A study measuring the impact of shared decision making in a human-robot team

Journal Title
XX(X):1–21

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Mohammad Q. Azhar[1] and Elizabeth I. Sklar[2]

Abstract

This paper presents the results of a user study in which the impact of sharing decision making in a human-robot team was measured. In the experiments outlined here, a human and robot play a game together in which the robot searches an arena for items, with input from the human, and the human-robot team earns points for finding and correctly identifying the items. The user study reported here involved 60 human subjects. Each subject interacted with two different robots. With one robot, the human acted as a *supervisor*: the human issued commands and the robot obeyed. With the other robot, the human acted as a *collaborator*: the human and robot shared decisions and were required to reach agreement about the robot's actions in the arena before any actions were taken, facilitated using *computational argumentation*. Objective performance metrics were collected and analysed for both types of human-robot team, as well subjective feedback from human subjects regarding attitudes toward working with a robot. The objective results showed significant improvement in performance metrics with the human-as-collaborator pairs versus the human-as-supervisor pairs. Subjective results demonstrated significant differences across many subjective measures and indicated a distinct preference for the human-as-collaborator mode. The primary contribution of this work lies in the demonstration and evaluation of a computational argumentation approach to human-robot interaction, particularly in proving the efficacy of this approach over a less autonomous mode of interaction.

Keywords

Shared decision making, argumentation-based dialogue, human-robot team

Introduction

Humans interact with each other in many different ways. Some modes of interaction are *supervisory*, where one person commands another, while other modes are *collaborative*. In a supervisory interaction, one person takes responsibility for making decisions about joint actions and actions that affect others, whereas in a collaborative interaction, partners share decision making. Collaborators exchange ideas and discuss options. Together, they reach *agreement* about actions that depend on or are related to each other. Agreement is facilitated through conversation, or *dialogue*, in which each partner communicates ideas and adjusts their beliefs according to new or altered ideas presented by others. Some *human-robot interaction (HRI)* modes are less autonomous and involve a human supervisor who maintains the locus of control and tells the robot what to do (at varying levels of detail). The human leader sets overall goals and assigns tasks to the robot designed to achieve those goals. The robot may then define its own series of subgoals in order to accomplish its assigned tasks. If, for example, a human tells a robot to go to a particular location, the robot will execute path-planning behaviours to select waypoints and motion behaviours to travel to each waypoint, as well as collision-avoidance behaviours in order to arrive safely.

The restrictive mode of supervisory interaction limits the potential robustness of a human-robot team because it does not take full advantage of the sensory, processing or reasoning capabilities of the robot. If a robot fails at its assigned task, it will generally only be able to report that failure has occurred and be unable to elaborate on the reason(s) for its failure. For example, if a robot cannot go to a destination assigned by its human supervisor because there is a fire blocking egress, the typical robot cannot engage the human in discussion about alternative goals. In addition, most robots cannot request assistance to accomplish tasks that they fail to complete. Some work has demonstrated the effectiveness of a robot that asks for help from a human when it determines that without help, it will fail to complete its assigned task (Rosenthal et al. 2010). In that work, the locus of control for task completion and responsibility for actions shifts from the robot to the human, allowing the robot to request assistance from the human and accomplish tasks it would otherwise have been unable to tackle.

We are interested in situations where the initiative for collaboration can emanate from *either* the human or the robot, where discussion about actions can ensue, *reasons* for and against taking particular actions can be passed back and forth, and the responsibility for actions can flow to either the human or the robot, until the task is completed. For example, suppose a human-robot team is asked to fetch an orange ball from an unfamiliar environment. If the human

[1]The City University Of New York, USA
[2]King's College London, UK

Corresponding author:
Dr. Elizabeth I. Sklar, Department of Informatics, King's College London, Strand, London, WC2R 2LS, UK.
Email: elizabeth.sklar@kcl.ac.uk

asks a robot to help her look for such a ball which she can pick up, the robot may wander around their environment taking pictures and sending them to the human, without knowing whether it has captured an image of a ball or a round fruit that is orange-coloured. Feedback from the human about the content of the image would improve the likelihood that the human-robot team has obtained a picture of an orange ball, because humans have better abilities to distinguish between items that closely resemble each other. Feedback from the robot about the location where the image was taken and directions for how to get there would improve the likelihood that the human-robot team can retrieve the ball, because robots have better abilities to map their environment and perform path-planning using the map. If the robot and human disagree about an aspect of their task, the ability to communicate the reasons for their individual opinions—perhaps the *evidence* that led each to reach their own *conclusion*—can be invaluable in resolving the conflict. Thus, rich discussion enables *shared decision making* about the content of candidate images, *agreement* about an image that indeed contains an orange ball and *guidance* for retrieving the ball.

In the study presented here, we seek to measure the differences for a human-robot team operating in two distinct modes. The first mode is the less autonomous "human-as-supervisor" relationship, where the human issues commands and the robot executes them. The second mode is a "human-as-collaborator" relationship, where the human and robot share decision making. This latter mode is enabled using *computational argumentation-based dialogue* (Walton and Krabbe 1995), which is adapted here to provide the ability for the human and robot to seek agreement concerning aspects of their joint task (Sklar and Azhar 2015; Sklar et al. 2013a). Objective performance metrics were collected and analysed for both types of human-robot team, as well subjective feedback from human subjects regarding attitudes toward working with a robot. The objective results showed significant improvement in performance metrics with the human-as-collaborator pairs versus the human-as-supervisor pairs. Subjective results demonstrated significant differences across many subjective measures and indicated a distinct preference for the human-as-collaborator mode. The primary contribution of this work lies in our demonstration and evaluation of a *computational argumentation approach to human-robot interaction (HRI)*. Not only are we the first to provide an implementation of *computational argumentation applied to HRI,* but also we are the first to evaluate this approach with physical robots and to compare our results with a simulated robot environment.

The remainder of this article is organised into seven sections, as follows. First, the **Background** section defines *computational argumentation*[*] and *argumentation-based dialogue,* particularly for those readers unfamiliar with this sub-area of artificial intelligence and philosopny. As explained in this section, we discuss the notion of "dialogue" exclusively in the sense of *computational argumentation-based dialogue;* we do not imply any application of or contribution to natural language dialogue. Second, the **Approach** section outlines the underlying methodology and implementation of our computational argumentation-based dialogue framework for human-robot interaction. Third, the **Experiments** section describes the user study that we conducted in order to demonstrate the effectiveness of our computational argumentation approach to shared decision making in a human-robot team. This is followed by the **Results** section, which presents outcomes of the user study. Then, the **Related Work** section highlights some relevant literature to illustrate our contributions. Finally, the **Conclusion** section closes with a summary, discussion of our results, and directions for future work.

Background

The theory of interaction adapted for the framework employed in the work presented here comes from **computational argumentation** (Rahwan and Simari 2009), which is a logic-based formal methodology for structuring evidence in support of (or attacking) specific conclusions and has its roots in philosophy and artificial intelligence. Building on this methodology, **computational argumentation-based dialogue** (Walton and Krabbe 1995; Hulstijn 2000; McBurney and Parsons 2002; Prakken 2006) is a formalism in which participants engage in goal-oriented exchange following specific protocols. In our work—as mentioned earlier—we consider the notion of "dialogue" exclusively in the sense of *computational argumentation-based dialogue;* we do not imply any application of or contribution to natural language dialogue[†]. Computational argumentation-based dialogue theory arises from studying people, where early research in the computational argumentation community identified six primary types of dialogue (Walton and Krabbe 1995). These are distinguished, based on participants' knowledge and their individual and shared goals, specifically: *information-seeking* (Walton and Krabbe 1995) (where one participant asks a question that she does not know the answer to and believes the other participant can answer), *inquiry* (McBurney and Parsons 2001b) (where both participants seek an answer to a question that neither knows the answer to), *persuasion* (Prakken 2006) (where one participant tries to alter the beliefs of another participant), *negotiation* (Rahwan et al. 2003) (where participants bargain over the allocation of a scarce resource), *deliberation* (McBurney and Parsons 2004) (where participants decide together on taking a particular action) and *eristic* (Walton and Krabbe 1995) dialogues (where partipants quarrel verbally). Other types of argumentation-based dialogue have followed in the literature, including: *command* (Girle 1996) (where one participant tells another what to do), *chance discovery* (McBurney and Parsons 2001a) (where a new idea arises out of the discussion between participants), and *verification* (Cogan et al. 2005) (where one participant asks a question that she already knows the answer to and she believes the other participant also knows the answer, so her aim is to verify her belief).

[*] *Computational argumentation* is also referred to as *logical argumentation*, or simply *argumentation*, in the artificial intelligence and philosophy literatures.

[†] While we agree that natural language dialogue will be important in any fully autonomous human-robot system, we have not chosen that area in which to focus our research.

Argumentation-based dialogue theory has been studied by researchers in a number of disciplines, including Law, Medicine, Artificial Intelligence and Multi-Agent Systems (Nielsen and Parsons 2006; Bench-Capon and Dunne 2007; Medellin-Gasque 2013). Argumentation-based dialogues have been developed to help two agents make decisions about their *goals* and *plans*. Two co-operative agents who share a goal will only accept plans that are aligned with their beliefs. Plan-based dialogue models have been developed using a belief, desire and intention (BDI) architecture (Wobcke et al. 2006) in the agent-oriented software engineering community. Belesiotis *et al.* (Belesiotis et al. 2010) developed an abstract argumentation-based protocol that allows two agents to discuss their proposals until an agreement is reached through the persuasion-aligned planning beliefs of the agents.

A number of researchers have investigated and applied specific dialogues to a range of different domains and problems. Black & Atkinson (Black and Atkinson 2011) developed a dialogue framework where two agents employ *persuasion* dialogue to discuss how to act. Black (Black 2007) demonstrated how a general argumentation-based dialogue framework can be applied in the medical domain where two doctors engage in *inquiry* dialogues to expand their knowledge and make better decisions. Black & Hunter (Black and Hunter 2009) developed this theoretical framework for two collaborative agents to engage in inquiry dialogue as a means to expand their joint knowledge. Tang *et al.* (Tang and Parsons 2005) developed formal mechanisms to employ *deliberation* dialogue for discussing actions. Agents decide what actions to undertake and the order in which the actions should be performed, combining both agents' knowledge and overlapping expertise. In later work, Tang *et al.* (Tang et al. 2010b) developed a formal argumentation model to generate plans for a team that operates in a non-deterministic environment. Medellin-Gasque *et al.* (Medellin-Gasque et al. 2012) developed a formal argumentation model for two common goal-sharing autonomous agents. In this model, the agents decide on a plan in which they will propose, justify and share information about plans, engaging in argumentation-based persuasion and negotiation dialogue.

Only a handful of projects have implemented interactive argumentation-based systems and tested them with human users. To our knowledge, nobody else has tested them with robots. Tang *et al.* (Tang et al. 2010a, 2011, 2012b) developed an argumentation engine called *ArgTrust*, which is based on a formal argumentation framework for agents to reason about their beliefs, their level of trust in information from which those beliefs are derived and in the source(s) of that information. Sklar *et al.* (Sklar et al. 2015) evaluated a prototype implementation of ArgTrust to study how people reason and make decisions in uncertain situations and how they explain their decisions. The results of the user study involving 22 participants indicate that an argumentation-based system such as ArgTrust can help humans carefully formulate their decisions. Toniolo *et al.* (Toniolo et al. 2015a,b) developed *CISpaces*, an argumentation-based tool designed to assist intelligence analysts in making sense of complex and incomplete information. The system allows analysts to exchange arguments encoded through *argument schemes* (Walton et al. 2008), which are formally coded common-sense patterns of reasoning.

We have applied computational argumentation theory to human-robot interaction, using argumentation-based dialogue as the means to provide support for answering a question, to aid in the diagnosis of errors due to hardware or software failure or unexpected changes in the environment, or to resolve conflicts (Sklar et al. 2013a; Sklar and Azhar 2015). We developed the *ArgHRI* software framework (Azhar 2015), which employs a version of the ArgTrust engine (Tang et al. 2012b) and a *control layer* (Parsons and McBurney 2003), which sits on top of ArgTrust, to manage multiple layers of argumentation-based dialogue. Next, we describe ArgHRI, followed by presentation of the experimental results that are the focus and primary contribution of this article. To our knowledge, we are the first to apply computational argumentation-based dialogue for human-robot collaboration, from implementation of the formal theory to evaluation with physical robots and human subjects.

