Explanation through Argumentation

Elizabeth I. Sklar
King's College London
London, UK
elizabeth.sklar@kcl.ac.uk

Mohammad Q. Azhar
Borough of Manhattan Community College,
City University of New York
New York, NY, USA
mazhar@bmcc.cuny.edu

ABSTRACT

Computational Argumentation is a logical model of reasoning that has its origins in philosophy and provides a means for organising evidence for (or against) particular claims (or decisions). Argumentation-based Dialogue is a related methodology that is used for structuring interactions between two (or more) agents and has been explored within the Multi-Agent Systems community as an extended form of negotiation where agents can not only exchange claims, but also their reasons for believing (or disbelieving) those claims. Recently, the Artificial Intelligence (AI) community has become intrigued by the notion of “Explainable AI”, in which intelligent systems are able to explain predictions or decisions to (human) users. There is a natural pairing between Explainable AI and Argumentation: the first requires the need to clarify and defend decisions and the second provides a method for linking any decision to the evidence supporting it. In this paper, we describe how the two are connected and illustrate the utility of argumentation-based dialogue as a technique for implementing Explainable AI in a human-robot system.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; • **Human-centered computing** → **User studies**;

KEYWORDS

Computational Argumentation, Explainable AI, Human-Robot Interaction

ACM Reference Format:

Elizabeth I. Sklar and Mohammad Q. Azhar. 2018. Explanation through Argumentation. In *6th Int'l Conference on Human-Agent Interaction (HAI'18)*, Dec. 15–18, 2018, Southampton, United Kingdom. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/10.1145/3284432.3284473>

1 INTRODUCTION

Recently, the Artificial Intelligence (AI) community has become intrigued by the notion of *Explainable AI*, in which intelligent systems are able to clarify predictions or decisions to (human) users.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HAI '18, December 15–18, 2018, Southampton, United Kingdom

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-5953-5/18/12...\$15.00

<https://doi.org/10.1145/10.1145/3284432.3284473>

The ideas explored in this paper were inspired by the example in [10], where the following questions are hypothesised as part of an imaginary interaction between a human and an AI system while jointly planning a mission:

- Q1. *Why did you do that?*
- Q2. *And why didn't you do something else (that I would have done)?*
- Q3. *Why is what you propose to do more efficient/safe/cheap than something else (that I would have done)?*
- Q4. *Why can't you do that?*
- Q5. *Why do I need to replan at this point?*
- Q6. *Why do I not need to replan at this point?*

For each of these questions, it is imagined that a human user of an AI system would like to be able to query the system as to its underlying reasoning with respect to particular decisions. Although the concept of Explainable AI has been discussed since at least 2004 [35], current motivation centres around the need to describe, as well as understand, the reasoning underlying model-free methodologies such as *deep learning*. In the new age of GDPR [11], automated systems are now legally required (in Europe) to be able to explain how recommendations were computed. Thus while data mining techniques have been widely employed to learn models from “big data” sets, the task of communicating the basis of those models in ways that people can comprehend remains a significant challenge.

The approach we advocate here is to employ *Computational Argumentation* [24] for structuring the reasons underlying the decision making in an AI system and *Argumentation-based Dialogue* [37] for structuring the interaction between the AI system and a human user. Computational Argumentation is a logical model of reasoning that has its origins in philosophy and provides a means for organising evidence for (or against) particular claims (or decisions). Argumentation-based Dialogue is a related methodology that is used to guide interchanges between two (or more) agents and has been explored within the *Multi-Agent Systems (MAS)* community as an extended form of negotiation where agents can not only exchange claims, but also their reasons for believing (or disbelieving) those claims. There is a natural pairing between Explainable AI and Argumentation: the first requires the need to explain decisions and the second provides a method for linking any decision to the evidence supporting it. In this paper, we describe how the two are connected and illustrate the utility of argumentation-based dialogue as a technique for implementing Explainable AI in a human-robot system.

The details of our approach are described in the first part of this paper, followed by an example in which our methodology is applied to illustrate the effectiveness of our approach. We provide a brief review of applications of argumentation to support decision making

and a short history of explanation in AI systems. We close with a summary of our contributions and directions for future work.

