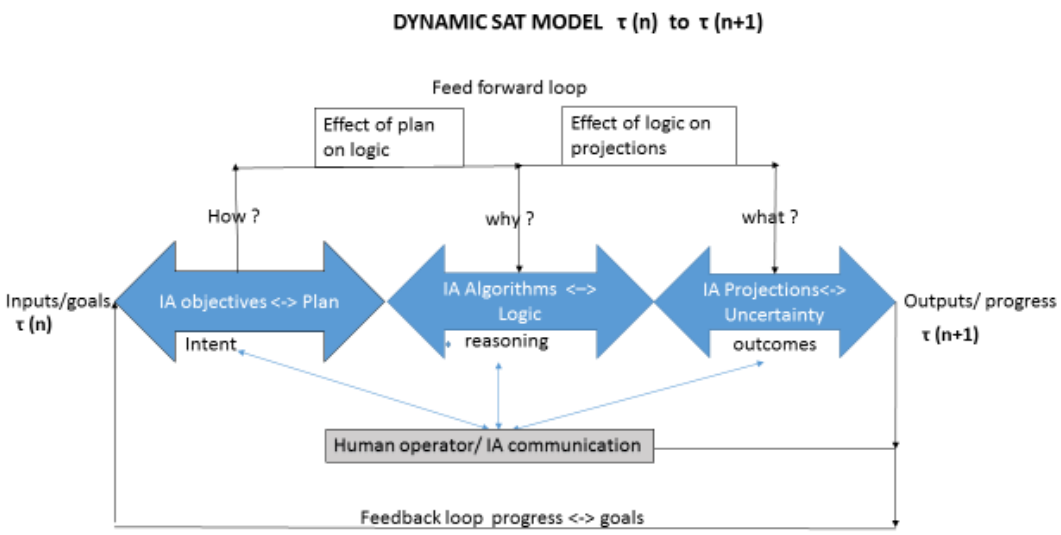

Fig. 2 Requirements for dynamic transparency for feed forward (time n – current changes) and feedback (time n+1 – future changes) (adapted from Chen et al. [2018])

4. Bidirectional Communication: Overview

Dialogue between agents consists of more than a linguistic interface. Salas et al. (2015) defined communication as "a reciprocal process of team members sending and receiving information that forms and re-forms a team's attitudes, behaviors, and cognitions". Thus communication is bidirectional, with the agent and the human connected by a symbol set, context, and grammar. Figure 3 outlines human–agent bidirectional communications that are interpreted by a natural language processing (NLP) component with the further assumption that graphical and nonverbal representations can be interpreted within a common linguistic framework. Further processing is done by the cognitive components, which may consist of a variety of AI techniques including cognitive architectures and machine learning (ML) components to augment the human partner's capabilities (Kelley and McGhee 2013; Kelley 2014; Barnes et al. 2017).

Just as speech and thought are integrally connected but also different, AI results must be interpretable in terms of the human's understanding of the communication output. Because AI reasoning can be opaque, it is sometimes necessary to have a specialized interpretation layer such as XAI (Chakraborty et al. 2017). In a dynamic environment, even fairly simple inquires ("what is the best route for an unmanned vehicle") can involve nontrivial processing and multiple iterations depending on the tradeoffs and uncertainties of both the human and the agent (Stowers et al. 2016; Chen et al. 2018). The interface in Fig. 3 is shown as the blue layer enabling multimodal inputs/outputs for bidirectional dialogues (Barber et al. 2013).

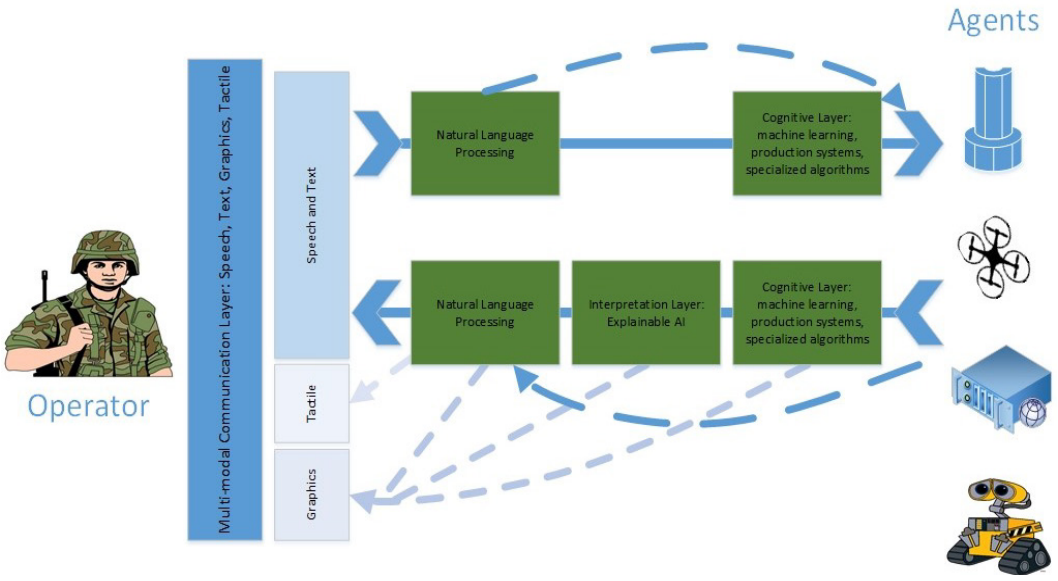

Fig. 3 Components of bidirectional communications

As Salas et al.'s (2015) definition implies not all communication is explicit; collaboration depends on a mutual understanding of intent. We explain each component in more detail, discussing the purpose and limitations of each as a conveyer of bidirectional information for mutual problem solving.