Approach

Our approach enables a human-robot team to expand and share knowledge and make decisions together during execution of a shared task. *ArgHRI* is the software framework that we have implemented to support these activities. The framework comprises:

- a **belief system** for describing the robot's environment and capabilities (domain dependent) and a rule-based system to maintain its beliefs (domain independent);
- a domain independent **argumentation engine** that applies computational argumentation to calculate the support for a specific conclusion based on the robot's beliefs *(ArgTrust)* (Tang et al. 2012b);
- a domain independent **argumentation-based dialogue system** that implements the dialogue theory described in the previous section;
- a **robot operating environment**, which provides a control architecture for deploying physical and simulated robots *(HRTeam)* (Sklar et al. 2011);
- a **game engine** that provides a domain-dependent facility for the experimental task domain, the *Treasure Hunt Game (THG)*; and
- a **user interface** and **game client** that enable interaction between a human user and the robot.

Each of these components is discussed in this section, in the order in which they are listed above.

Belief system

ArgHRI employs a belief system structured around a simplified ontology for describing the robot's beliefs and a rule-based methodology for maintaining the robot's beliefs. The ontology includes a set of predicates that describe the robot's internal state, the state of its environment (Table 1) and its possible actions and capabilities (Table 2). The rule-based methodology is a formal *dialogue game* (McBurney and Parsons 2002), which allows the robot to manage its beliefs in a structured way and is expressed using

At(t, loc)	which is $true$ if, at time t, the robot is in location loc
Battery(t, s)	which is $true$ if, at time t, the robot reports battery status s (low, high); here the battery status is set to low when it only has 20% battery and set to high when it has more than 80% battery.
Found(t, obj)	which is $true$ if, at time t, the robot senses object obj
ObjectAt(t, loc, obj)	which is $true$ if, at time t, object obj is at location loc

Table 1. Example predicates that describe the robot's beliefs.

GoTo(t, loc)	t is the time at which the action begins; loc is a location
Stop(t)	t is the time at which the robot ceases motion
Sense(t)	t is the time at which the robot senses an object

Table 2. Example predicates that describe the robot's actions.

fundamental elements of computational argumentation as specified in related work (Parsons et al. 2003; Sklar and Parsons 2004). These elements include:

$R.\Sigma$	=	robot's beliefs about itself and its environment
$R.\Gamma(H)$	=	robot's beliefs about the human
$R.CS$	=	robot's *commitment store* (record of utterances)
$H.CS$	=	human's commitment store
$R.\Delta$	=	full set of beliefs that the robot can reason with: $R.\Sigma \cup R.\Gamma(H) \cup R.CS \cup H.CS$

The robot's set of internal beliefs ($R.\Sigma \cup R.\Gamma(H)$) is considered *private*, whereas its commitment store, $R.CS$, contains all its utterances in a dialogue; so we consider the commitment store to be *public*. The robot has access not only to its own commitment store, but also to the commitment store of the human with whom it is engaged in a dialogue. Both $R.CS$ and $H.CS$ are stored by the robot, which means that the robot can update its beliefs about the human, $R.\Gamma(H)$, based on $H.CS$, what the human has said in the dialogue. The full set of beliefs that the robot has access to for reasoning, $R.\Delta$, comprises not only its own beliefs, but also $H.CS$, which allows the robot, during a dialogue, to repeat utterances the human makes or ask for clarification about them, without formally committing them to the robot's internal set of beliefs ($R.\Sigma$). This formal commitment step, in which the robot's beliefs are *revised* and $R.\Sigma$ is updated, occurs at the end of a dialogue. The rules for belief revision are incorporated in the protocols specific to each dialogue locution, detailed in the next section.

There are two important features of the ArgHRI belief system. First, we do not make any attempt to create or maintain a true and complete model of the human's beliefs. We only model what the *robot believes the human believes*, $R.\Gamma(H)$, based on $H.CS$, what the human has said in the dialogue. For the purposes of the experiments described here, we assume that the human tells the truth, though we acknowledge that such an assumption puts limitations on our system for real-world deployment. There is a growing literature on the use of argumentation to model untruths (Caminada 2009), one area for future investigation. Second, following the rules of dialogue games (McBurney and Parsons 2002), participants can only utter beliefs that they can *support* in the formal sense of argumentation. This means that a participant is allowed to utter b, where b is represented as an atomic *fact* in the participant's internal set of beliefs, or c, where c is a conclusion that can be drawn from a set of arguments (S, c). Here, S is considered the *support* for the conclusion c, such that each element of S is either an atomic fact (like b) or can be derived from rules that are part of the participant's set of beliefs. For example, the rule:

$$\mathsf{GoTo}(t, loc) \rightarrow \mathsf{At(t+1, loc)}$$

says that if the predicate on the left is true, then the predicate on the right can be concluded as the result of applying the rule. Note that an atomic fact, b, can represent fixed knowledge about the world or about the domain or the robot's environment, all of which could be established *a priori* and remain unchanged during a robot's mission; as well, b can represent information about the robot or its environment that does change during a mission, e.g., the output of a Sense(t) action. This restriction on legal utterances is why it is important that $R.\Sigma$ includes $H.CS$.

Argumentation engine

ArgHRI employs the computational argumentation engine *ArgTrust* (Tang et al. 2012b), which is a partial implementation of the formal system from (Tang et al. 2012a), to calculate the support for a specific conclusion based on the robot's beliefs, as exemplified above. This is a critical component of ArgHRI because it is used by the robot to reason about how a goal might be achieved or to resolve conflicts found within its internal set of beliefs ($R.\Sigma$) or between its beliefs and its beliefs about the human's beliefs ($R.\Gamma(H)$). Conflicts can be formally computed in two ways (Sklar and Azhar 2015): *undermining*, where the human's conclusion of an argument conflicts with the conclusion of the robot's arguments or vice versa; and *rebuttal*, where the conclusion of the human's argument conflicts with some element in the support of the robot's argument or vice versa.

ArgHRI communicates with ArgTrust using an API in which ArgHRI sends a set of inputs (beliefs and rules) and a query (a conclusion to assess) to ArgTrust. ArgTrust searches the inputs for evidence that supports, or attacks, the conclusion. The inputs include not only the robot's beliefs about the specific situation, but also knowledge about the domain (e.g., a map of the robot's environment) and world knowledge (e.g., "North is north of South"). The output from ArgTrust is an *accept* predicate, containing the evidence that supports the conclusion; a *defeat* (*reject*) predicate, containing the evidence that attacks the conclusion; or an *undecided* predicate, containing the conflicting evidence that prevents the system from either accepting or defeating the conclusion.

Two comments should be interjected here. First, the choice of ArgTrust for computing arguments in ArgHRI was made largely because the system is one of the few implementations of formal argumentation that can easily be integrated into another system as a standalone computation engine. Second, the specific way in which beliefs are represented provokes questions about the scalability of the system. However, this is a general problem for AI-based systems, many of which employ knowledge representations that will not easily scale. Addressing this issue is beyond the scope of the research focus here.

Argumentation-based dialogue system

The argumentation-based dialogue system developed for ArgHRI is an implementation of a *dialogue game* (McBurney and Parsons 2002), which originates from well-founded argumentation-based theory and provides an alternative approach to structuring dialogue, supporting less restrictive conversation policies and more efficient communication (Black 2007). Our formal model is adapted from (McBurney and Parsons 2002; Parsons et al. 2003), which prescribes a structure for defining *protocols* for utterances associated with each type of dialogue, essentially a list of legal moves that each participant in the dialogue game is allowed to select from. In related work, we have defined the protocols for each of the types of dialogue implemented in ArgHRI (Sklar and Azhar 2015; Sklar et al. 2013a), including state machines that control the possible sequences of locutions, from start state to termination state. One of the advantages of employing argumentation-based dialogue games is that it has been proven, formally, that the rules for each type of dialogue guarantee termination (Parsons et al. 2003).

Three types of argumentation-based dialogue have been implemented in ArgHRI:

- **information-seeking:** where one participant seeks answers to questions from another participant, who is believed by the initiating participant to know the answers;
- **inquiry dialogue:** where the participants collaborate to answer a question or questions whose answers are not known to any participant; and
- **persuasion dialogue:** where one participant seeks to persuade another party with a different opinion to adopt a belief or point-of-view.

Here are some examples of human-robot scenarios where each type of dialogue may be applied. The robot could ask the human for information that the robot does not have and believes that the human has, in order to prevent errors; this is an example of an *information-seeking* dialogue. The robot and human may agree to seek an answer to an unknown query because neither of them has enough information to make an informed decision; this is an example of an *inquiry* dialogue. The human could suggest that the robot follow her plan and discard its own plan, in order to pre-empt failure predicted by the human; this is an example of a *persuasion* dialogue. The robot might discover information that the human does not possess or that contradicts something that the robot believes the human believes, in order to correct the human's misconception(s) and pre-empt possible failure; this is another example of a *persuasion* dialogue.

Table 3 lists the conditions under which the robot can initiate each of these three types of dialogue. The conditions consider a belief, b, and make a decision based on the membership of b in two elements of the robot's belief set: its internal beliefs ($R.\Sigma$) and its beliefs about the human's beliefs ($R.\Gamma(H)$).

	$b \in R.\Gamma(H)$	$\neg b \in R.\Gamma(H)$	$?b \in R.\Gamma(H)$
$b \in R.\Sigma$	agreement *(no dialogue)*	disagreement *(persuasion)*	lacking knowledge *(information-seeking)*
$\neg b \in R.\Sigma$	disagreement *(persuasion)*	agreement *(no dialogue)*	lacking knowledge *(information-seeking)*
$?b \in R.\Sigma$	lacking knowledge *(information-seeking)*	lacking knowledge *(information-seeking)*	shared lack of knowledge *(inquiry)*

Table 3. Cases for different types of dialogues (from (Sklar and Azhar 2015)). The notation $b \in R.\Sigma$ means that the robot believes b (or $\neg b$, accordingly). If the membership is in $R.\Gamma(H)$, then the meaning is that the robot believes that the human believes b (or $\neg b$). The concept that a participant does not have any knowledge about b, i.e., is unable to decide whether they believe b or $\neg b$, is represented by $?b$.

A *control layer* manages the dialogue game and incorporates the following components outlined by (McBurney and Parsons 2003, 2009):

- **commencement rules:** a set of rules that defines the *pre-conditions* or circumstances under which the dialogue can begin;
- **locutions:** the complete set of possible moves consisting of statements (*utterances* or *locutions*) issued by one player (participant) and directed toward the other player;
- **combination rules:** a set of rules (*protocols*) that governs which moves a player can make in each dialogical context;
- **commitments:** a set of rules that define the circumstances under which each player expresses commitment to a proposition and a public commitment store for each player (i.e., $R.CS$);
- **speaker order:** a set of rules that defines the order in which a speaker may make utterances; and
- **termination rules:** a set of rules that enable a dialogue to reach a termination condition, where either both players agree, by *accepting* the same proposition, or both players reach a *stalemate*, by failing to accept the same proposition exhausting all possible moves.

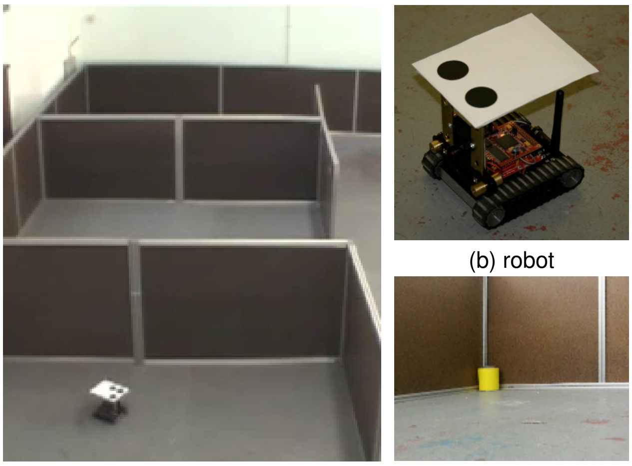
Note that *acceptance* can mean that both players agree to either b or $\neg b$. An individual player cannot commit to both b and $\neg b$. A commitment store is updated after every utterance ($R.CS$ and $H.CS$), and the robot's beliefs ($R.\Sigma$

and $R.\Gamma(H)$) are updated after the termination of each dialogue.

Robot operating environment

The robot actualised for ArgHRI is controlled through the *HRTeam* robot operating environment (Sklar et al. 2011, 2013b). HRTeam was developed to support experimental research in human/multi-robot interaction in both physical and simulated environments. This software framework was employed in ArgHRI in a single-robot mode. The HRTeam framework is structured around a client-server architecture that comprises an *agent* layer (for intelligent reasoning), a robot layer (for actuating robot behaviours) and a centralised server for passing messages between nodes on each layer. The robot layer is built on *Player/Stage* (Gerkey et al. 2003), which supports easy transfer between systems employing physical or simulated robots. As detailed in the Experiments section, later in this article, ArgHRI employed both the physical and simulation operating modes of HRTeam.