2 BACKGROUND: COMPUTATIONAL ARGUMENTATION

Computational Argumentation [24] is a logic-based methodology for knowledge representation and reasoning in which evidence is structured as either supporting or attacking specific conclusions. Computational argumentation (hereafter simply *argumentation*) is a powerful mechanism for modelling an agent's beliefs and process of making decisions. For at least two decades, argumentation has been considered in the context of MAS and formalised for two-party interaction in the concept of argumentation-based dialogue, where agents exchange information (in the form of "arguments") with the goal of reaching *joint conclusions*. Formal argumentation-based *dialogue games* are mechanisms that define rules for such exchanges. In this section, we provide a very brief, high-level description of the fundamental formalisms in argumentation and then go into more detail with respect to argumentation-based dialogue, in order to demonstrate the use of dialogue games as a mechanism for agents to formulate and provide *explanation* about the reasons behind the decisions they make.

2.1 Argumentation

The logic underlying argumentation begins by defining a language, $\mathcal{L}$, in which all knowledge relevant to the situation at hand is represented. An argument, $\mathcal{A}$, consists of a set of *beliefs*, $\mathcal{S}$, and a *conclusion*, c . The set of beliefs comprise the evidence necessary to prove, or *derive* or *infer*, the conclusion, frequently represented as $\mathcal{S} \vdash c$. The symbol $\vdash$ is often referred to as an *inference mechanism*. All the bits are taken together to form an argument: $\mathcal{A} = (\mathcal{S}, c)$.

A belief, $b \in \mathcal{S}$, is typically an atomic piece of knowledge, such as "It is raining". This can be represented as a (logic) proposition (without variables): $b = \text{It is raining today}$, or as a (logic) predicate (with variables): $b = \text{Weather}(V1, V2)$. The predicate can be instantiated with values for the variables, such as: $V1 = \text{raining}, V2 = \text{today}$, to form the instantiated predicate: $\text{Weather}(\text{raining}, \text{today})$. In traditional systems, propositions and instantiated predicates are computed as either True or False.

We represent an agent's set of beliefs (or *knowledge base*) as Σ , and argumentation provides a structure over Σ in which beliefs can be organised as supporting, attacking or unrelated to other beliefs. Traditional argumentation considers beliefs (as above) and the relationships between beliefs to be binary, so we could say $b1 \rightarrow b2$ or $b1 \rightarrow \neg b2$, to indicate that belief $b1$ supports or attacks $b2$, respectively. More recent applications of argumentation take into account the *strength* of beliefs and support (or lack of support) for relationships between beliefs, often in the form of *confidence* about a belief and *trust* in the source of the information (i.e., data) from which the belief was derived. These are frequently represented as real values between 0 (meaning no confidence or trust) and 1 (maximum confidence or trust).

2.2 Dialogue Games

Computational argumentation-based **dialogue** is a formalism in which participants engage in structured, rule-guided, goal-oriented

exchange according to specific protocols [15, 18, 22, 37]. The rules of such dialogues provide guarantees concerning completion and ensure that infinite dialogues cannot occur (for example, by restricting a participant from uttering the same argument more than once in a dialogue). Here, we employ the term *dialogue* exclusively to refer to *computational argumentation-based dialogue*; we do not imply any application of or contribution to natural language dialogue¹.

The theory underlying computational argumentation-based dialogue is derived from modelling human interaction. Early research in the computational argumentation community identified six primary types of dialogue [37] and related work over the ensuing decade produced a number of other types. Each type is distinguished by participants' knowledge and their individual and shared goals:

- *eristic* [37]—participants quarrel verbally;
- *information-seeking* [37]—one participant asks a question that she does not know the answer to and believes the other participant can answer;
- *command* [12]—one participant tells another what to do;
- *inquiry* [17]—both participants seek an answer to a question that neither knows the answer to;
- *chance discovery* [16]—a new idea arises out of exchange between participants;
- *negotiation* [23]—participants bargain over the allocation of a scarce resource;
- *deliberation* [19]—participants decide together on taking a particular action;
- *education* [29]—one participant wants to teach another something;
- *verification* [8]—one participant asks a question that she knows the answer to and believes another does too, so aims to verify her belief; and
- *persuasion* [22]—one participant tries to alter the beliefs of another.