5. Natural Language Processing (NLP)

NLP uses software to understand and generate language. Although usually characterized by spoken language or text translations, any symbolic representation that is shared by the sender and the receiver needs to be integrated into NLP components. In order to understand (generate) meaningful symbols, the software must disambiguate morphology (structure of the symbol), semantics (meaning), syntax (rules), and pragmatics (context). Translations that depend on processing individual symbols are not practical because of the inherent ambiguity of language (Jurafsky and Martin 2009). To understand inputs, an agent must consider the preceding dialogue and environmental constraints to interpret even a modest stream of symbols. This means that nonverbal instances, such as tactile or graphic inputs, must fit into the general language framework in order to become part of the ongoing dialogue between humans and agents (Barber et al. 2013; Tal-Oron Gilad 2014). Pragmatics require the listener to understand the intent of a symbolic input. For example, commanding a robot (agent) to "put a glass of water down" does not mean to drop the glass or even to put the water glass on the floor. A human agent would search for a table or some safe place to put the glass and only ask if the solution was not obvious. Thus to communicate, an agent must be able to interpret the

purpose of the dialogue as well infer possible solutions to queries that go beyond *shallow* translations (Jurafsky and Martin 2009). Popular speech-based systems such as *Siri* and *Alexa* are useful, but they have a limited repertoire of knowledge and do not have the sophisticated world model and domain knowledge necessary for collaboration in a complex environment.

On one extreme, open-ended NLP may be too cumbersome for simple commands to a robot wherein a limited specialized vocabulary would be more efficient (Pettit et al. 2013; Barber et al. 2015). Look-up tables and simple rules, however, are not sufficient for more complex human–agent interactions (Barnes et al. 2017). An intermediate solution is *Controlled English* (CE), which is a specialized language processing approach that incorporates computational linguistics, specialized lexicons, and AI techniques for specific domains (Giammanco et al. 2015). Researchers are using CE with AI components to create agents that support humans in domains such as military intelligence, civil affairs, and UAS operations (Giammanco et al. 2015; Xue et al. 2015; McNeese et al. 2017). For example, using an inference engine that was trained by interacting with a military intelligence officer, CE software was able to interact with a human partner to collaborate on intelligence analyses in a laboratory setting (Mott et al. 2015). This approach avoids the all-encompassing processing issues of open-ended NLP by focusing on specialized environments, but it remains to be seen if it is flexible enough to adjust to the changeable verities of real-world situations. In particular, by delimiting processing to a specialized domain, human–agent interaction may be brittle as new elements are introduced into complex environments.

Research in NLP is currently investigating more complex human–agent interactions such as two-way communications, mixed initiative situations wherein the agent initiates dialogues, agent interpretation of human affective states, and the use of multimodal communications implying that more human-like communication will be possible in future systems (Mavridis 2015). An important issue related to NLP is the problem of having agents understand spatial relationships because two-way interactions between humans are often communicated more efficiently using graphical displays (Chen and Barnes 2014; Oron-Gilad 2014). For example, Tellex et al. (2011) demonstrated how generalized grounding graphs could be used to translate verbal descriptors into spatial representations (see also Skubic et al. [2004]). Also, a common linguistic framework can enable an agent to interpret tactile and gesture commands that are particularly useful in military environments when voice communications are not always feasible (Elliott et al. 2010; Barber et al. 2013, 2015; Mavridis 2015).

6. Cognitive Component

The processing capabilities of the agent can be complex, consisting of various types of AI systems working in concert to address different aspect of the environment. The actual mechanics of the AI is beyond the scope of this report; however, to effectively interact with humans, the AI architecture must have modules that can sense changes and have world models that interpret the meaning of inputs in terms of task objectives (Russell and Norvig 2009). The agent should have executive AI functions that can correlate its world model with inputs from its human counterpart and other external sources as well generate meaningful outputs (Chen et al. 2018). Especially when the agent is assuming a role similar to a human team member, a possible approach is the use of cognitive architectures such as Soar and Adaptive Control of Thought—Rational (ACT-R) to develop agents that emulate human cognitive characteristics such as short-term memory, perception, and so on (Laird et al. 2011; Kelley and McGhee 2013; McNeese et al. 2017). Whereas cognitive architectures are structured to be similar to human thought processes, the underlying world model can be diverse, utilizing a combination of techniques ranging from neural nets to rule-based systems to newer ML techniques (Kelley and McGhee 2013).