The architecture of ArgHRI is integrated into HRTeam and provides the intelligent agent layer as the implementation of the decision-making capabilities of the robot. In other words, the R entity mentioned above is implemented in this layer. Figure 1 shows the HRTeam arena (a) and robot (b) that were employed in ArgHRI and used for the experiments described here.



(a) arena (b) robot (c) treasure

Figure 1. The ArgHRI + HRTeam arena and robot.

Game domain

We have adapted the *Treasure Hunt Game (THG)* (Jones et al. 2006; Sklar and Azhar 2015) as an experimental domain for our research. The motivation is to emulate a controlled urban-search-and-rescue-like environment for studying human-robot interaction (Lewis et al. 2003; Marge et al. 2009).

Our version of the THG (Sklar and Azhar 2015) involves two players, a human and a robot, who function together as a team. Their task is framed as a real-time strategy game in which they must locate objects, or "treasures," in an arena that is accessible to the robot but not to the human. An example treasure is shown in Figure 1(c). This is a game because the team achieves a *score* based on their performance, time is a factor and the robot has limited resources. The robot cannot simply perform an exhaustive search of the arena to find all the treasures. Thus, the human and the robot have to decide how best to make use of those resources and locate as many treasures as possible in order to maximize their score.

The robot operates *inside* the arena with the ability to move around the arena, use sensors (e.g., cameras) to gather information about the arena and remotely communicate with the human player. The human operates *outside* the arena and has the ability to receive limited information about the arena from the robot and to communicate with the robot. Thus the type of human-robot interaction in our THG is categorized as a *remote interaction*, since the human and the robot are in different locations and not in each other's line of sight (Goodrich and Schultz 2007).

The robot has an energy level associated with it that decreases as the robot performs the following actions:

- When the robot moves, it expends energy and its health points decrease.
- When the robot gathers sensor data, it expends energy and its health points decrease.
- When the robot transmits sensor data to the human, it expends energy and its health points decrease.

The shared mission of the THG is for the human-robot team to find and correctly identify as many treasures as possible before the robot loses all of its health points. The human-robot team's score in the game is the number of points earned by correctly identifying treasures. They lose points by incorrectly identifying treasures.

A *game engine* runs independently of the ArgHRI system, to keep track of which treasures are placed where in the arena. The human-robot team can submit guesses (i.e., $\langle \mathsf{treasure}, \mathsf{location} \rangle$ tuples) to the game engine. The game engine returns a value (true or false) indicating if the guess is correct or not and updates the team's score.

User interface

The ArgHRI user interface, pictured in Figure 2, consists of the following panels:

(a) **Map Panel**: This panel displays a map of the arena, i.e., a 2D visualization of the physical arena. The interface module draws the robot's up-to-date location on the map.
(b) **Image Panel**: This panel displays the five most recent images received from the robot, taken in the room last visited by the robot.
(c) **Dialogue History Panel**: This panel displays the history of current and past dialogues between the robot and the human. This is like a "chat log" and contains participants' commitment stores ($R.CS$ and $H.CS$).
(d) **Dialogue Panel**: This panel provides the input facility, through which the human interacts with the robot. This is a constrained interface, to avoid having to work with the complexities of natural language (an obvious area for expansion in future work, but beyond the scope of the present research).
(e) **Game Status Panel**: This panel displays up-to-date game score information, the robot's "health points", and a tally of the treasures found during each game.

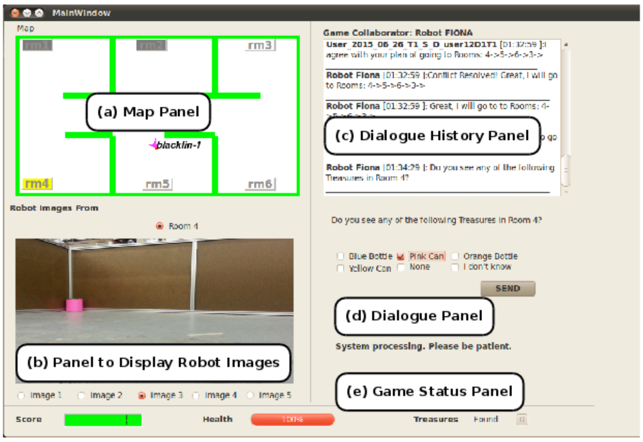


Figure 2. The ArgHRI user interface.

Experiments

The primary goal of this article is to present the results of a user study which we conducted in order to demonstrate the effectiveness of our *computational argumentation-based dialogue* approach to shared decision making in a human-robot team. The study investigates two factors. The first factor considers two different interaction modes: *human-as-collaborator* (where computational argumentation-based dialogue is employed to share decisions) versus *human-as-supervisor* (the control, where decisions are not shared). The second factor considers two different operating conditions: *physical* versus *simulated* robots.

The first factor—*human-as-collaborator vs human-as-supervisor*—was tested using a *within-subjects* design; each human subject interacted with two different robots, one implementing each interaction mode. Testing this factor was the primary purpose of our study: to demonstrate the effectiveness of our computational argumentation-based dialogue approach to shared decision making in a human-robot team, as compared to a "control" mode in which decisions were not shared.

The second factor—*physical vs simulated*—was tested using a *between-subjects* design; each human subject interacted either with two physical or two simulated robots. There were two reasons for assessing the impact of this factor. One was a practical consideration: human subjects could participate with robots operating in the simulated condition from several test sites, whereas in order to work with the physical robots, participants had to travel to a remote lab facility. The second purpose was because it is very common in robotics research to test theories with simulated robots. We were interested to ascertain whether significantly different outcomes would result from working with physical versus simulated robots.

The outcomes are reported and discussed in detail in the **Results** section. But first, in this section, we explain the experimental setup in detail.

Modes

As above, the experiments described here compare the performance of a human-robot team playing the THG in two different interaction modes:

- *human-as-collaborator* mode: the human and robot interact as collaborating peers, sharing decisions about what the robot should do, using computational argumentation-based dialogue, and reaching agreement before the robot takes any actions; and
- *human-as-supervisor* mode: the human and robot do not share decisions, and the human interacts with the robot in a supervisory capacity, providing commands to the robot which the robot obeys without question.

These two modes were implemented and presented to human subjects as two different robots. Robot "Fiona" enacted the *human-as-collaborator* mode, sharing decisions with the human using computational argumentation-based dialogues where the human and robot had to reach agreement about the robot's actions before the robot executed any actions. Robot "Mary" enacted the *human-as-supervisor* mode, obeying the human but not sharing decisions: the human dictated tasks to the robot and Robot Mary performed the tasks she was given. According to the hierarchy defined by (Parasuraman et al. 2000), Robot Fiona provides multi-faceted mid-level automation support, by offering "decision/action alternatives", helping to narrow down the selection and executing "suggestions if the human approves" (levels 2–5) (Parasuraman et al. 2000, p.287). In contrast, Robot Mary provides low-level automation of decision and action selection, wherein the human "must take all decisions" (level 1) (Parasuraman et al. 2000, p.287) (though the robot in our case takes all the actions). We note that Robot Mary is not tele-operated. Both Fiona and Mary are autonomous robots that perform their own low-level path-planning and motion decisions; the contrast is in the higher-level goal setting, where Fiona shares decisions about goals with her human teammate whilst Mary accepts goals as set by her human teammate without sharing decision-making.

Each human subject was exposed to both modes, following a *within-subjects* experiment design. To mitigate *learning* and *order* effects, half the human subjects played games in the collaborative mode first (with Robot Fiona); the other half played games in the supervisory mode first (with Robot Mary).

Conditions

As above, experiments were conducted using two different operating conditions. Under one condition, human subjects played games with *physical* robots operating in a physical arena. Under the other condition, human subjects played games with *simulated* robots operating in a virtual version of the same arena used for the physical experiments. The two different operating conditions were facilitated using the HRTeam framework (described earlier) which supports this dual functionality such that the robot controller software is the same, whether the robot is a physical entity (as pictured in Figure 1b) or simulated. Each human subject was exposed to one condition, following a *between-subjects* experiment design.

Decision points

In order to conduct a controlled experiment, we engineered a series of three decision points to test the shared decision-making capabilities of the human-robot teams. First, the team decides *where to go*—which rooms should be visited in order to look for treasure (since an exhaustive search is not possible given the robot's energy limitations, as engineered for the experiments). Second, the team decides *how to get there*—the order in which the agreed-upon rooms should be visited. Third, the team decides *what is found there*—whether the sensor data collected by the robot (e.g., images) contain treasures.

In the *human-as-collaborator* mode, for the first and second decision points, the human and the robot begin by each independently planning a travel sequence (the robot uses the A* path planner (Hart et al. 1968)). Then, they engage in computational argumentation-based dialogue to share the decision about which plan the robot should follow. This allows the teammates to identify any conflicts in their respective plans and reach agreement about how to resolve those conflicts and arrive at a mutually agreeable plan—which is the one that the robot actually executes. For the third decision point, the robot sends images to the human and they use computational-argumentation based dialogue to share decisions about whether any treasures are contained in the images. Again, mutual agreement must be reached about the content of the image before a guess is submitted to the game engine.

In the *human-as-supervisor* mode, for the first and second decision points, the human plans the robot's sequence of travel using a visual representation of the map and the robot's current location in the map (without any computational assistance from the system, e.g., numeric coordinates or distance estimates), based on the assumption that human participants are capable of spatial reasoning and utilizing common sense to determine feasible robot paths within a simple static environment. Note that once these two decisions are made (i.e., which rooms should be visited and the order in which they should be visited), the robot employs the A* path planner to compute its shortest path from one room to the next. For the third decision point, the robot sends images to the human, but the human determines the content of each image and decides alone whether to submit any guesses to the game engine.

The first two decision points comprise a *deliberation* phase, measuring the time it takes for the team (in either mode) to decide on a travel plan. These two decision points are reached only once per game, in order to conduct a controlled experiment. After the travel plan has been formulated, the robot starts moving. The time from when the robot begins moving until the game is over is referred to as the *execution* phase. The third decision point could be reached multiple times per game, as participants sought to identify multiple treasures during each game. Detailed analysis of the decision points and dialogues that were used in this study is beyond the scope of this article, but has been presented elsewhere (Azhar and Sklar 2016).

Metrics

There were several types of data collected during the user study, including objective performance metrics and subjective survey responses. The *objective performance metrics* described how well the team played the game. These include: (i) deliberation time; (ii) execution time; (iii) the length of the path traversed by the robot; and (iv) the score in the game. Two sets of *subjective survey responses* were collected during the experiment: one after playing a game with Robot Mary and one after playing a game with Robot Fiona. The second survey (after completing both games) included some additional questions about the user's overall experience. The subjective survey responses were presented using a Likert response format (Likert 1932) with seven alternatives, where 7 was the most favourable, 4 was neutral and 1 was the least favourable. Ten questions asked about the user's perception with respect to how much each robot helped the human to complete their task successfully, how easy it was to collaborate with each robot, how much the user trusted each robot, and how much the user was affected by the dialogue. Sample questions are listed in Table 4. The questions were presented to users in random order.

Thus our experiments involved two *independent variables*: interaction mode (human-as-collaborator (col) or human-as-supervisor (sup)) and operating condition (physical (phy) or simulated (sim) robots); and multiple *dependent variables*: four performance metrics and six survey responses (detailed below).

Research hypotheses

In our experiments, we expected to see statistically significant differences between the two interaction modes, but no significant differences between the two operating conditions.

We pose the following hypotheses about interaction mode with respect to the objective performance metrics:

H1. Deliberation time will be longer for human-as-collaborator than human-as-supervisor mode.
(H_0 : col = sup; H_A : col > sup)

H2. Execution time will be faster for human-as-collaborator than human-as-supervisor mode.
(H_0 : col = sup; H_A : col < sup)

H3. Distance travelled will be shorter for human-as-collaborator than human-as-supervisor mode.
(H_0 : col = sup; H_A : col < sup)

H4. Game score will be larger for human-as-collaborator than human-as-supervisor mode.
(H_0 : col = sup; H_A : col > sup)

The reasoning behind hypothesis H1 is that it will take more time for the human and robot to discuss options and reach agreement using our argumentation-based dialogue system than were the human simply to send commands to the robot. The reasoning behind the remaining hypotheses (H2–H4) is that the human and robot will combine abilities and reach agreement about the best plan, which will result in faster execution time, shorter distance and higher game score.