We have implemented three types of argumentation-based dialogue to support shared decision making in a human-robot team: *inquiry*, *information-seeking* and *persuasion*. Figure 1 illustrates which of these dialogue types are appropriate given the beliefs of each participant. In our case, there are two participants: a robot (R) and a human (H). We represent each participant's knowledge base as a set of beliefs. Each participant may have knowledge that something is true ("believes b ") or false ("believes $\neg b$ "), or may have no knowledge about it ("does not know b "). We can use the beliefs in a participant's knowledge base to construct arguments for or against particular decisions. When one participant shares their beliefs, through argumentation-based dialogue, they can alter the knowledge base of the other participant, either by adding new beliefs (through *inquiry* and *information-seeking*) or changing existing beliefs (through *persuasion*).

Each dialogue game is structured as a series of "moves" in which the participant who initiates the dialogue makes a statement about a belief (b). The other participant can respond by providing information about b that supports (agrees with) or attacks (disagrees

¹Though the overlap between NLP and Argumentation is not discussed here, recent work in *argument mining* coming from the NLP community is developing automated methods for constructing arguments from natural language corpora. The interested reader is referred to the *ArgMining* workshop series, most recently here: <https://www.research.ibm.com/haifa/Workshops/argmining18/index.shtml>.

	H believes b	H believes $\neg b$	H does not know b
R believes b	agreement (no dialogue)	disagreement (persuasion)	lack of info (info-seek)
R believes $\neg b$	disagreement (persuasion)	agreement (no dialogue)	lack of info (info-seek)
R does not know b	lack of info (info-seek)	lack of info (info-seek)	shared lack (inquiry)

Figure 1: Cases for different types of dialogues (from [27]).

with) the opinion of the first participant. An example illustrating the structure of a *persuasion* dialogue is shown in Figure 2. The opening move is an *assert(b)* statement in which the first participant states that they believe b . The second participant has three legal moves: they can agree with the first participant (*accept(b)*), they can disagree (*assert($\neg b$)*), or they can ask for explanation about the first participant's assertion (*challenge(b)*), essentially asking why the first participant believes b .

Figure 2: Persuasion dialogue (from [27]).

It is exactly this third type of move which is our focus here—because the response to a challenge is to *explain* the reasons behind the belief that was challenged. In the remainder of this paper, we first describe an implementation, experimental setup and evaluation that we produced previously with the goal of analysing the impact of employing argumentation-based dialogue to support human-robot decision making (Section 3). But here, we focus our analysis not on performance metrics (as in previous analyses), but rather on *explanation*. So, second, we examine how frequently participants asked for explanation, whether the explanation was accepted, and whether the explanation had any impact on the ultimate performance of the human-robot team (Section 4). Our contribution here lies in this demonstration of the efficacy of argumentation as a means to facilitate explanation in a Human/AI System. The next section describes our framework and experimental environment.

Figure 3: The ArgHRI experimental setup (from [3]).

3 OUR APPROACH: ARGHRI FRAMEWORK AND EXPERIMENTAL SETUP

In earlier work, we developed *ArgHRI*, a formal methodology [27] which we subsequently implemented as a framework [2] that supports human-robot interaction through the use of argumentation-based dialogue. We evaluated ArgHRI and its impact on human decision making by undertaking a comparative user study, aspects of which we have documented previously [2–4]. Here we elaborate on the ability of the framework to *explain*, a feature of our methodology that we have not previously described in detail.

ArgHRI consists of a *dialogue manager* [2], which interacts with human users and employs the *ArgTrust* [32, 33] engine to perform reasoning, and a *robot controller*, which employs the *HRTTeam* framework [9, 28, 31] for managing robot behaviour and interacting with physical or simulated robots. The dialogue manager controls all human-robot dialogues and dialogue-related events, including selection of appropriate argumentation-based dialogue type (illustrated in Figure 1), dialogue move generation (an example is shown in Figure 2), translation into scripted chat-style content for presentation to humans (examples are contained in Figures 4, 5 and 6) and maintenance of dialogue history. All dialogue events follow the protocols and theoretical framework detailed in [27]. Figure 3 shows (a) the HRTTeam arena where the robots operate, (b) one of our (physical) robots, (c) an example of the type of target that the robots seek, and (d) the user interface that was used for the experiments described here.