For example, more recent AI approaches use ML techniques to solve difficult real-world problems that were previously ill suited to algorithmic approaches based on deductive principles (Everitt and Hutter 2018). These ML paradigms utilize solution sets and feature layers to inductively learn the most efficient mapping between the two based on multiple reinforcement trials. However, a single feature layer could lead to a shallow solution set because there could be multiple mappings to the same solution, whereas *deep learning* approaches have hidden layers, in order to map external inputs to unique solutions (Goodfellow et al. 2016). Hidden layers take advantage of underlying complexity by using Markov models and feedback loops to capture the stochastic nature of real-world processes (Pynadath et al. 2018). Thus, ML approaches take advantage of subtle cues and possible interactions to converge on accurate solutions somewhat like humans discovering patterns in their environment even when they cannot articulate exactly why (Goodfellow et al. 2016). For example, a ML algorithm using a deep learning approach was able to *master* a number of Atari games in a relatively short period of time by uncovering hidden patterns in the games (Everitt and Hutter 2018). In summary, the repertoire of AI techniques is broad enough that agents can use a variety of approaches to help solve problems that human–agent teams encounter in real-world settings. The delimiting factor is integrating the cognitive component into an agent that can articulate its intent to its human partner.

One example of blending various reasoning systems was developed as part of a tri-service project entitled *Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies* (IMPACT) as part of the larger Department of Defense program the *Autonomous Research Pilot Initiative* (ARPI). IMPACT researchers investigated the synergy between a human controller and various intelligent software approaches for planning military missions. IMPACT researchers were able to integrate multiple intelligent software approaches including IAs, ML, and an automated route planner using the *Fusion* architecture developed by the US Air Force Research Laboratory (Draper et al. 2018). They developed their own interface architecture enabling the human operator to interact with the fused AI systems using text, graphics, and NLP voice systems (Calhoun et al. 2018). The end-of-program simulation used active military subject-matter experts to demonstrate the utility of combining human and AI systems to conduct realistic planning missions involving tri-service scenarios (Draper et al. 2018).

The most difficult challenge for IMPACT proved to be integrating the various reasoning modules into a fused architecture so that the human could interact with IMPACT as a cohesive unit. Similarly, McNeese et al. (2017) found that integrating the human and machine element was the chief difficulty in implementing a synthetic UAS crewmember. One aspect of the mismatch between human understanding and agent communications is lack of transparency of AI outputs (Chen et al. 2018).

7. Interpretability Software

There are two paradoxes related to human–agent teaming. The first paradox is that the more proficient the automated component, the more likely the human is to rely on it, even when it is incorrect (Parasuraman and Manzey 2010). The probability of being incorrect may be small, but the results can be disastrous (Parasuraman and Riley 1997). The second paradox has to do with the ability of ML to solve real-world problems. (Goodfellow et al. 2016). The more sophisticated the ML approach is, the more likely the underlying process is to be opaque, resulting in either overtrusting the systems because of its proficiency or mistrusting the system because of its opaqueness (Chakraborty et al. 2017; Pynadath et al. 2018).

Thus, whereas deep learning approaches can result in greater accuracy, they can also lead to greater opacity. A number of recent approaches to XAI are investigating methods to extract the pertinent cues from deep leaning approaches to interpret their meaning for the operator (Chakraborty et al. 2017). In particular, Pynadath et al. (2018) suggest a possible XAI strategy to extract agent SAT information from a ML paradigm that includes hidden Markov (probabilistic) layers.

In summary, bidirectional communication requires a common language framework, mutually interpretable cognitive processes, and an understanding of each other's roles in relation to task objectives (Lyons 2013; Chen et al. 2018). However, as we discuss in the following section, communication protocols that accurately transfer information between humans and agents are necessary for collaboration but are not sufficient. Human-to-human communication depends on implicit understanding of each partner's information requirements to ensure the ideal mix of push and pull of information that defines a well-integrated team (Cooke 2015). As Salas et al. (2015) point out, "teams that communicate effectively may alternate between explicit communication, or overt transmission and acknowledgment of messages, and implicit communication, whereby information is more passively conveyed".

8. Bidirectional Collaboration and Understanding without Communication

Human teams can coordinate their actions and anticipate each other's intentions without overt communications not only because of a common model of the tasking environment but also each team member has a *mental model* of each other's roles and actions (Chen and Barnes 2014; Chen et al. 2018). Specifically, what makes human teammates interact naturally is that each member has a *theory of mind* (TOM), which is the ability "to recognize and attribute mental states—thoughts, perceptions, desires, intentions, feelings—to oneself and to others", allowing humans to internalize what they believe are the thought processes of the other team members (Pedersen 2018). TOM depends not only on innate human traits but also on repeated interactions with their teammates (Astington and Edward 2010; Mahey et al. 2014). It is nontrivial to create a software agent that not only reacts to overt communication cues but that can anticipate its partner's information and action requirements (McNeese et al. 2017).