We pose the following hypotheses about interaction mode with respect to the subjective survey metrics:

H5. The user's perception of the *success of human-robot games* will be more positive for human-as-collaborator than human-as-supervisor mode.
(H_0 : col = sup; H_A : col > sup)

H6. The user's perception of the *ease of collaboration* in human-robot games will be more positive for human-as-collaborator than human-as-supervisor mode.
(H_0 : col = sup; H_A : col > sup)

H5	*I think that I can collaborate successfully with a robot in the treasure hunt game.*
H6	*I think that collaborating with a robot will make my task easier than working on the task alone.*
H7	*I think that a robot can be a trustworthy collaborator.*
H8	*I don't think that I have to expend a lot of effort to communicate with a robot.*
H9	*How well would you say you understood the task while interacting with Robot Mary/Fiona?*
H10	*How mentally demanding was the task while interacting with Robot Mary/Fiona?*

Table 4. Sample survey questions and corresponding research hypotheses.

H7. The user's perception of their *level of trust* in the robot for human-robot games will be higher for human-as-collaborator than human-as-supervisor mode.
(H_0 : col = sup; H_A : col > sup)

H8. The user's perception of the *effort to engage in dialogue* to support human-robot games will be higher for human-as-collaborator than human-as-supervisor mode.
(H_0 : col = sup; H_A : col > sup)

H9. The user's perception of their *understanding of the task* will be higher for human-as-collaborator than human-as-supervisor mode.
(H_0 : col = sup; H_A : col > sup)

H10. The user's perception of their *mental demand* will be higher for human-as-collaborator (col) than human-as-supervisor (sup) mode.
(H_0 : col = sup; H_A : col > sup)

In addition, the survey given after both games included some questions about the whole experiment. These questions attempted to obtain users' feedback on the complexity of the task they were asked to solve; their explicit preferences for one robot or the other, in the context of simple or complex tasks; and their perception of how much the human-as-collaborator robot helped them. The questions were as follows:

FS1. "Given a simple task, I prefer *Robot Mary/Fiona*."
FS2. "Given a complex task, I prefer *Robot Mary/Fiona*."
FS3. "How difficult was each scenario to understand?"
FS4. "How hard was it to make a decision (or come up with a plan to solve the game)?"
FS5. "How much did the Robot Fiona's feedback help you *resolve problems?*"
FS6. "How much did the Robot Fiona's feedback help your *decision making?*"

Some of these survey questions were inspired by the NASA-TLX measurement instrument (Hart and Staveland 1988; Hart 2006), particularly to assess participants' perceived levels of mental demand and task complexity.

Participants

Our study involved 60 participants. There were 27 participants (20 male and 7 female) who played games with physical robots. There were 33 participants (22 male and 11 female) who played games with simulated robots. Most of the 60 participants ranged in age from 18 to 24 (74%), while the remaining 26% of the participants were aged 25 to 39 years old. About 97% of the participants were undergraduate students. Overall, about half of the participants had no prior experience with robots. Those participants who interacted with a robot previously had less than one year of experience.

Each human subject received monetary compensation (US$10) for participating for an hour. Additional monetary compensation (US$10) was given for travel time and transportation fare for the human subjects who played games with physical robots in our laboratory (because they were required to travel some distance to get to the lab).

Results

This section presents the results from the experiments described above. We analysed the effects of the two experimental factors (interaction mode and operating condition), investigating whether any differences in outcomes were due to interaction mode (col vs sup) or operating condition (phy vs sim), or combinations thereof, as illustrated in Figure 3. Our analysis is presented in three sections: (1) *objective performance* metrics; (2) *subjective survey* data; and (3) a transcript of an example *dialogue sequence*, in order to demonstrate the type of interactions that resulted.

	operating condition:	
interaction mode:	physical (phy)	simulation (sim)
collaborator (col)	col-phy	col-sim
supervisor (sup)	sup-phy	sup-sim

Figure 3. Factor analysis (2×2) of experimental data.

Objective metrics

Before performing statistical tests on the objective performance metrics, we used the Shapiro-Wilk test (Shapiro and Wilk 1965) to assess the data for normality. The results are shown in Table 5, for both experimental factors taken together and separately. In most cases, there is a greater than 95% probability that the sample data we have collected is drawn from a normal distribution ($p < 0.05$). The only cases where the normality is really questionable is with respect to distance travelled (H3), particularly in the *human-as-supervisor* interaction mode. The mean and standard deviation for each objective metric are plotted in Figure 4. The left column plots the data grouped into 2×2 factors for analysis, as shown in Figure 3. The right column plots the data for each factor independently. In general, we are looking for statistically significant differences between the two interaction modes (*human-as-collaborator* (col) and *human-as-supervisor* (sup)), but not between the two operating modes (*physical* (phy) and *simulated* (sim)).

We conducted pairwise comparisons of the means of the four objective metrics using 2-way analysis of variance (ANOVA). In all cases, the critical value of $F_{crit} = 6.86$ was used, for $(1, 116)$ degrees of freedom and $\alpha = 0.01$,

deliberation time (H1)			
phy-col	phy-sup	sim-col	sim-sup
0.83	0.70	0.95 (0.155)	0.83
col	sup	phy	sim
0.91	0.80	0.84	0.93
execution time (H2)			
phy-col	phy-sup	sim-col	sim-sup
0.91	0.90	0.96 (0.208)	0.96 (0.257)
col	sup	phy	sim
0.96	0.90	0.95	0.92
distance travelled (H3)			
phy-col	phy-sup	sim-col	sim-sup
0.93 (0.058)	0.96 (0.464)	0.94 (0.063)	0.96 (0.217)
col	sup	phy	sim
0.93	0.99 (0.756)	0.98 (0.319)	0.92
game score (H4)			
phy-col	phy-sup	sim-col	sim-sup
0.86	0.93 (0.066)	0.92	0.95 (0.146)
col	sup	phy	sim
0.92	0.96 (0.073)	0.93	0.95

Table 5. W scores for Shapiro-Wilk tests (the probability that the sample data we have collected is drawn from a normal distribution). Cases where $p > 0.05$ are noted in parentheses.

	F (df)	
col vs sup	$55.33\ (1, 116)$	significant
phy vs sim	$1.10\ (1, 116)$	not significant
int vs ope	$1.73\ (1, 116)$	not significant

Table 6. ANOVA results comparing deliberation time (H1).

indicating that if the F values computed from the data are greater than F_{crit}, then the results are statistically significant (i.e., there is a greater than 99% probability that the results did not occur by chance).

Deliberation time (H1). Table 6 contains the ANOVA results for comparing the *deliberation time* objective metric. The hypothesis that col>sup in both operating conditions (phy and sim) holds, as shown in the first row of the table. This demonstrates that the *deliberation time* was significantly longer in the human-as-collaborator mode, where the human and robot discussed actions prior to executing them, as opposed to the human-as-supervisor mode, where no discussion took place. No statistically significant difference was detected between the two operating conditions, as shown in the second row of the table. No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table.

Execution time (H2). Table 7 contains the ANOVA results for comparing the *execution time* objective metric. The hypothesis that col<sup in both operating conditions (phy and sim) holds (as shown in the first row of the table). This says that the *execution time* was significantly shorter in the human-as-collaborator mode, where actions were discussed, as opposed to the human-as-supervisor mode, where the robot had no inputs into its action choices. Statistically significant differences were also found between the two operating conditions, as per the second row of the table. In both interaction modes, the overall execution time is shorter in the physical condition than the simulated condition. These results indicate that the timing of the motion model in the simulation condition is not well calibrated to the physical condition. No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table. Thus the difference between operating conditions has no impact on our research hypothesis about execution time as affected by interaction mode.

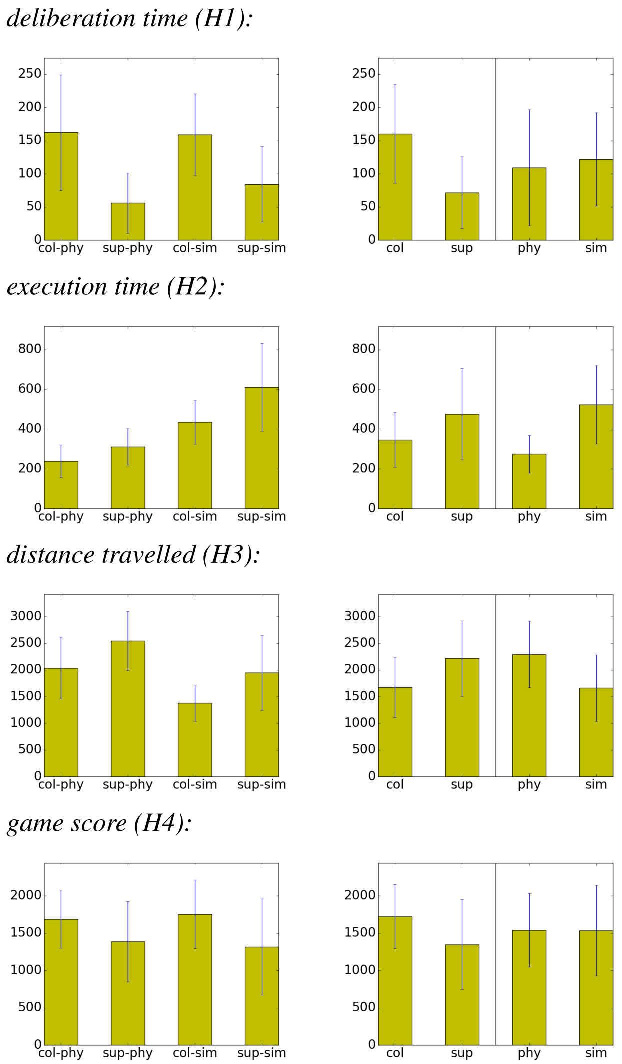


Figure 4. Mean and standard deviation of objective metrics.

	F (df)	
col vs sup	$24.25\ (1, 116)$	significant
phy vs sim	$87.90\ (1, 116)$	significant
int vs ope	$3.91\ (1, 116)$	not significant

Table 7. ANOVA results comparing execution time (H2).

Distance travelled (H3). Table 8 contains the ANOVA results for comparing the *distance travelled* objective

	F (df)	
col vs sup	$27.32\ (1, 116)$	significant
phy vs sim	$36.51\ (1, 116)$	significant
int vs ope	$0.08\ (1, 116)$	not significant

Table 8. ANOVA results comparing distance travelled (H3).

	F (df)	
col vs sup	$15.02\ (1, 116)$	significant
phy vs sim	$0.00\ (1, 116)$	not significant
int vs ope	$0.52\ (1, 116)$	not significant

Table 9. ANOVA results comparing game score (H4).

metric. The hypothesis that col<sup in both operating conditions (phy and sim) holds (as shown in the first row of the table). This says that the *distance travelled* was significantly shorter in the human-as-collaborator mode, where actions were discussed, as opposed to the human-as-supervisor mode, where the robot had no inputs into its action choices. Statistically significant differences were also found between the two operating conditions, as shown in the second row of the table. In both interaction modes, the overall distance travelled is longer in the physical condition than the simulated condition. This is because the robot's localisation in the physical condition is noisier than in the simulated condition (where the robot has perfect information about its position in the arena). Thus the simulated robot travels along a smoother trajectory because it does not need to correct its position, whereas the physical robot has to adjust its position and thus travels a more jagged trajectory. No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table. Thus the difference between operating conditions has no impact on our research hypothesis about distance travelled as affected by interaction mode.

Game score (H4). Table 9 contains the ANOVA results for comparing the *game score* objective metric. The hypothesis that col>sup in both operating conditions (phy and sim) holds (as shown in the first row of the table). This says that the *overall score* was significantly higher in the human-as-collaborator mode as opposed to the human-as-supervisor mode. No statistically significant difference was found between the two operating conditions (as per the second row of the table). No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table.

Summary of Objective metrics (H1–H4). In summary, all four objective metrics showed statistically significant results when comparing the human-as-collaborator (Robot Fiona) to the human-as-supervisor (Robot Mary) mode. Even if we choose to apply the conservative Bonferroni correction (Shaffer 1995), the differences are still significant for $\alpha = 0.001$[‡]. While the Fiona-human teams took longer to deliberate than the Mary-human teams, the Fiona-human teams completed games faster, travelled less distance and achieved higher game scores than the Mary-human teams.

In the cases of the deliberation time and game scores, there were no statistically significant differences between the games played with physical versus simulated robots. However, the execution times and distances travelled were different, due to poor calibration of the simulator and noise in localisation for the physical robots.

Subjective metrics

The subjective metrics are analysed in two parts. First, we consider the responses to survey questions that correspond to hypotheses H5-H10. Second, we consider the responses to the six questions on the final survey which asked participants for feedback about the whole experiment (FS1-FS6).