We designed a comparative user study in order to evaluate the effectiveness of our framework which provides explanation, in contrast with a baseline which does not provide explanation. In our

experiments, we employed the ArgHRI framework to facilitate a *Treasure Hunt Game (THG)* [27] in which a human and a robot work together as a team to find “treasures” in an arena (Figure 3a). The robot moves around inside the arena and uses sensors to gather data and communicate with the human user. The human operates remotely from the robot, receives data about the robot’s environment from the robot and communicates with the robot to make *shared decisions* about tasks the robot will perform. The robot does not have enough energy to conduct an exhaustive search of the arena in order to find all the treasures, so the shared mission for the human-robot team is to locate and identify as many treasures in the arena as possible before the robot runs out of energy.

For our experiments, we designed a series of three “decision points” in order to control a sequence of interactions so that we could evaluate the impact of ArgHRI on shared decision making:

- (1) *Where to go?* The human and robot decide together which rooms should be searched for treasures.
- (2) *How to get there?* The human and robot decide together a high-level path for visiting the agreed-upon rooms.
- (3) *What is found there?* The human and robot decide together whether the images captured by the robot contain treasures.

As detailed in earlier work [4], we analysed our experimental results and showed that the argumentation-dialogue-enabled mode of interaction outperformed a traditional supervisory mode of interaction, both in terms of performance metrics (such as game score and robot energy efficiency) and user satisfaction metrics (such as confidence in decisions and trust in the robot as a collaborator). Here, we analyse the interactions between the human user and the robot through the lens of *explainability* and examine the dialogues in detail for situations where the user asked for explanation and whether the explanation resulted in improved performance and/or user satisfaction.

Our user study involved 60 participants teaming with either physical or simulated robots. See Table 1 for a summary of the demographics of the participants. Each participant played two Treasure Hunt Games, under two conditions. In the baseline condition, participants engaged with *Robot Mary*, whose behaviour was that of a robot responding to a human *supervisor*: the human issued tasks and Mary performed them. In the comparison condition, participants engaged with *Robot Fiona*, whose behaviour was that of a robot collaborating with a human *peer*: the human and robot discussed tasks for Fiona to perform, and Fiona only performed them once agreement was reached. The order of games was mixed, so that half of the participants played a game with Mary first and the others played a game with Fiona first.

	total	male	female
physical robots	27	20	7
simulated robots	33	22	11

Table 1: Participants’ demographics. Most were undergraduates (97%). Two-thirds were 18-24 years old (74%); the rest were 25-39 years old. About half had no prior experience with robots; the rest had less than a year’s experience.

Objective performance metrics included distance travelled by robots (a proxy for energy usage) and the score obtained by each team. As detailed in [4], the performance metrics were statistically significantly better for the human-Fiona (collaborative) versus the human-Mary (supervisory) teams. In addition, a comparison of subjective measures, obtained through surveys in which users indicated their level of satisfaction and trust working with the robot, were also significantly better for the human-Fiona teams.

As above, our goal in this paper is not to review the performance results in any detail, since we have already done that in earlier papers [2–4], but rather to explore the extent to which *explanation* played a role in these results. The dialogue game structure described in Section 2.2 provides the opportunity for either player in the game to *challenge* the other, effectively asking the other to explain their reasoning. In the next section, we delve into the dialogues which occurred during the experiments and analyse their content through this lens.

4 ARGUMENTATION AS A MEANS FOR PROVIDING EXPLANATION

Table 2 contains the total numbers of dialogues which occurred for each of the three decision points. This data illustrates the distribution of the different types of dialogue. The percentages for the first three columns of data are computed as the ratio of the number of dialogues of each type to the total number of dialogues. The percentage in the fourth column (“total”) is computed as the ratio of the total number of dialogues to the total number of participants, representing the proportion of the Treasure Hunt Games that were played with Robot Fiona where any dialogue occurred. Refer back to Figure 1, where the opportunities for dialogue are explained. If the human user and the robot are in agreement with the opening move of a dialogue game, then dialogue does not (need to) occur. These numbers indicate the large proportion of the times (60%–83% of the time) when each human-robot team disagreed when reaching each decision point and were able to employ the dialogue facility in order to find agreement. This statistic alone justifies the need for a dialogue facility to support collaboration in a human-robot team.