Fortunately, it is not necessary to duplicate human consciousness in order to develop a world model that enables software agents to interact with its human partners. Pynadath and Marsella (2005) developed the PsychSim architecture to emulate TOM for agents in an urban school environment showing that agents (teacher, students, and a bully) could use perceived attributes of themselves and other agents to determine their interaction patterns. PsychSim allowed the software agents to better understand the intent of the multiple other agents in the scenario. Wang et al. (2016) embedded PsychSim in a predictive framework (partially observable Markov decision process) to enable the robot agents to make recommendations and provide explanations during a complex reconnaissance mission. Human participants trusted agents who were not always reliable if the agent was able to explain its suggestions, indicating that even an imperfect SMM

was better than an opaque agent. An important requisite is that each team member understand each other's intent in terms of achieving a common goal state (Rouse et al. 1992; Evans et al. 2017). Recent research indicates that robot agents can learn social cues and task requirements by repeated interactions with humans, suggesting that self-learning algorithms based on repeated predictive interactions might be a useful way to train agents to interact without overt communications (Kwon 2018). In summary, an agent does not have to have a fully articulated TOM in order to interact with humans. The computer science community is investigating various paradigms such as PsychSim that attempt to emulate human-to-human common understanding of the tasking environment for bidirectional communication (Wang et al. 2018; Pynadath et al. in press). Moreover, self-learning algorithms offer the possibility of training agents to interact with their human operator when trigger events occur in the tasking environment. However, fostering a natural interaction of when to exchange information between humans and IAs during open-ended real-world situations remains an important research objective (McNeese et al. 2017).

Especially when dealing with a physical robot, making the IA more human-like may have a positive effect on human–agent communication. Schaefer et al. (2017) suggest a number of nontechnical agent attributes to harmonize human–agent communication including anthropomorphic features such as robotic gaze, expression, and speech patterns to engender trust and empathy between agents. Similarly, Parasuraman and Miller (2004) suggest using communication styles based on human etiquette rules to improve the human's perception of software agents as trustworthy.

However, it is not certain that making an agent more human-like will have a long-term positive effect versus human perception of the agent's reliability and the human's insight into an agent's reasoning processes (Hancock et al. 2011; Wright et al. 2016). Anthropomorphizing agents can have negative as well as positive effects and, as mentioned, operators are influenced to a greater degree by their perception of agent reliability and transparency rather than depending solely on the human-like qualities of the agent (Hancock et al. 2011; Meyer and Lee 2013).

9. Transparency and Intent Displays

Understanding the intent of their joint task enables the human and the agent to define both their unique and joint roles required to complete the task successfully (Evans et al. 2017; Schaeffer et al. 2017; Chen et al. 2018) The dictionary defines intent in terms of "usually clearly formulated or planned intention" (Merriam-Webster 2018), implying that intent is not only an objective but also a plan to obtain an objective. However, in the military, the term *commander's intent* is used to

convey the commander's overarching objective rather than a specific plan to achieve an objective (Holder 2018). A military intent is usually multifaceted (e.g., obtain an objective in the shortest time with the least loss of life), requiring the command staff to develop various options to reflect tradeoffs during the planning stage, and the term implies that the chosen plan can be disregarded if a better solution or even an alternative objective manifests itself in the heat of battle (Holder 2018).

IMPACT displays were developed to enable IAs and human operators to jointly develop intent displays (Schaefer et al. 2017; Calhoun et al. 2018). IMPACT design philosophy depended on Flexible Levels of Execution – Interface Technologies (FLEXIT), in which, based on circumstances, planning can be done manually, jointly as a human–agent team, or triggered automatically (Miller and Parasuraman 2007; Calhoun et al. 2018). Joint planning using the IMPACT paradigm is initiated by human operators defining a general mission framework (*a play*) and priorities in terms of asset timeliness, sensors, and weapons, thus inputting a multifaceted intent. The autonomous agents then develops plans that are optimized to reflect operators' priorities and then fed back to the operator as a display that shows how each plan option compares to the operator's intent criteria.

Figure 4 depicts an example of a transparent SAT display (L1+2+3) discussed previously delineating two agent plans (A and B) based on the software agent's depiction of the operator's input criteria, allowing operators to choose the final plan closest to their intent (Stowers et al. 2016; Holder 2018). The purpose of the displays is to make the agent's understanding of intent and its resulting COAs transparent to its operator in terms of the plan itself, its rationale, and predicted outcomes (including uncertainties), thus enabling the operator to change the plan or chose an alternate plan generated by the IA depending on the operator's SA of the changing military situation (Mercado et al. 2016; Chen et al. 2018). It is important to note that what is conveyed to the agent is the general outlines of the plan and its intent in terms of multifaceted criteria rather than a fully articulated plan, causing the agent to generate feasible options based on AI techniques to depict the consequences of its generated options using graphics. As the tactical situation changes, communication between the agent and humans allows the humans to change their outcome criteria causing the agent to devise a new plan showing its consequences in terms of the updated tactical situation (e.g., new enemy activity near east gates) (Chen et al. 2018). The joint planning enables the agent's proficiency in rapidly developing a plan based on the operator's priorities and the human's understanding of command intent to devise a plan that is both practical and tactically relevant.

Fig. 4 SAT visualization of alternatives generated by the IAs in terms of each plan's COAs, reasoning, and predicted outcomes/uncertainty (Stowers et al. 2016)

SAT visualizations are dependent on the context of the mission constraints. The Autonomous Squad Member (ASM) is a simulated small robot supporting an infantry squad and similar to IMPACT was also part of the ARPI (Selkowitz et al. 2016). The context of the ASM scenario was geared toward the immediate situation because an infantry squad has to react instantaneously to a volatile combat environment. The intent visualization (Fig. 5) was sparser than the IMPACT displays reflecting that it was designed for immediate reactions and status-at-a-glance SAT information. The pictorial format showed the trajectory of the ASM and its squad whereas icons showed temporal information, the rationale in terms of the chief motivator for the projected squad route, and an uncertainty indicator. As mentioned previously, experimental results showed that SAT information improved operators' calibrated trust and SA (Chen et al. 2018).