The first set of questions were presented to participants directly after playing a game with each robot, so we can compare how each user felt about the two robots they interacted with. Figure 5 illustrates the differences in responses between *human-as-collaborator* (col) and *human-as-supervisor* (sup), computed as col$-$sup and normalised over the number of responses. Positive values (> 0) indicate that a higher percentage of participants provided more favourable responses in human-as-collaborator mode than in human-as-supervisor mode. Detailed analysis of the differences corresponding to each hypothesis appears below.

In order to assess the statistical significance of these differences, we examine the distributions of raw responses, illustrated in Figure 6. The left column plots the data grouped into 2×2 factors for analysis, as shown in Figure 3. The right column plots the data for each factor independently. In general, we are looking for statistical significance in the differences between the two interaction modes (*human-as-collaborator* and *human-as-supervisor*). We expect the two operating conditions (*physical* and *simulated*) to have similar responses.

We conducted pairwise comparisons of the means of the four subjective metrics using 2-way analysis of variance (ANOVA). Although a Likert response format was used in the survey and the data collected is ordinal in nature, there is evidence in the literature defending the robustness of ANOVA for analysing Likert response format data (Glass et al. 1972; Carifio and Perla 2007). The critical value of $F_{crit} = 6.86$ was used for degrees of freedom $= (1, 116)$ and $\alpha = 0.01$, indicating that if the F values computed from the data are greater than F_{crit}, then the results are statistically significant (i.e., there is a greater than 99% probability that the results did not occur by chance). For some metrics, the degrees of freedom were higher than $(1, 120)$ in the second term, in which case $F_{crit} = 6.635$ was used for $\alpha = 0.01$.

Now we discuss each hypothesis in turn.

Success of human-robot games (H5). From Figure 5 we see that participants' perceived success of the games

[‡]For the subjective metrics, the Bonferroni correction would adjust the value of α from 0.01 to $0.01/4 = 0.0025, > 0.001$.

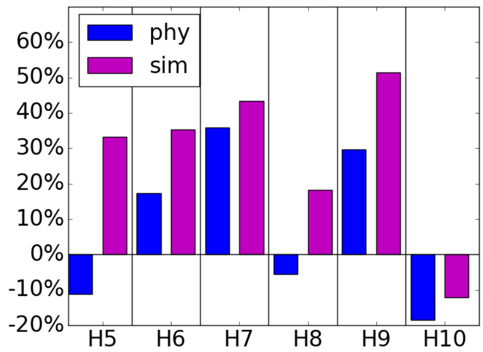


Figure 5. Differences in responses between *human-as-collaborator* (col) and *human-as-supervisor* (sup), computed as col−sup, normalised over the number of responses. Positive values (> 0) indicate that a higher percentage of participants provided more favourable responses in human-as-collaborator mode than in human-as-supervisor mode.

they played was slightly more favourable for human-as-supervisor mode under the physical operating condition, but more favourable for human-as-collaborator mode under the simulated operating condition. Table 10 contains the ANOVA results for determining statistical significance across the 2×2 experimental factors. Results supporting the hypothesis that col>sup in both operating conditions (phy and sim) are not statistically significant. The hypothesis that the two operating conditions (phy vs sim) are not different does hold. No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table.

	F (df)	
col vs sup	4.31 $(1, 236)$	not significant
phy vs sim	0.53 $(1, 236)$	not significant
int vs ope	6.03 $(1, 236)$	not significant

Table 10. ANOVA results comparing perceived success (H5).

Ease of collaboration (H6). From Figure 5 we see that participants' perceived ease of collaborating with the robot was more favourable for human-as-collaborator mode under both operating conditions. Table 11 contains the ANOVA results for determining statistical significance across the 2×2 experimental factors. The hypothesis that col>sup in both conditions (phy and sim) holds (as shown in the first row of the table). This demonstrates that the perceived *ease of collaboration* was significantly higher in the human-as-collaborator mode as opposed to the human-as-supervisor mode. In other words, participants felt that it was easier to collaborate with the robot operating in human-as-collaborator mode. The hypothesis that there is no significant difference between phy and sim also holds, as shown in the second row of the table. No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table.

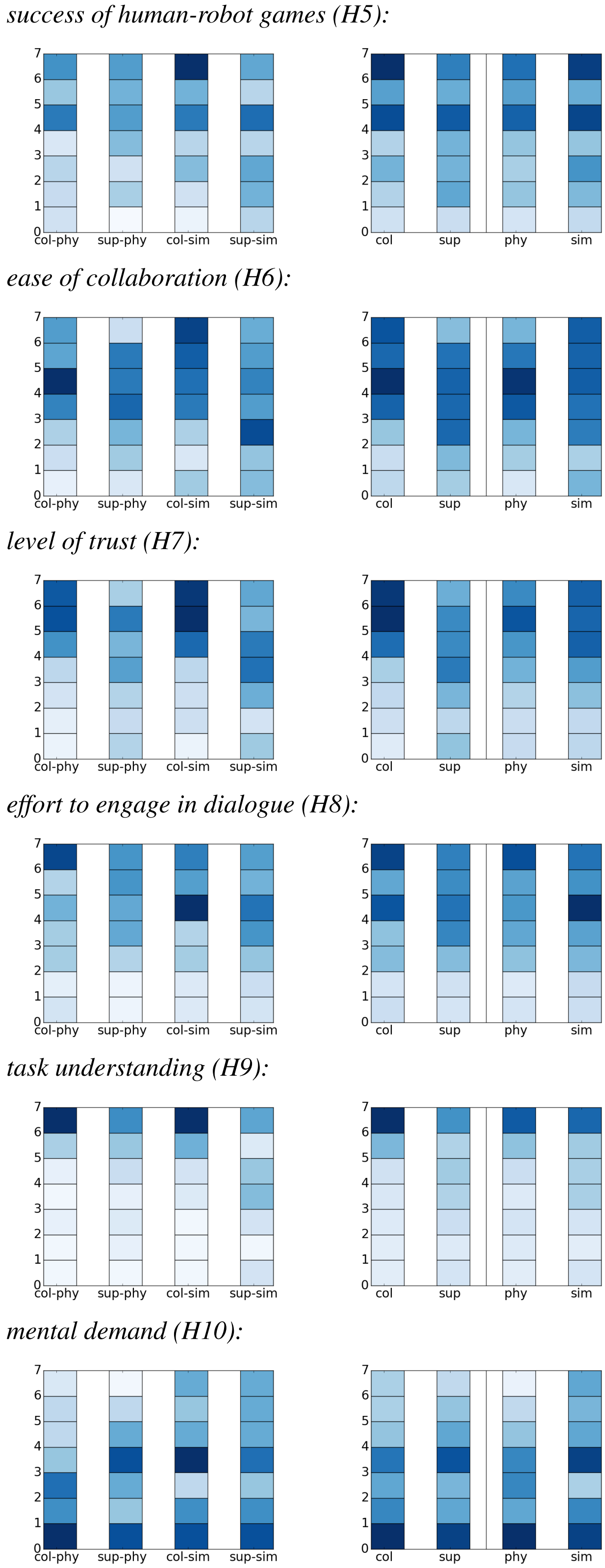


Figure 6. Heatmaps for subjective metrics. Darker cells indicate more popular responses. The subjective survey responses were presented using a Likert response format (Likert 1932) with seven alternatives, where 7 was the most favourable, 4 was neutral and 1 was the least favourable.

Level of trust (H7). From Figure 5 we see that participants' level of trust in the robot was more favourable for human-as-collaborator mode under both operating conditions.

	F (df)	
col vs sup	18.47 (1, 356)	significant
phy vs sim	0.70 (1, 356)	not significant
int vs ope	0.85 (1, 356)	not significant

Table 11. ANOVA results comparing ease of collaboration (H6).

Table 12 contains the ANOVA results for determining statistical significance across the 2×2 experimental factors. The hypothesis that col>sup in both conditions (phy and sim) holds, as shown in the first row of the table. This demonstrates that the *level of trust* was significantly higher in the human-as-collaborator mode as opposed to the human-as-supervisor mode. The hypothesis that there is no significant difference between phy and sim also holds, as shown in the second row of the table. No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table.

	F (df)	
col vs sup	52.50 (1, 356)	significant
phy vs sim	0.42 (1, 356)	not significant
int vs ope	0.07 (1, 356)	not significant

Table 12. ANOVA results comparing level of trust (H7).

Effort to engage in dialogue (H8). From Figure 5 we see that participants' perceived amount of effort required to engage in dialogue with the robot was more favourable for human-as-supervisor mode under the physical operating condition, but more favourable for the human-as-collaborator mode under the simulated operating condition. Table 13 contains the ANOVA results for determining statistical significance across the 2×2 experimental factors. The hypothesis that col>sup in both conditions (phy and sim) does not hold, as shown in the first row of the table. This demonstrates that the *level of effort to engage in dialogue* was not significantly different for the human-as-collaborator mode compared to the human-as-supervisor mode—participants felt that they did not have to expend more effort in order to engage in the dialogue afforded by the human-as-collaborator mode in comparison with the human-as-supervisor mode. This result is unexpected, because we thought that participants would feel that they were having to try harder in the human-as-collaborator mode (just as it can feel more difficult to work with a colleague who voices their opinion versus one who just does whatever you ask); but the results show the contrary. The hypothesis that there is no significant difference between phy and sim also holds, as shown in the second row of the table. No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table.

Task understanding (H9). From Figure 5 we see that participants' perceived understanding of the tasks undertaken in the games was more favourable for human-as-collaborator mode under both operating conditions.

	F (df)	
col vs sup	0.71 (1, 236)	not significant
phy vs sim	2.42 (1, 236)	not significant
int vs ope	1.56 (1, 236)	not significant

Table 13. ANOVA results comparing effort for dialogue (H8).

Table 14 contains the ANOVA results for determining statistical significance across the 2×2 experimental factors. The hypothesis that col>sup in both conditions (phy and sim) holds, as shown in the first row of the table. This demonstrates that the participants' perceived *understanding of the task* was significantly higher for the human-as-collaborator mode compared to the human-as-supervisor mode. The hypothesis that there is no significant difference between phy and sim also holds, as shown in the second row of the table. No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table.

	F (df)	
col vs sup	21.36 (1, 116)	significant
phy vs sim	5.59 (1, 116)	not significant
int vs ope	2.51 (1, 116)	not significant

Table 14. ANOVA results comparing task understanding (H9).

Mental demand (H10). From Figure 5 we see that participants' perceived mental demand was greater for human-as-supervisor mode under both operating conditions. Table 15 contains the ANOVA results for determining statistical significance across the 2×2 experimental factors. There are no statistically significant differences between the two interaction modes (col and sup) under either operating condition (phy or sim). This demonstrates that the participants' perceived *mental demand* was not more stressed when working with the robot in human-as-collaborator mode as compared to the human-as-supervisor mode; this result was also unexpected, as with the result for H8. No statistically significant effect was found between the two experimental factors, interaction mode (int) and operating condition (ope), as shown in the third row of the table.

	F (df)	
col vs sup	0.18 (1, 116)	not significant
phy vs sim	3.77 (1, 116)	not significant
int vs ope	0.14 (1, 116)	not significant

Table 15. ANOVA results comparing mental demand (H10).

Summary of Subjective metrics (H5–H10). In summary, the six objective metrics showed mixed results. Those that showed statistically significant differences between the human-as-collaborator versus human-as-supervisor mode were the metrics reflecting ease of collaboration, level of trust and task understanding: in both these cases, the human-as-collaborator interaction mode scored more favourably. As with the objective metrics, if we choose to apply the Bonferroni correction, the differences are still significant for $\alpha = 0.001$[§].

Other results were generally positive (perceived success, effort to engage in dialogue, task understanding and mental demand), but not statistically significantly different when analysed across the 2×2 factors.

Final survey questions. Finally, we turn to the results from the final survey questions (FS1-6), listed at the end of the previous section. Statistical analysis of these results are contained in Table 16. The table lists the *mode* (most popular) answer for each ordinal metric. Data for FS3-6 was reported using a 7-point Likert response format, where 7 was the most positive, 4 was neutral and 1 was the most negative. There were no statistically significant differences found between the physical (phy) and simulated (sim) operating conditions.

We make the following observations from the results:

- For both simple (FS1) and complex (FS2) tasks, under both operating conditions, participants overwhelmingly preferred the *human-as-collaborator* mode (col).
- Participants found the task environment *easy*, with a slight preference toward easier from the users who participated in the *physical* user study (FS3, 1 vs 3)
- Statistically significant differences between operating conditions were noted with respect to participants' responses about how hard it was to come up with a plan to solve the game: those working with physical robots reported that it was *somewhat easy* (3), whereas those working with simulated robots reported that it was *somewhat difficult* (5). But since both of these responses are just one mark either side of *neutral*, the differences are not strong.
- The majority of participants found the *human-as-collaborator* mode, *very helpful* (7) in resolving problems (FS5), in both operating conditions.
- Participants found the *human-as-collaborator* mode *helpful* in resolving problems (FS6), with a preference toward *very helpful* in the simulated operating condition (5 vs 7).