Next we analyse the opportunities for explanation. As described earlier, when a participant issues a *challenge()* move in an argumentation-based dialogue game, this indicates that the participant is seeking explanation for the assertion previously offered by the other dialogue participant. Data is presented in Tables 3, 4 and 5. Note that for decision point 2 (Table 4, *How to get there?*), there were no challenges that occurred in any inquiry dialogues, so those lines are omitted from the table. As well, for decision point 3 (Table 5, *What is found there?*), there were no persuasion dialogues.

The results show that, for the first two decision points (Table 3, *Where to go?* and Table 4, *How to get there?*), when a challenge was issued, most were accepted after explanation was provided (acceptance rates range from 75 – 100%). However, for the third decision point (Table 5, *What is found there?*), when a challenge was issued, most were rejected after explanation was provided (acceptance rates range from 0 – 42%). This observation is discussed later in this section.

<i>Where to go?</i>				
	<i>persuasion</i>	<i>info-seek</i>	<i>inquiry</i>	<i>total</i>
physical	10 (56%)	5 (28%)	3 (17%)	18 (67%)
simulated	16 (59%)	11 (41%)	0	27 (81%)
<i>How to get there?</i>				
	<i>persuasion</i>	<i>info-seek</i>	<i>inquiry</i>	<i>total</i>
physical	11 (65%)	6 (35%)	0	17 (63%)
simulated	20 (91%)	2 (9%)	0	22 (67%)
<i>What is found there? (first 4 opportunities)</i>				
	<i>persuasion</i>	<i>info-seek</i>	<i>inquiry</i>	<i>total</i>
physical	0	75 (67%)	16 (14%)	91 (84%)
simulated	0	84 (69%)	21 (17%)	105 (80%)

Table 2: Number of dialogues of each type, for each decision point. The percentages in the first three columns are the ratio of the number of dialogues of each type to the total number of dialogues. The percentage in the fourth column is the ratio of the total number of dialogues to the total number of decisions—if the human and robot agree when they reach a decision point, then dialogue is not necessary.

<i>Where to go?</i>			
	<i>persuasion</i>		
	<i>challenges</i>	<i>accepts</i>	<i>rejects</i>
physical	6 (60%)	5 (83%)	1 (17%)
simulated	10 (63%)	9 (90%)	1 (10%)
	<i>information-seeking</i>		
	<i>challenges</i>	<i>accepts</i>	<i>rejects</i>
physical	4 (80%)	4 (100%)	0
simulated	6 (55%)	6 (100%)	0
	<i>inquiry</i>		
	<i>challenges</i>	<i>accepts</i>	<i>rejects</i>
physical	2 (67%)	2 (100%)	0
simulated	0	0	0

Table 3: Number of challenges at the 1st decision point. The percentages in the first column are the ratio of the number of challenges to the number of dialogues of that type. The percentages in the second two columns are the ratios of the number of accepts or rejects, respectively, to the number of challenges.

4.1 Example Dialogue Analysis.

In order to better understand the opportunities for explanation and how these unfolded for the experiments described here, we now present an example dialogue and discuss it in detail. This dialogue is taken from one of the experiments with a simulated robot and is representative of the exchanges from both simulated and physical robot teams. We selected a simple exchange for illustration here, but note that our framework supports more complex dialogue structures, such as nested dialogues (where one dialogue is “paused” while another occurs) or multiple dialogues in parallel (where locutions from two or more different dialogues are interleaved).

Figure 4 contains the transcript from the first decision point. Line 1 contains the standard opening statement by Robot Fiona at

<i>How to get there?</i>			
	<i>persuasion</i>		
	<i>challenges</i>	<i>accepts</i>	<i>rejects</i>
physical	8 (73%)	6 (75%)	2 (25%)
simulated	12 (60%)	9 (75%)	3 (25%)
	<i>information-seeking</i>		
	<i>challenges</i>	<i>accepts</i>	<i>rejects</i>
physical	1 (17%)	1 (100%)	0
simulated	1 (50%)	1 (100%)	0

Table 4: Number of challenges at the 2nd decision point. The percentages in the first column are the ratio of the number of challenges to the number of dialogues of that type. The percentages in the second two columns are the ratios of the number of accepts or rejects, respectively, to the number of challenges.