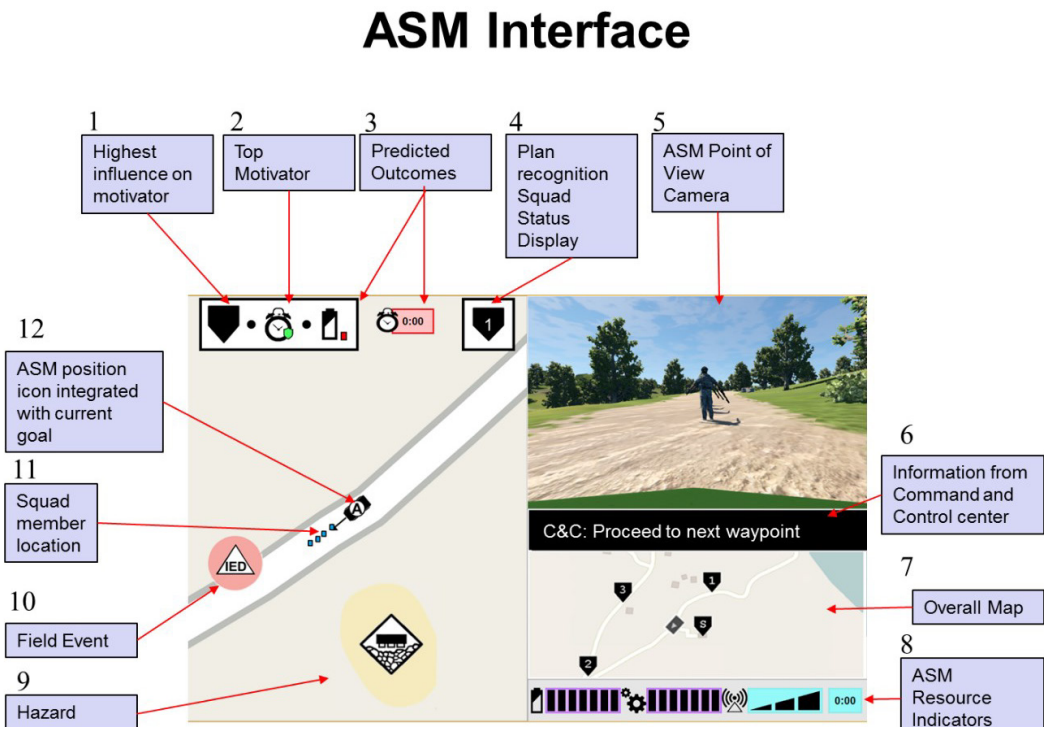

Fig. 5 ASM SAT visualization with annotation

Schaefer et al. (2017) reviewed the importance of intent for agent technology in driverless cars. Engineering intent criteria are used early in the design process to ensure that strict safety and performance objectives are met before the vehicle is introduced to the public The underlying agent technology must not only interact with its user but also with other vehicles and pedestrians whose actions may violate the assumptions of the driverless agent. During driving, an intent display suggesting an immediate action may be counterproductive causing the driver to intervene inappropriately or change from a passive observer to active driver too slowly to avert an accident (cf., Bainbridge 1983). Intent displays in these situation should be more strategic and anticipatory showing possible traffic conditions ahead and pedestrian alerts so as to improve the humans' general SA rather than require humans to override autonomy under extreme time constraints. In general, rapid responses to emergencies are probably best automated, whereas strategic decisions such as avoiding traffic jams or slowing down because there are pedestrian in the street are best supervised by the human (Chen and Barnes 2014; Wright et al. 2016). The efficacy of intent displays depend on their environment: planning displays should be outcome oriented and operational displays should require status-at-glance SA, whereas intent for autonomous vehicles should focus on the anticipating the near term rather than overriding time-sensitive automated responses such as collision avoidance. Parasuraman and Manzey (2010) presciently observed that the human's role in automation is not so much diminished as it is changed.

10. Conclusion

Communication between humans and agents, whether by voice, graphics, or multimodal devices, is best conveyed by communication strategies predicated on understanding common intent (Evans et al. 2017; Schaeffer et al. 2017; Holder 2018). For example, SAT planning displays are formulated to show how the intent of an agent's proposed COAs are manifested in terms of how the proposed plan, its logic, and its projected outcome (and uncertainties) compare to the human's original intent (Mercado et al. 2016; Stowers et al. 2016). Other approaches assume communication is based on a synthetic SMM enabling a robotic agent to explain its COA suggestions in terms of the intents of both human and artificial agents (Pynadath and Marsella 2005; Chen and Barnes 2014; Wang et al. 2016). During a mission, both the human and the agent will need to have the ability to propose new COAs based on their perception of command intent as the combat situation evolves. (Draper et al. 2018). However, creating a seamless flow of information between humans and agents in dynamic environments is still a matter of intense research interest (Chen et al. 2018; Calhoun et al. 2018; McNeese et al. 2017; Wang et al. 2018).