The responses to FS6 coincide with the responses to FS3 and FS4. Participants using the simulated operating condition found the task more difficult than the participants using the physical operating condition. We surmise that the participants using the physical operating condition, who saw the arena and the robots *in situ* before they played any games, had a better conception of the task environment. However, note that these participants played games in a remote location from the arena (in an adjacent room), so even though robots were moving around next door, these participants did not have line-of-sight to those robots and were only able to perceive and interact with them through the same interface (Figure 2) as the participants using the simulated condition.

Dialogue sequence

We conclude the presentation of results by sharing an example scene for a *persuasion dialogue* between one of the human participants (User12) from the physical operating condition and Robot Fiona at the "how to get there" decision point. Robot Fiona's belief state is as follows:

beliefs	*description*
$b \in R.\Sigma$	Robot Fiona believes that she should search $R4 \to R5 \to R6 \to R3$
$\neg b \in R.\Gamma(H)$	Robot Fiona believes that the human believes that the robot should $R3 \to R4 \to R6 \to R5$

Robot Fiona recognizes a *conflict*, and so it initiates a *persuasion dialogue* because the human and the robot have differing beliefs.

dialogue move	*scripted text in chat-style interface*
control layer	There is a conflict about Search Order
assert(R, H, b)	Robot Fiona: There is a conflict in our plans, we need to reach an agreement. I would like to go to Rooms: $4 \to 5 \to 6 \to 3 \to$. To get the reason for why I disagree with your plan, click "Why". If you want to continue with my plan, click "Agree". If you would like to continue with your plan, click "Disagree".

The user then interface provides the human collaborator with the following three choices from the possible dialogue moves for the persuasion dialogue:

dialogue move	*scripted text in chat-style interface*
control layer	Do you agree with the robot? ○ Agree ○ Disagree ○ I would like to know why?

Here according to the persuasion dialogue protocol, the human's choice would affect Robot Fiona's actions. If Robot Fiona's agenda is different from the human's agenda, then an explanation is given. Robot Fiona determines an efficient path plan in which to visit all

	mode		Fisher	
	phy	sim	*odds* (p)	
FS1	col	col	1.84 (0.397)	not significant
FS2	col	col	1.23 (1.000)	not significant
FS3	1	3	3.00 (0.195)	not significant
FS4	3	5	3.08 (0.042)	not significant
FS5	7	7	0.80 (1.000)	not significant
FS6	5	7	1.24 (1.000)	not significant

Table 16. Fisher exact tests to determine if there were statistically significant differences in the results obtained under the physical (phy) versus simulated (sim) operating condition for final survey questions (FS1-FS6). The mode (most popular) values are listed in the first two columns, followed by the Fisher score $(odds)$ and p values.

[§]For the objective metrics, the Bonferroni correction would adjust the value of α from 0.01 to $0.01/6 = 0.00167, > 0.001$.

rooms that it needs to visit using the A* algorithm (Hart et al. 1968). Robot Fiona's agenda is based on its current location and amount of battery power. The human then has the option of agreeing or disagreeing with the robot's agenda. If the human accepts the robot's belief, by selecting "Agree," then Robot Fiona would execute its plan. If the human rejects the robot's belief, by selecting "Disagree," then Robot Fiona would accept and execute the human's suggested plan¶.

In this example, the human challenges Robot Fiona's belief by first selecting "I would like to know why?," and the persuasion dialogue continues as follows:

dialogue move	*scripted text in chat-style interface*
challenge (H, R, b)	User12: I would like to know why you would like to visit Rooms: $4 \rightarrow 5 \rightarrow 6 \rightarrow 3$
assert $(R, H, S(b))$	Robot Fiona: According to my calculation, going to Rooms: $4 \rightarrow 5 \rightarrow 6 \rightarrow 3$ has a lower cost than your plan of going to Rooms: $3 \rightarrow 4 \rightarrow 6 \rightarrow 5$. Please "Agree" or "Disagree" with my plan.

Here, Robot Fiona provides supporting evidence for its belief in response to the human's challenge. The human accepts the robot's supporting evidence by selecting "Agree," and the persuasion dialogue continues as follows:

dialogue move	*scripted text in chat-style interface*
accept(H, R, b)	User12: I agree with your plan of going to Rooms: $4 \rightarrow 5 \rightarrow 6 \rightarrow 3$.
control layer	persuasion dialogue terminates

At the conclusion of the dialogue, Robot Fiona's beliefs and its beliefs about the human's beliefs are updated as follows:

beliefs	*description*
$b \in R.\Sigma$	Robot Fiona believes that she should search rooms $4 \rightarrow 5 \rightarrow 6 \rightarrow 3$.
$b \in R.\Gamma(H)$	Robot Fiona believes that she should search rooms $4 \rightarrow 5 \rightarrow 6 \rightarrow 3$

Robot Fiona successfully persuaded the human that its path was more efficient and would reduce the task completion time. The persuasion dialogue terminated in agreement. In contrast, Robot Mary (*human-as-supervisor* mode) would have agreed with the human's initial inefficient plan of searching room 4 first, then rooms $5, 6$ and 3.

Related Work

HRI environments can be divided into two general categories based on a time/space matrix (Dix et al. 2004), according to when and where a human and a robot work together: *proximate interaction* and *remote interaction* (Goodrich and Schultz 2007). Proximate interaction takes place when a human and robot are co-located in each other's line of sight. Remote interaction is when a human and a robot are in different locations and not in each other's line of sight. HRI developed for social applications are considered proximate interactions, since both human and robot are co-located and interact with each other face-to-face. Human-robot communication for proximate interaction needs to consider both non-verbal (e.g., gesture, gaze) and verbal (e.g., message content) aspects. In contrast, HRI during a search-and-rescue operation is considered a remote interaction, since the human and robot teammates are typically in different locations and out of sight of each other. Thus human-robot communication for remote interaction does not necessarily require support for non-verbal cues, but is primarily dependent on a rich exchange of information that can aid teamwork.

A successful human-robot team with a common goal needs to support interaction where humans and robots can complement each other's expertise and seek each other's help (Groom and Nass 2007). For example, a robot in a search-and-rescue scenario can seek a human's help with identifying human victims. Communication is an absolute requirement for successful human-robot collaboration and a very challenging problem (Goodfellow et al. 2010)(Hoffman and Breazeal 2004). As humans, we naturally use dialogue to communicate with other humans. Thus, dialogue has been viewed as a natural means of communication for humans to interact with robots. A human may interact with robots as a supervisor, operator, mechanic, peer or bystander (Scholtz 2003). In order for general deployment of multi-purpose robots that can collaborate with humans at work or home, a spoken dialogue interface will be necessary to provide valuable feedback to untrained and non-technical partners (Cakmak and Takayama 2014). Current issues in the human-robot dialogue domain could be divided into three major categories, which are: the "when to say it" problem, the "how to say it" problem, and the "what to say" problem. The "when to say it" problem deals with the *timing* of dialogue delivery (e.g., turn taking (Chao and Thomaz 2016; Thomaz and Chao 2011; Jonsdottir et al. 2008)). The "how to say it" problem addresses the best ways for a robot to *deliver* that content (e.g., using text, gestures, embodied cues (Mutlu 2011)(Simmons et al. 2011), speech or different modalities). The "what to say" problem addresses ways to determine the *concepts* that should be conveyed during dialogue.

One of the challenges of NL parsers for robotics is that they require a large amount of corpus data to train the system, which is lacking in the domain of human-robot collaboration (Scheutz et al. 2011).

Human-robot dialogue may also benefit from generating dialogue content based on a user's level of experience (e.g., novice, expert) (Torrey et al. 2006; Fong et al. 2001). Fischer (Fischer 2011) studied how robot dialogue could be designed to reduce uncertainty about the capabilities of the robot and the collaborative task addressed by human partners. However, the author used a Wizard-of-Oz study and didn't show how a dialogue framework can support

¶This decision was made in observance of Asimov's Three Laws of Robotics (Asimov 1950). A non-fiction reason for favoring the human in unresolved conflicts is that in the real-world, a human is held legally responsible for the actions made by a robot; so we make that explicit here.

such dialogue. Experimental robots have been deployed in a museum as a tour guide (Thrun et al. 2000), in the classroom as a tutor (Castellano et al. 2013; Krause et al. 2014), and in the office as a receptionist (Gockley et al. 2005). These examples primarily utilize limited (if any) choices about what concept(s) to discuss and scripted dialogue content for human-robot communication, thereby avoiding the "what to say" problem. However, such models do not scale for robots that operate in highly dynamic HRI environments (e.g., search-and-rescue).

During human-robot collaboration, a human and robot work together to make decisions about their joint actions. In this article, joint actions, or shared tasks, are those actions in which both the robot and the human communicate as a team to achieve a common goal (Hoffman and Breazeal 2004). As with two humans, human-robot team communication requires sharing information and taking initiative. The style of communication varies based on collaborative task, interaction, and environment. The communication requirements for human and robot collaboration involving dialogue differ in types of shared tasks for human-robot interaction (Fong et al. 2001). In (Yanco and Drury 2002), the authors defined a taxonomy of human-robot interaction. The categories were based on autonomy level/amount of intervention, ratio of people to robots, and level of shared interaction among human-robot teams. The autonomy level indicates the robots' level of autonomy, and the intervention level measures human intervention during a human-robot interaction. The authors suggest that the sum of robot autonomy and human intervention measurements should equal 100 percent. For instance, tele-operated robots have the least amount of autonomy (0 percent) and the greatest degree of human intervention (100 percent). On the other hand, museum tour-guide robots have full autonomy and require almost no human intervention (Nourbakhsh et al. 2005).

Humans may interact with robots as a supervisor, operator, mechanic, peer or bystander (Scholtz 2003). *Supervisory interaction* is the same as when one human supervises another human. The interaction is monitoring robots and evaluating their actions to achieve some goal(s). Here the robot software automatically generates actions. Supervisors, however, may step in to refine the robot's planning system, goals, and intentions to achieve any desired goals. *Operator interaction* allows an operator to choose robot-appropriate control mechanisms, behavior, or takes over full control to tele-operate the robot using the software. Scholtz pointed out the fact that the operator cannot change the goal or intention. Thus interaction support is needed for action, perception, and evaluation levels. *Mechanic interaction* refers to the role in which a human physically changes robot hardware (e.g., fixing a camera). It is similar to an operator role except for the hardware part. When changes have been made, software and hardware need to be observed to validate the robot's desired behavior that requires support for actions, perceptions, and evaluation. *Bystander interaction* refers to an implicit interaction with robots (e.g., interacting passively with Roomba, a home-cleaning robot, or museum tour guide robot). A robot might have some available controls for bystanders. The research on emotion and social interaction investigates how to make available robot capabilities evident to bystanders. The *Peer interaction* assumes that supervisors have control over only changing the goals and intentions. Then teammates can give commands to robots in order to achieve higher goals and intentions. For observations, we need support for the perception and evaluation levels. Human members interaction will not involve low-level robot behaviors (i.e., obstacle avoidance), but rather high-level behaviors (i.e., follow me). In case of emergency, a peer can take the role of operator or have the ability to hand off problems to a more qualified operator.

The Human-Robot Interaction Operating System (HRI/OS) (Fong et al. 2006), an interaction infrastructure based on a collaborative control model (Fong et al. 2001), was introduced to provide a framework for humans and robots to work together. The software framework supports human and robot engagement in a task-oriented dialogue about each others' abilities, goals, and achievements. HRI/OS was designed to support the performance of operational tasks, where tasks were well-defined and narrow in scope. In space exploration, operational tasks include: shelter and work hangar construction, habitat inspection, and in-situ resource collection and transport. HRI/OS is an agent-based system that incorporates embodied agents (humans and robots) and software agents employing a goal-oriented Open Agent Architecture for inter-agent communication and delegation (Cohen et al. 1994). The Open Agent Architecture (OAA) (Cohen et al. 1994) introduces the Inter-agent Communication Language (ICL) for interface, communication, and task coordination using a language shared by all agents (Cohen et al. 1994) regardless of platform and the low-level languages in which they are programmed. (Fong et al. 2006) have identified robots as capable of resolving issues, rather than immediately reporting task failure, through dialogue with humans in cases where robots lack skills or their resources have proved inadequate to the task. For example, a robot that has difficulty interpreting camera data might ask a human to lend visual processing ability to the task. This often allows tasks to be completed in spite of limitations of autonomy. (Fong et al. 2005) have investigated how *peer interaction* can help communication and collaboration, and the authors concluded that engaging in a dialogue where robots can ask task-oriented questions of humans through remote interaction such as teleoperation can be beneficial.