<i>What is found there? (first 4 opportunities)</i>			
	<i>information-seeking</i>		
	<i>challenges</i>	<i>accepts</i>	<i>rejects</i>
physical	12 (17%)	3 (25%)	9 (75%)
simulated	12 (15%)	5 (42%)	7 (58%)
	<i>inquiry</i>		
	<i>challenges</i>	<i>accepts</i>	<i>rejects</i>
physical	4 (5%)	0	4 (100%)
simulated	9 (11%)	3 (33%)	6 (67%)

Table 5: Number of challenges at the 3rd decision point. The percentages in the first column are the ratio of the number of challenges to the number of dialogues of that type. The percentages in the second two columns are the ratios of the number of accepts or rejects, respectively, to the number of challenges.

the start of a Treasure Hunt Game. In line 3, the system recognises that the robot’s goal (to visit 4 rooms) is different from the human’s goal (to visit 1 room). A *persuasion* dialogue ensues, in which the robot tries to convince the human to change their mind (belief). Line 3 contains an *assert(b)* move (see Figure 2), made by the robot who is the initiator of the dialogue. Line 4 contains a *challenge(b)* move, the human’s response. This response is met by one or more *assert(s)* moves in which the robot provides to the human the evidence behind the claim which the robot asserted earlier (i.e., each $s \in S$ is part of the *support* for the conclusion of the argument (S, b)). Line 5 contains the *explanation*—this evidence (i.e., that the robot had inside knowledge from the “Game Master” which supports its belief that 4 rooms should be visited). Line 6 contains the human’s agreement (acceptance). This exchange is scored in Table 3 for *persuasion* dialogue as one of the 10 challenges which resulted in one of the 9 accepts.

Figure 5 contains the transcript from the second decision point for the same game. Note that in between the dialogue in Figure 4 and this dialogue, the robot and human agreed that the robot should explore rooms 2, 4, 5 and 6 in the arena. Line 1 contains the statement by Robot Fiona which opens the exchange for the second

1	Robot>	Welcome to Treasure Hunt Game! I am Robot Fiona. How many rooms would you like me to visit?
2	Human>	I would like to visit 1 room.
3	Robot>	Our goals are different. I believe that we should visit 4 rooms instead of visiting 1 room. Do you "Agree" or "Disagree" ?
4	Human>	I would like to know why you would like to visit 4 rooms.
5	Robot>	It is because I received clues from the Game Master. Please "Agree" or "Disagree" with my goal and click "Send".
6	Human>	I agree with your proposed goal to visit 4 rooms.
7	Robot>	Great! We will follow my goal and visit 4 rooms.

Figure 4: Example transcript for 1st decision point, *Where to go?* A *persuasion* dialogue occurred. A challenge and explanatory response appear in lines 4 and 5.

decision point. Since the human responds in line 2 with "I don't know", this represents lack of information on the part of the human. Since the robot responds in line 3 that it has a plan, then the situation calls for an *information-seeking dialogue* (see Figure 1). The robot opens with asserting its plan (to visit the rooms in this order: 4 → 5 → 2 → 6). The human challenges (line 4) this assertion by asking *why?* The robot responds (line 5) with its evidence: that it has calculated a lower path cost for visiting rooms in that sequence. The human accepts this argument (line 6). This exchange is tallied in Table 4 for *information-seeking* and shows the 1 challenge which was issued and accepted in the experiments with simulated robots.

1	Robot>	What is the the order you want me to explore rooms?
2	Human>	I don't know. Do you have a plan?
3	Robot>	Yes. I would like to go to Rooms 4->5->2->6.
4	Human>	I would like to know why you would like to visit Rooms 4->5->2->6.
5	Robot>	According to my calculations, going to Rooms 4->5->2->6 has a lower path cost. Please "Agree" or "Disagree" with my plan.
6	Human>	OK, I agree with you. Please go to Rooms 4->5->2->6.

Figure 5: Example transcript for 2nd decision point, *How to get there?* An *information-seeking* dialogue occurred. A challenge and explanatory response appear in lines 4 and 5.