Characterizing humans and intelligent software as teams is useful as long as it is understood as an imperfect metaphor for human–human teams. (Barnes et al. 2017). Because agents lack a TOM, human and agent interactions without explicit information exchanges are still problematic (Salas et al. 2015; McNeese et al. 2017; Kwon 2018). During real-world missions, there are many technical NLP and AI problems related to creating more natural human–like interactions still to be resolved. Sophisticated AI solutions proposed by an agent are useful only if its human partner understands their implications, making XAI and SMM research important considerations (Chakraborty et al. 2017; Pynadath et al. 2018; Wang et al. 2018). The greater precision of AI for specific tasks, the ability of software to respond instantaneously, and the continuous improvements in techniques such as NLP all auger well for human–agent synergistic benefits in the future (Mavridis 2015). However, there are still many research issues both in computer science and human science to be resolved before human–agent communication reaches its full potential as a basis for collaboration during complex real-world missions (Chen et al. 2018).

11. References

Astington JW, Edward MJ. The development of a theory mind in early childhood. Montreal (Canada): Encyclopedia of early childhood; 2010. p. 1–5.

Barber D, Reinerman-Jones L, Matthews G. Toward a tactile language for human–robot interaction: two studies of tacton learning and performance. Hum Factors. 2015;57(3):471–490.

Barber D, Lackey S, Reinerman-Jones L, Hudson I. Visual and tactile interfaces for bi-directional human robot communication. SPIE Defense, Security, and Sensing. 2013:87410U.

Barnes MJ, Chen JYC, Hill SD. Humans and autonomy: implications for shared decision making in military operations. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2017 Apr. Report No.: ARL-TR-7919.

Bradshaw JM, Feltovich P, Johnson M, Breedy M, Bunch L, Eskridge T, Jung H, Lott J, Uszok A, van Diggelen J. From tools to teammates: joint activity in human–agent–robot teams. Paper presented at: International Conference on Human Centered Design; 2009 July 24–26; San Diego, CA.

Calhoun GL, Ruff H, Behymer KJ, Frost EM. Human-autonomy teaming interface design considerations for multi-unmanned vehicle control. TIES. 2018;19(3):321–352.

Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, Srivastava M, Preece A, Julier S, Rao R, et al. Interpretability of deep learning models: a survey of results. Presented at: IEEE Smart World Congress 2017 Workshop: DAIS 2017—Workshop on Distributed Analytics InfraStructure and Algorithms for Multi-Organization Federations; 2017 Aug 7–8; San Francisco, CA.

Chen JYC, Barnes MJ. Human–agent teaming for multirobot control: a review of human factors issues. IEEE T Hum Mach Syst. 2014;44(1):13–29.

Chen JYC, Barnes MJ. Agent transparency for human-agent teaming effectiveness. Proceedings of the IEEE conference on Systems, Man, and Cybernetics (SMC); 2015 Oct; Hong Kong.

Chen JYC, Procci K, Boyce M, Wright J, Garcia A, Barnes MJ. Situation awareness-based agent transparency. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2014 Apr. Report No.: ARL-TR-6905.

Chen JYC, Lakhmani SG, Stowers K, Selkowitz A, Wright J, Barnes M. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. TIES. 2018;19(3):259–282. doi: 10.1080/1463922X.2017.1315750.

Cooke NJ. Team cognition as interaction. CDPS. 2015;24(6):415–419.

Draper M, Rowe A, Douglass S, Calhoun G, Spriggs S, Kingston D, et al. Realizing autonomy via intelligent adaptive hybrid control: adaptable autonomy for achieving UxV RSTA team decision superiority (also known as Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies (IMPACT). Dayton (OH): Air Force Research Laboratory (US); 2018. Report No.: AFRL-RH-WP-TR-2018-0005.

[Economist] March of the machines: a special report on artificial intelligence. The Economist. 2016 July.

Elliott L, van Erp J, Redden E, Duistermaat M. Field-based validation of a tactile navigation device. IEEE Trans Haptics. 2010;3(2):78–87.

Endsley MR. Situation awareness misconceptions and misunderstandings. JCEDM. 2015;9:4–15

Evans A, Marge M, Stump E, Warnell G, Conroy J, Summers-Stay D, et al. The future of human robot teams in the Army: factors affecting a model of human-system dialogue towards greater team collaboration. In: Savage-Knepshield P, Chen J, editors. Advances in human factors in robots and unmanned systems. Cham, Switzerland: Springer Open; 2017. p. 197–209. (Advances in Intelligent Systems and Computing. vol. 499).

Everitt T, Hutter M. Universal artificial intelligence. In: Abbass M, Scholz J, editors. Foundations of trusted autonomy. Cham (Switzerland): Springer Open; 2018. p. 15–47.

Giammanco C, Mott M, McGowan R. Controlled English for critical thinking about the civil-military domain. Presented at the International Technology Alliance in Network and Information Sciences Annual Fall Meeting; 2015.

Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge (MA): MIT Press; 2016. http://www.deeplearningbook.org.