Research in both mixed-initiative dialogue and grounded dialogue are important to enhance peer-based collaboration. An early example of mixed-initiative dialogue for human-robot collaboration employs a Bayesian network (Hong et al. 2007) and demonstrates the benefits of human-robot engagement in a collaborative conversation to address ambiguity in natural language. Peltason and Wrede (Peltason and Wrede 2010) introduce *Pamini,* a pattern-based, mixed-Initiative human-robot interaction framework, to support flexible dialogue modelling that adopts task-state protocols with dialogue acts and interaction patterns. The research in grounded dialogue investigates how to generate dialogue in context and (Knepper et al. 2015) find evidence of benefits when a robot seeks help during failure. The aid comes from a human team member when the robot does not know how to solve a problem during human-robot collaboration.

Recent research in mixed human-robot teams (Gombolay et al. 2015) compares human interaction between human-robot collaboration and human-human collaboration to identify human-robot team efficiency and considers how to maximize the human desire to work with robot team members. One of Gombolay's experiments controlled the level of human decision-making authority for three different scenarios: (1) manual control—the human makes all the decisions (e.g., supervisory mode); (2) semi-autonomous—human and robot make decisions together (i.e., collaborative mode); and (3) autonomous—robot makes all the decisions. The results from the experiments support our finding that semi-autonomous (e.g., collaborative) decision-making can benefit the human-robot team during collaboration.

Butchibabu (Butchibabu 2016) proposes a *Maximum Entropy Markov model (MEMM)* to support dialogues during planning and task allocation and demonstrates that goal-based information sharing requires more context-based communication (e.g., can you help me find the red ball?) than reactive-implicit information (e.g., no red ball found). Our work presented here, however, does not address task planning and task allocation, but solely investigates the impact of shared-decision making that employs computational argumentation.

The domain of human-robot collaboration also lacks comprehensive experiments that investigate users' perceptions of a robot operating as a peer, a collaborator that can argue, but agree. For HRI systems to be truly collaborative, participants must be able to exchange ideas and engage in opportunistic dialogue that can adjust dynamically as the situation unfolds, including spontaneously changing which peer is leading the dialogue. For example, upon experiencing or anticipating failure or discovering new opportunities, the robot needs to be able to take the *initiative* (Carbonell 1971; Horvitz 1999) in an ongoing or new conversation. Our approach applies *argumentation-based dialogue* to enable two-way feedback for human-robot peer collaboration.

Conclusion

The research presented here does not claim that computational argumentation-based dialogue is the only methodology appropriate for enabling shared decision making in human-robot teams. Rather, this work suggests that computational argumentation-based dialogue is a strong candidate to facilitate shared decision making and that it can aid people in human-robot settings. This research also does not claim that the computational argumentation-based dialogue presented here is the only possible implementation of a computational argumentation-based theoretical framework. The system described here applies a formal logic-based argumentation dialogue framework to the practical domain of human-robot collaboration, a contribution to both the fields of human-robot interaction and computational argumentation.

In this research, we accept that there are many applications in HRI that do not require peer interaction. In many cases, supervisory interaction will be sufficient. Collaborative interaction will be required of an HRI system that needs robot autonomy with guidance from humans, in situations such as urban search-and-rescue, humanitarian demining and nuclear power plant inspection. Such interaction can also be useful in domestic and healthcare domains, where robots of the future will be caregivers in private homes and residential care facilities. Human-robot collaborators will benefit from systems that have the ability to seek information from each other in order to minimize uncertainty, expand individual and shared knowledge, and challenge or persuade each other.

Our research suggests several opportunities for application of computational argumentation-based dialogue for collaborative interaction in the a variety of human-robot scenarios:

- *Complex tasks*: When a robot and human collaborator do not know how to address all the issues in a complex task, both parties can engage in an **inquiry dialogue** to explore problems surrounding the complex task.
- *Computationally expensive decisions*: A robot is better equipped to perform computationally expensive tasks faster than a human collaborator. The human collaborator can gather information from the robot employing an **information-seeking dialogue**.
- *Lack of knowledge*: Humans may lack knowledge about a robot's capabilities or information about its physical environment. Similarly, the robot may lack knowledge that humans may have. The human collaborator can employ an **information-seeking dialogue** to inquire about information from the robot collaborator and vice-versa.
- *Dynamic environments*: Humans can seek information using an information-seeking dialogue or use an inquiry dialogue from the robot when dynamic changes occur in the environment (e.g., sudden appearance of an unexpected obstacle). Dialogues can be embedded inside an **inquiry dialogue** to explore changes in a task due to the dynamic nature of an operating environment.
- *Multiple unknowns*: A task can be considered complex where there are multiple unknowns. In this case, a human and robot may need to engage in **information-seeking** and **inquiry dialogues**.
- *Conflict resolution*: When the human user and the robot hold opposing beliefs, thereby causing a conflict, there is an opportunity for **persuasion dialogue**. For example, when a human and a robot have different agendas (see the dialogue sequence at the end of the previous section), the robot will attempt to convince the human of the efficacy of its agenda by providing an "effective agenda" justification.

We have described a user study that evaluates the impact of computational argumentation-based dialogue as a means to support shared decision making for human-robot teams performing in a Treasure Hunt Game scenario. Our underlying computational argumentation-based dialogue framework, ArgHRI, was outlined. The user study is discussed in detail. Overall, 60 human subjects participated and two interaction modes were compared: humans *collaborating* with a robot that employed argumentation-based dialogue to facilitate interaction ("Robot Fiona"), and humans *supervising* another robot that did not use dialogue ("Robot Mary"). Results showed that, with respect to all of the objective performance metrics collected and some of the subjective survey results, the collaborative mode (with Robot Fiona) was statistically significantly better than the supervisory mode (with Robot Mary). The subjective metrics which showed statistically significant differences reflected ease of collaboration, level of trust and task understanding, where the human-as-collaborator mode was preferred.

We can conclude that even for a simple task like the one presented here, an HRI system capable of supporting collaboration through argumentation-based dialogue can be beneficial to system performance and user experience. This result comes despite expectations that supervisory interaction would be preferred for simple tasks in which human collaborators do not require much help. The work presented here demonstrates that human-robot collaboration, enabled through argumentation-based dialogue, can support expansion of individual or shared knowledge. It can aid in the resolution of disagreements and thus prevent human or robot errors. It may reduce task completion time and increase success for a collaborative task.

Future work will involve extending this research in several directions. There is a need for human-robot dialogue support for more complex task domains in which there are common goals. There is a need for human-robot dialogue support for collaborative interaction between humans and multi-robot teams that have different goals. There is a need for bridging the computational argumentation-based dialogue framework presented here to natural language research and exploring natural language implementations of argumentation-based dialogue.

Human-robot dialogue that can aid shared decision making, supports the expansion of individual or shared knowledge, and resolves disagreements between collaborative human-robot teams will be much sought after as human society transitions from a world of robot-as-a-tool to robot-as-a partner. The work presented here demonstrates a version of collaborative interaction enabled through argumentation-based dialogue, allowing humans and robots to work together as partners.

Acknowledgments

The authors are grateful to the anonymous reviewers for their insightful comments that have helped improve the presentation of this work. The authors would also like to thank Dr Isabel Sassoon for her advice on aspects of the statistical analysis.

This work was partially funded by the US National Science Foundation (NSF) under grants #IIS-1116843, #IIS-1338884 and #CNS-1117761, by the US Army Research Office under the Science of Security Lablet grant (SoSL), by the US Army Research Laboratory under the Network Science Collaborative Technology Agreement, by a University of Liverpool (UK) Research Fellowship, and by a US-UK Fulbright-King's College London Scholar Award. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders.

References

Asimov I (1950) *I, Robot*. New York, NY, USA: Bantam Dell.

Azhar MQ (2015) *Toward an argumentation-based dialogue framework for human-robot collaboration*. PhD Thesis, City University of New York, New York, NY, USA.

Azhar MQ and Sklar EI (2016) Analysis of empirical results on argumentation-based dialogue to support shared decision making in a human-robot team. In: *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. New York, NY, USA: IEEE, pp. 861–866.

Belesiotis A, Rovatsos M and Rahwan I (2010) Agreeing on plans through iterated disputes. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, volume 1. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, pp. 765–772.

Bench-Capon TJM and Dunne PE (2007) Argumentation in artificial intelligence. *Artificial intelligence* 171(10–15): 619–641.

Black E (2007) *A Generative Framework for Argumentation-Based Inquiry Dialogues*. PhD Thesis, University College London, London, UK.

Black E and Atkinson K (2011) Choosing Persuasive Arguments for Action. In: *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, volume 3. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, pp. 905–912.

Black E and Hunter A (2009) An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems* 19(2): 173–209.

Butchibabu A (2016) *Anticipatory communication strategies for human robot team coordination*. PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Cakmak M and Takayama L (2014) Teaching people how to teach robots: The effect of instructional materials and dialog design. In: *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction (HRI)*. Bielefeld, Germany: New York, NY, USA: ACM Press, pp. 431–438.

Caminada MWA (2009) Truth, Lies and Bullshit; distinguishing classes of dishonesty. In: *Proceedings of the Workshop on Social Simulation at the International Joint Conference on Artificial Intelligence (IJCAI)*. Pasadena, CA, USA, pp. 39–50.

Carbonell JR (1971) Mixed-Initiative Man-Computer Instructional Dialogues. Technical report, Bolt Beranek and Newman Incorporated (BBN), Cambridge, MA, USA.

Carifio J and Perla RJ (2007) Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences* 3(3): 106–116.

Castellano G, Paiva A, Kappas A, Aylett R, Hastie H, Barendregt W, Nabais F and Bull S (2013) Towards Empathic Virtual and Robotic Tutors. In: *Artificial Intelligence in Education, Lecture Notes in Computer Science Volume 7926*. Berlin: Springer, pp. 733–736.

Chao C and Thomaz A (2016) Timed petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration. *International Journal of Robotics Research* 35(11): 1330–1353.

Cogan E, Parsons S and McBurney P (2005) What Kind of Argument Are We Going to Have Today? In: *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Ütrecht, The Netherlands: New York, NY, USA: ACM Press, pp. 544–551.

Cohen PR, Cheyer A, Wang M and Baeg SC (1994) An open agent architecture. In: *AAAI Spring Symposium on Software Agents*. Stanford University, CA, USA: Palo Alto, CA, USA: AAAI Press, pp. 197–204.

Dix A, Finlay J, Abowd G and Beale R (2004) *Human-Computer Interaction*. 3rd edition. Harlow, England: Pearson Education.

Fischer K (2011) How People Talk with Robots: Designing Dialogue to Reduce User Uncertainty. *AI Magazine* 32(4): 31–38.

Fong T, Nourbakhsh I, Kunz C, Flückiger L, Schreiner J, Ambrose R, Burridge R, Simmons R, Hiatt LM and Schultz A (2005) The peer-to-peer human-robot interaction project. In: *AIAA SPACE*. Long Beach, CA, USA: American Institute of Aeronautics and Astronomics.

Fong TW, Kunz C, Hiatt L and Bugajska M (2006) The human-robot interaction operating system. In: *Proceedings of the 1st ACM SIGCHI/SIGART conference Human-Robot Interaction (HRI)*. Salt Lake City, UT, USA: New York, NY, USA: ACM Press, pp. 41–48.

Fong TW, Thorpe C and Baur C (2001) Collaboration, dialogue, and human-robot interaction. In: *Proceedings of the 10th International Symposium of Robotics Research (ISRR)*, volume Springer Tracts in Advanced Robotics (STAR) 6. Lorne, Victoria, Australia: Berlin: Springer-Verlag, pp. 255–268.

Gerkey B, Vaughan RT and Howard A (2003) The Player/Stage Project: Tools for Multi-Robot and Distributed Sensor Systems. In: *Proceedings of the International Conference on Advanced Robotics (ICAR)*, volume 1. University of Coimbra, Portugal: IEEE, pp. 317–323.

Girle R (1996) Commands in Dialogue Logic. In: Gabbay DM and Ohlbach HJ (eds.) *Practical Reasoning: Proceedings of the First International Conference on Formal and Applied Practical Reasoning (FAPR), Lecture Notes in Artificial Intelligence*, volume 1085. Berlin: Springer, pp. 246–260.

Glass GV, Peckham PD and Sanders JR (1972) Consequences of failure to meet assumptions underlying the analyses of variance and covariance. *Review of Educational Research* 42: 237–288.