Figure 6 contains the transcript from the third decision point for the same game. This point is reached after the robot and human have agreed on the rooms to visit and the robot's path (order) for exploring the rooms (agreed in the first and decision points, respectively). For each room the robot visits, it performs a *sensor-sweep* task and captures a series of images around the room. It then sends the images to the human so that the team can collaboratively decide if there are any items of treasure in the room. For the purposes of

1	Robot>	Do you see any of the following Treasures in Room 4?
2	Human>	I believe that there is a pink can in Room 4.
3	Robot>	Great. You are Correct! The Game Master confirmed that the pink can is in room 4.
4	Robot>	Do you see any of the following Treasures in Room 5?
5	Human>	I believe that there is no treasure in Room 5.
6	Robot>	You are correct. The Game Master confirmed that there is no treasure in room 5.
7	Robot>	Do you see any of the following Treasures in Room 2?
8	Human>	I don't know anything about the treasure in Room 2. Do you?
9	Robot>	OK. But I only guess the colour of the treasure in the room. Do you see any blue treasure in this room?
10	Human>	I would like to know why.
11	Robot>	According to my colour sensor, I believe that there may be blue treasure in this room. Please "Agree" or "Disagree" and click "Send".
12	Human>	I disagree with you. I don't see any blue colour treasure in this room.
13	Robot>	Do you see any of the following Treasures in Room 6?
14	Human>	I believe that there is no treasure in Room 6.
15	Robot>	You are correct. The Game Master confirmed that there is no treasure in room 6.

Figure 6: Example transcript for 3rd decision point, *What is found there?* In this example, there were 4 opportunities for dialogue, because the robot visited 4 rooms (Room 4, 5, 2 and 6, in that order). Only the third room visited (Room 2) resulted in a dialogue, in this case, an *inquiry* dialogue. A challenge and explanation appear in lines 10 and 11.

our experiments, we gave the robot the ability to detect colours, but not shapes. So, the robot can identify that there is a blue object in the room, but cannot distinguish whether it is the blue can or the blue bottle; so the robot needs help from the human to make the final identification. For each room the robot visits (4 rooms in this example), the robot asks the human if they can see any items of treasure in the images collected by the robot and shared with the human (lines 1, 4, 7 and 14, for rooms 4, 5, 2 and 6, respectively). A challenge occurs when discussing the images taken in room 2 (lines 7–12). This is an instance of an *inquiry* dialogue, because neither the robot nor the human has the answer to the question about whether there are treasures in room 2. So they must work together to find the answer. In line 9, the robot asserts that it detects blue treasure. In line 10, the human asks why, and the robot's explanation ensues in line 11. The human disagrees with the robot's conclusion (line 12) and thus rejects it. This exchange is tallied in Table 5 for *inquiry* dialogue as one of the 9 challenges which resulted in one of the 6 rejections.

4.2 Does Explanation Help?

We examine the results in an attempt to ascertain whether explanation can be correlated with improved performance and/or user satisfaction. The two performance metrics we consider are *game score*, an indication of how well the team performed the task of correctly identifying treasures, and *distance travelled*, an indication of how efficient the team was. If our hypothesis is correct—that explanation helps to improve performance and conserve energy—then the teams that took advantage of the explanation feature by challenging each others' assertions would have higher game scores and lower distance measures. The user satisfaction metric we consider is an average of two scores that we obtained from users through a survey which they completed at the end of the two games they played. They were asked a number of questions about their experience, which they responded to on a 7-point Likert scale. These included whether they felt that Robot Fiona helped them resolve problems and whether they thought that Robot Fiona helped them with their decision making. The average of these scores is analysed here, as a proxy for *user satisfaction*.

We tallied the number of challenges each user engaged in during their games with Robot Fiona, divided into the number of dialogues engaged in. The overall average rate of challenges (across all three dialogue types) was 28% for physical robot teams and 30% for simulated. We labeled each user's rate of challenges as "high" or "low", depending on whether their tally was above or below the average, respectively. Then we computed statistics across these samples, to see if there are differences. The results are shown in Table 6. While there are discernible differences, these are not statistically significant (assessed using t-tests for independent samples). We also ran linear regression tests on the raw challenge percentage rates vs the game scores, distance travelled values and user satisfaction metrics. There are no identifiable correlations.