Hancock PA, Billings DR, Schaefer KE, Chen JYC, de Visser E, Parasuraman R. A meta-analysis of factors affecting trust in human-robot interaction. Hum Factors. 2011;53(5):517–527.

Holder E. Defining soldier intent in soldier-natural language processing interaction. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2018 Apr. Report No.: ARL-TR-8919.

Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River (NJ): Prentice-Hall; 2009.

Kahneman D. Thinking, fast and slow. New York (NY): Farrar, Straus & Giroux; 2011. p. 499.

Kelley TD, McGhee S. Combining metric episodes with semantic event concepts within the symbolic and sub-symbolic robotics intelligence control system (SS-RICS). SPIE Defense, Security, and Sensing. 2013 May. p. 87560L.

Kelley TD. Robotic dreams: a computational justification for the post hoc processing of episodic memories. Int J Mach Conscious. 2014;6(2):109–123.

Kwon D. Machines that learn like children provide deep insight into how the mind and body interact together to bootstrap knowledge and skills. Sci Am. 2018 Mar. p. 27–32.

Laird J, Derbinsky N, Voigt J. Performance evaluation of declarative memory systems in soar. Proceedings of the Behavior Representation in Modeling and Simulation Conference; 2011 Mar 21–24; Sundance, UT. p. 33–40.

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

Lyons JB. Being transparent about transparency: a model for human–robot interaction. In: Sofge D, Kruijff GJ, Lawless WF, editors. Trust and Autonomous Systems: Papers from the AAAI Spring Symposium (Technical Report SS-13-07). p. 48–53. Menlo Park (CA): AAAI Press; 2013.

Oron-Gilad T. Scalable interfaces for operator control units: common display to conduct MOUT operations with multiple video feeds (final research report). Be'er Sheva (Israel): Ben Gurion University of the Negev; 2014.

Mahey CEV, Moses LJ, Pfiefer JH. How and why: theory of mind in the brain. Dev Cogn Neuros 9. 2014;68–81.

Marathe AR, Metcalfe JS, Lance BJ, Lukos J, Jangraw D, Kuan-Ting L, Touryan J, et al. Privileged sensing framework: a principled approach to improved human-autonomy integration. TIES. 2018;19(3);283–320.

Mavridis N. A review of verbal and non-verbal human–robot interactive communication. Robotics Autonomous Sys. 2015;63(1):22–35.

McNeese N, Mustifa D, Cooke N, Myers C. Teaming with a synthetic teammate: insights into human-autonomy teaming. Hum Factors. 2017;60(2):262–273.

Mercado J, Rupp M, Chen J, Barber D, Procci K, Barnes M. Intelligent agent transparency in human–agent teaming for multi-UxV management. Hum Factors. 2016;58(3):401–415.

Miller C, Parasuraman R. Designing for flexible interaction between humans and automation: delegation interfaces for supervisory control. Hum Factors. 2007;49:57–75.

[Merriam-Webster] Intent. Springfield (MA): Merriam-Webster; 2018 [accessed 2018 July 24]. https://www.merriam-webster.com/dictionary/intent.

Mott D, Shemanski R, Giammanco C, Braines D. Collaborative human–machine analysis using a controlled natural language. SPIE Next-Generation Analyst III. 2015 Jan 1;9499. doi: https://doi.org/10.1117/12.2180121.

Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. Hum Factors. 1997;39:230–253.

Parasuraman R, Miller C. Trust and etiquette in high-criticality automated systems. Commun ACM. 2004;47:51–55.

Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. 2010;52(3):381–410.

Parasuraman R, Barnes M, Cosenzo K. Adaptive automation for human robot teaming in future command and control systems. International Journal of Command and Control. 2007;1(2):43–68.

Pedersen T. Theory of mind. Newburyport (MA): Psych Central; 2018 Dec 10 [accessed 2018 July 18]. https://psychcentral.com/encyclopedia/theory-of-mind.

Pettit RA, Redden ES, Carstens CB, Hooper D. Scalability of robotic controllers: effects of progressive autonomy on intelligence, surveillance, and reconnaissance robotic tasks. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2013 Jan. Report No.: ARL-TR-6182.

Pynadath DV, Marsella SC. PsychSim: Modeling theory of mind with decision-theoretic agents. Proceedings of the International Joint Conference on Artificial Intelligence, 2005. p. 1181–1186.

Pynadath DV, Barnes MJ, Wang N, Chen JYC. Transparency communication for machine learning in human-automation interaction. In: Zhou J, Chen F, editors. Human and machine learning. Human–computer interaction series. Springer, Cham; 2018.

Pynadath DV, Wang N, Barnes MJ. Presented at the 23rd European Conference on Artificial Intelligence; in press; Stockholm Sweden.

Rouse WB, Cannon-Bowers JA, Salas E. The role of mental models in team performance in complex systems. IEEE Transactions on Systems, Man, & Cybernetics. 1992;22:1296–1308.

Russell SJ, Norvig P. Artificial intelligence: a modern approach. 3rd ed. Saddle River (NJ): Upper Prentice Hall; 2009.