Gockley R, Bruce A, Forlizzi J, Michalowski M, Mundell A, Rosenthal S, Sellner B, Simmons R, Snipes K, Schultz A and Wang J (2005) Designing robots for long-term social interaction. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Edmonton, Canada: IEEE, pp. 1338–1343.

Gombolay MC, Gutierrez RA, Clarke SG, Sturla GF and Shah JA (2015) Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots* 39(3): 293–312.

Goodfellow IJ, Koenig N, Muja M, Pantofaru C, Sorokin A and Takayama L (2010) Help me help you: Interfaces for personal robots. In: *Proceedings of the 5th ACM/IEEE international conference on Human-Robot Interaction (HRI)*. Osaka, Japan: New York, NY, USA: ACM Press, pp. 187–188.

Goodrich MA and Schultz AC (2007) Human-robot interaction: a survey. *Foundations and Trends in Human-Computer Interaction* 1(3): 203–275.

Groom V and Nass C (2007) Can robots be teammates? Benchmarks in human-robot teams. *Interaction Studies* 8(3): 483–500.

Hart PE, Nilsson NJ and Raphael B (1968) A formal basis for the heuristic determination of minimal cost paths. *IEEE Transactions on Systems Science and Cybernetics* 4(2): 100–107.

Hart SG (2006) NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50(9): 904–908.

Hart SG and Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52: 139–183.

Hoffman G and Breazeal C (2004) Collaboration in human-robot teams. In: *Proceedings of the AIAA 1st Intelligent Systems Technical Conference*. Chicago, IL, USA. DOI: 10.2514/6.2004-6434.

Hong JH, Song YS and Cho SB (2007) Mixed-initiative human-robot interaction using hierarchical bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37(6): 1158–1164.

Horvitz E (1999) Principles of mixed-initiative user interfaces. In: *Proceedings of the ACM CHI 99 Human Factors in Computing Systems Conference*. Pittsburgh, PA, USA: New York, NY, USA: ACM Press, pp. 159–166.

Hulstijn J (2000) *Dialogue Models for Inquiry and Transaction*. PhD Thesis, University of Twente, Enschede, Netherlands.

Jones EG, Browning B, Dias MB, Argall B, Veloso M and Stentz A (2006) Dynamically formed heterogeneous robot teams performing tightly-coordinated tasks. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 570–575.

Jonsdottir GR, Thorisson KR and Nivel E (2008) Learning smooth, human-like turntaking in realtime dialogue. In: *Intelligent Virtual Agents*. Berlin: Springer, pp. 162–175.

Knepper RA, Tellex S, Li A, Roy N and Rus D (2015) Recovering from failure by asking for help. *Autonomous Robots* 39(3): 347–362.

Krause E, Zillich M, Williams T and Scheutz M (2014) Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In: *Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence*. Québec City, Québec, Canada: Palo Alto, CA, USA: AAAI Press, pp. 2796–2802.

Lewis M, Sycara K and Nourbakhsh I (2003) Developing a testbed for studying human-robot interaction in urban search and rescue. In: *Proceedings of the 10th International Conference on Human Computer Interaction*. Crete, Greece: Berlin: Springer, pp. 22–27.

Likert R (1932) A Technique for the Measurement of Attitudes. *Archives of Psychology* 140: 44–55.

Marge MR, Pappu AK, Frisch B, Harris TK and Rudnicky AI (2009) Exploring Spoken Dialog Interaction in Human-Robot Teams. In: *Proceedings of Robots, Games, and Research: Success stories in USARSim IROS Workshop*. St. Louis, MO, USA: New York, NY, USA: ACM Press, pp. 126–133.

McBurney P and Parsons S (2001a) Chance discovery using dialectical argumentation. *New Frontiers in Artificial Intelligence* : 414–424.

McBurney P and Parsons S (2001b) Representing epistemic uncertainty by means of dialectical argumentation. *Annals of Mathematics and Artificial Intelligence* 32(1–4): 125–169.

McBurney P and Parsons S (2002) Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language, and Information* 11(3): 315–334.

McBurney P and Parsons S (2003) Dialogue Game Protocols. In: Huget MP (ed.) *Communication in Multiagent Systems*, *Lecture Notes in Computer Science*, volume 2650. Berlin: Springer, pp. 269–283.

McBurney P and Parsons S (2004) A Denotational Semantics for Deliberation Dialogues. In: *Proceedings of the 3rd International Joint Conference of Autonomous Agents and Multiagent Systems (AAMAS)*. pp. 86–93. DOI:10.1109/AAMAS.2004.11.

McBurney P and Parsons S (2009) Dialogue games for agent argumentation. In: *Argumentation in Artificial Intelligence*. New York: Springer US, pp. 261–280.

Medellin-Gasque R (2013) *Argumentation-based Dialogues over Cooperative Plans*. PhD Thesis, University of Liverpool, Liverpool, UK.

Medellin-Gasque R, Atkinson K, McBurney P and Bench-Capon T (2012) Arguments over co-operative plans. In: *Theory and Applications of Formal Argumentation*. Berlin: Springer, pp. 50–66.

Mutlu B (2011) Designing Embodied Cues for Dialog with Robots. *AI Magazine* 32(4): 17–30.

Nielsen SH and Parsons S (2006) A generalization of Dung's Abstract Framework for Argumentation: Arguing with Sets of Attacking Arguments. In: *Proceedings of the Workshop on Argumentation in Multiagent Systems (ArgMAS)*.

Nourbakhsh I, Hamner E, Dunlavey B, Bernstein D and Crowley K (2005) Educational results of the personal exploration rover museum exhibit. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 4278–4283.

Parasuraman R, Sheridan TB and Wickens CD (2000) A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man and Cybernetics–Part A: Systems and Humans* 30(3): 286–297.

Parsons S and McBurney P (2003) Argumentation-based dialogues for agent coordination. *Group Decision and Negotiation* 12(5): 415–439.

Parsons S, Wooldridge M and Amgoud L (2003) Properties and complexity of formal inter-agent dialogues. *Journal of Logic and Computation* 13(3): 347–376.

Peltason J and Wrede B (2010) Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns. In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. University of Tokyo, Japan: Association for Computational Linguistics, pp. 229–232.

Prakken H (2006) Formal systems for persuasion dialogue. *Knowledge Engineering Review* 21(2): 163–188.

Rahwan I, Ramchurn SD, Jennings NR, McBurney P, Parsons S and Sonenberg L (2003) Argumentation-based negotiation. *The Knowledge Engineering Review* 18(4): 343–375.

Rahwan I and Simari GR (eds.) (2009) *Argumentation in Artificial Intelligence*. Berlin: Springer-Verlag.

Rosenthal S, Biswas J and Veloso M (2010) An Effective Personal Mobile Robot Agent Through Symbiotic Human-Robot Interaction. In: *Proceedings of the Ninth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, volume 1. Toronto, Canada: International Foundation for Autonomous Agents and Multiagent Systems, pp. 915–922.

Scheutz M, Cantrell R and Schermerhorn P (2011) Toward Humanlike Task-Based Dialogue Processing for Human Robot Interaction. *AI Magazine* 32(4): 77–84.

Scholtz J (2003) Theory and Evaluation of Human Robot Interactions. In: *Hawaii International Conference on System Science (HICSS)*, volume 36. Big Island, HI, USA: IEEE. DOI:10.1109/HICSS.2003.1174284.

Shaffer JP (1995) Multiple Hypothesis Testing. *Annual Review of Psychology* 46: 561–584.

Shapiro SS and Wilk WB (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52(3/4): 591–611.

Simmons R, Makatchev M, Kirby R, Lee MK, Fanaswala I, Browning B, Forlizzi J and Sakr M (2011) Believable Robot Characters. *AI Magazine* 32(4): 39–52.

Sklar E and Parsons S (2004) Towards the Application of Argumentation-based Dialogues for Education. In: *Proceedings of the 3rd International Conference of Autonomous Agents and Multi Agent Systems (AAMAS)*. New York, NY, USA: New York, NY, USA: ACM Press, pp. 1420–1421.

Sklar EI and Azhar MQ (2015) Argumentation-based dialogue games for shared control in human-robot systems. *Journal of Human-Robot Interaction* 4(3): 120–148.

Sklar EI, Azhar MQ, Parsons S and Flyr T (2013a) Enabling Human-Robot Collaboration via Argumentation (Extended Abstract). In: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. St. Paul, MN, USA: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1251–1252.

Sklar EI, Özgelen AT, Muñoz JP, Gonzalez J, Manashirov M, Epstein SL and Parsons S (2011) Designing the HRTeam Framework: Lessons Learned from a Rough-and-Ready

Human/Multi-Robot Team. In: *Workshop on Autonomous Robots and Multirobot Systems (ARMS) at Autonomous Agents and MultiAgent Systems (AAMAS)*. pp. 232–251.

Sklar EI, Parsons S, Li Z, Salvit J, Perumal S, Wall H and Mangels J (2015) Evaluation of a trust-modulated argumentation-based interactive decision-making tool. *Autonomous Agents and Multi-Agent Systems* 30: 1–38.

Sklar EI, Parsons S, Özgelen AT, Schneider E, Costantino M and Epstein SL (2013b) Hrteam: A framework to support research on human/multi-robot interaction. In: *Proceedings of the International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*. St Paul, MN, USA: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1409–1410.

Tang Y, Cai K, Sklar E, McBurney P and Parsons S (2010a) A system of argumentation for reasoning about trust. In: *Proceedings of the 8th European Workshop on Multi-Agent Systems*. Paris, France.

Tang Y, Cai K, Sklar E, McBurney P and Parsons S (2012a) Using argumentation to reason about trust and belief. *Journal of Logic and Computation* 22(5): 979–1018.

Tang Y, Cai K, Sklar E and Parsons S (2011) A prototype system for argumentation-based reasoning about trust. In: *Proceedings of the 9th European Workshop on Multiagent Systems*. Maastricht, The Netherlands.

Tang Y, Norman TJ and Parsons S (2010b) A model for integrating dialogue and the execution of joint plans. In: *Argumentation in multi-agent systems*. Berlin: Springer, pp. 60–78.

Tang Y and Parsons S (2005) Argumentation-based dialogues for deliberation. In: *Proceedings of the Fourth International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. Ütrecht, The Netherlands: New York, NY, USA: ACM Press, pp. 552–559.

Tang Y, Sklar EI and Parsons S (2012b) An argumentation engine: ArgTrust. In: *Ninth International Workshop on Argumentation in MultiAgent Systems (ArgMAS)*. Valencia, Spain.

Thomaz A and Chao C (2011) Turn-Taking Based on Information Flow for Fluent Human-Robot Interaction. *AI Magazine* 32(4): 53–63.

Thrun S, Beetz M, Bennewitz M, Burgard W, Cremers AB, Dellaert F, Fox D, Hähnel D, Rosenberg C, Roy N, Schulte J and Schulz D (2000) Probabilistic algorithms and the interactive museum tour-guide robot minerva. *International Journal of Robotics Research* 19(11): 972–999.

Toniolo A, Li H, Norman TJ, Oren N, Ouyang RW, Srivastava M, Dropps T, Allen JA and Sullivan P (2015a) Enabling intelligence analysis through agent-support: The cispaces toolkit. In: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Istanbul, Turkey: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1907–1908.

Toniolo A, Norman TJ, Oren N, Etuk A, Dropps T, Allen JA, Cerutti F, Ouyang RW, Srivastava MB and Sullivan P (2015b) Supporting reasoning with different types of evidence in intelligence analysis. In: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Istanbul, Turkey: International Foundation for Autonomous Agents and Multiagent Systems, pp. 781–789.

Torrey C, Powers A, Marge M, Fussell SR and Kiesler S (2006) Effects of adaptive robot dialogue on information exchange and social relations. In: *Proceedings of the 1st ACM Conference on Human-Robot Interaction (HRI)*. Salt Lake City, UT, USA: New York, NY, USA: ACM Press, pp. 126–133.

Walton D and Krabbe ECW (1995) *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. Albany, NY, USA: State University of New York Press.

Walton D, Reed C and Macagno F (2008) *Argumentation Schemes*. Cambridge, UK: Cambridge University Press.

Wobcke W, Ho V, Nguyen A and Krzywicki A (2006) A bdi agent architecture for dialogue modelling and coordination in a smart personal assistant. In: *Proceedings of the 2005 NICTA-HCSNet Multimodal User Interaction Workshop - Volume 57*, MMUI '05. Sydney, Australia: Darlinghurst: Australian Computer Society, Inc. ISBN 1-920-68239-2, pp. 61–66.

Yanco H and Drury J (2002) A taxonomy for human-robot interaction. In: *Proceedings of the AAAI Fall Symposium on Human-Robot Interaction, AAAI Technical Report*, volume FS-02-03. Falmouth, MA, USA: Palo Alto, CA, USA: AAAI Press.