Salas E, Shuffler ML, Thayer AL, Bedwell WL, Lazzara EH. Understanding and improving teamwork in organizations: a scientifically based practical guide. In: Salas E, Rico R, Passmore J, editors. The Wiley Blackwell handbook of the psychology of team working and collaborative processes. NY: Wiley, 2015.

Scerri P, Pynadath D, Johnson L, Rosenbloom P, Si M, Schurr N, Tambe M. A prototype infrastructure for distributed robot–agent–person teams. AAMAS03: Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems; 2003 July 14–18; Melbourne, Australia.

Schaefer KE, Straub ER, Chen JYC, Putney J, Evans AW. Communicating intent to develop shared situation awareness and engender trust in human–agent teams. Sci Dir Cog Sys Res. 2017;46:26–39.

Selkowitz A, Lakhmani S, Larios C, Chen JYC. Agent transparency and the autonomous squad member. Presented at: 2016 International Annual Meeting of the Human Factors and Ergonomics Society; 2016 Sep 19–23; Washington DC.

Skubic M, Perzanowski D, Blisard S, Schultz A, Adams W, Bugajska M, Brock D. Spatial language for human–robot dialogs. IEEE Trans Systems Man Cybernetics Part C App Rev. 2004;34(2):154–167.

Stowers K, Kasdaglis N, Rupp M, Chen J, Barber D, Barnes M. Insights into human–agent teaming: intelligent agent transparency and uncertainty. Presented at: Applied Human Factors and Ergonomics (AHFE) Conference; 2016 July 27–31; Orlando, FL.

Tellex S, Kollar T, Dickerson S, Walter M, Banerjee A, Teller S, Roy N. Understanding natural language commands for robotic navigation and mobile manipulation. AAAI 2011. Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence; 2011 Aug 7–11; San Francisco, CA: Palo Alto (CA): AAAI Press; c2011. p. 1507–1514.

Wang N, Pynadath D, Rovira E, Barnes MJ, Hill SJ. Is it my looks? Or something I said? The impact of explanations, embodiment, and expectations on trust and performance in human–robot teams. International Conference on Persuasive Technology; 2018: Springer; c2018. p. 56–69.

Wang N, Pynadath D, Hill S. Trust calibration within a human–robot team: comparing automatically generated explanations. Proceedings of the 11th ACM/IEEE International Conference on Human Robot Interaction; 2016 Mar 7–10; New Zealand. p. 109–116.

Wright JL, Chen JYC, Barnes MJ, Hancock PA. Agent reasoning transparency: the influence of information level on automation-induced complacency. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2016 June. Report No.: ARL-TR-8044.

Xue P, Mott D, Giammanco C, Copestake A. An approach to handling linguistic and conceptual difference across coalition partners. Presented at: International Technology Alliance in Network and Information Sciences Annual Fall Meeting; 2015 Sep; College Park, MD.

Yena J, Fana S, Suna S, Hanratty T, Dumerb J. Agents with shared mental models for enhancing team decision makings. Dec Sup Sys. 2006;41:634– 653.

List of Symbols, Abbreviations, and Acronyms

AI	artificial intelligence
ARPI	Autonomous Research Pilot Initiative
ASM	Autonomous Squad Member
CE	Controlled English
COA	course of action
FLEXIT	Flexible Levels of Execution – Interface Technologies
IA	intelligent software agents
IMPACT	Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies
L1	agent's perception of its plan
L2	agent's perception of its logic
L3	agent's perception of its perceived likelihood
L	transparency levels
ML	machine learning
NLP	natural language processing
SAT	SA-based Agent Transparency
SA	situation awareness
SMM	shared mental model
TOM	theory of mind
UAS	unmanned aerial system
XAI	explainable AI

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

1 (PDF) GOVT PRINTG OFC A MALHOTRA

2 (PDF) DIR ARL IMAL HRA RECORDS MGMT RDRL DCL TECH LIB

1 (PDF) ARMY RSCH LAB – HRED RDRL HRB B T DAVIS BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898-7290

1 (PDF) USA ARMY G1 DAPE HSI M SAMS 300 ARMY PENTAGON RM 2C489 WASHINGTON DC 20310-0300

1 (PDF) USAF 711 HPW 711 HPW/RH K GEISS 2698 G ST BLDG 190 WRIGHT PATTERSON AFB OH 45433-7604

1 (PDF) USN ONR ONR CODE 341 J TANGNEY 875 N RANDOLPH STREET BLDG 87 ARLINGTON VA 22203-1986

1 (PDF) USA NSRDEC RDNS D D TAMILIO 10 GENERAL GREENE AVE NATICK MA 01760-2642

1 (PDF) OSD OUSD ATL HPT&B B PETRO 4800 MARK CENTER DRIVE SUITE 17E08 ALEXANDRIA VA 22350

ABERDEEN PROVING GROUND

15 (PDF) ARL RDRL HR J LANE Y CHEN P FRANASZCZUK K MCDOWELL K OIE RDRL HRB A F MORELLI RDRL HRB C L GARRETT RDRL HRB D D HEADLEY RDRL HRB DE M Y BARNES E HOLDER RDRL HRF A A DECOSTANZA RDRL HRF B A EVANS RDRL HRF C J GASTON RDRL HRF D A MARATHE S LAKHMANI