4.3 Discussion.

We discuss three observations made when analysing these results. First, the high rejection rate for challenges occurring at the third decision point (Table 5) was surprising. Two types of dialogue occurred at the third decision point: *information-seeking* and *inquiry*. In both cases, the human engages in dialogue in order to *add* to their knowledge base. With information seeking, the human believes

that the robot has knowledge that they would like to acquire. With inquiry, the human believes that the robot also does not have any knowledge of the question being posed and that the team can find the answer jointly. But at this decision point, unlike the first two, the human has the upper hand with respect to being able to analyse images. At the first two decision points, the human is likely guessing: Where to go? The robot knows better because it has received "hints" from the Game Master (oracle component of the system). How to get there? The robot's belief is backed up by computational methods (the A* path-planning algorithm [14]) to determine the shortest path through the rooms that the human-robot team selected. At the third decision point, the human's ability to process images and identify objects is superior to that of the robot, so it is more likely—as shown—that the human's belief is accepted.

Second, the lack of correlation between performance, satisfaction and rate of issuing challenges was disappointing. The performance scores and user satisfaction metrics were computed once, at the conclusion of each team's experimental run. But during each run, there were many opportunities for dialogue and challenges; and we do not have any performance metrics that can be assessed for individual dialogues. This means that we cannot clearly attribute the instance of an explanation to a particular performance metric.

Finally, some users never challenged (5 out of 27 participants in the physical robot experiments, and 8 out of 33 participants in the simulated experiments). We re-computed our statistics in Table 6, not counting the users who never issued a challenge; and we also computed statistics for participants grouped into "never challenged" versus "challenged", and the outcome was the same—there were no statistically significant differences in performance metrics when the users are given the ability to ask for explanation. We suspect that this phenomena says more about the way the experiments were designed originally and less about whether explanation helps or not. In future work, we will run a similar experiment with human subjects, but this time we will design the experiment to look explicitly at the question of whether explanation helps or hinders users. Some users find too much information overwhelming, so this should also be investigated with respect to explanation.

5 RELATED WORK

Computational Argumentation has been used in a variety of domains, including decision support [1], intelligence analysis [34] and legal reasoning [36]. Argumentation-based dialogue theory has similarly been studied by researchers across a wide variety of disciplines, including Law, Medicine, Artificial Intelligence and Multi-Agent Systems [6, 21]. 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 [38] in the agent-oriented software engineering community. Belesiotis *et al.* [5] 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. Previous work [25, 30], has demonstrated that argumentation can be an effective technique to support human

physical			
	<i>game score</i>	<i>distance travelled</i>	<i>user satisfaction</i>
low challenge rate	1720 (361)	2153 (637)	5.47 (1.20)
high challenge rate	1646 (447)	1895 (515)	5.71 (1.18)
simulated			
	<i>game score</i>	<i>distance travelled</i>	<i>user satisfaction</i>
low challenge rate	1759 (345)	1245 (302)	5.44 (1.64)
high challenge rate	1750 (573)	1504 (348)	5.97 (1.17)

Table 6: Mean (and standard deviation) of game score, distance travelled and user satisfaction metrics for games, grouped by low and high challenge rates.

decision making. Another early system [7] demonstrated agents interacting with people to help them understand plans that have been created jointly by multiple collaborating actors.

An early paper on *Explainable AI* [35] describes the concept as growing out of *expert systems* in the 1970's, where interactive intelligent systems aided human users in searching for information in large (at the time) databases which attempted to encode the knowledge and expertise of people experienced in particular fields, such as medicine [26]. The methods discussed in [35] are applied to a military domain for analysing battlefield decisions. A current survey of the field [20] highlights several approaches taken in the literature: *contrastive explanation*, where two alternatives are compared; *attribution theory*, where cause-and-effect relationships are identified; and *explanation selection*, where a minimal set of causes for an event are chosen (because people rarely want to know *all* the reasons something happened). Further, [20] describes two important areas for investigation: *explanation evaluation*, where studies assess the impact of different approaches to explanation, such as simple vs complex or general vs specific; and *explanation as conversation*, where different forms of human interaction are studied as models for conveying explanations to users. More recently, the community has identified explainability as a key challenge problem that needs to be addressed by AI research [13]. The aim is for such capability to accompany *machine learning* systems and be able to translate the reasoning behind complex models such as *deep learning*. The methods described here, which propose the use of Computational Argumentation to facilitate Explainable AI are complementary to this goal: arguments can be mined from data, and that data-mining process can incorporate a wide range of machine learning methods, including model-free techniques such as deep learning. Then the process of explaining is completed with argumentation-based dialogue in order to ensure appropriate and productive interaction with human users.

6 SUMMARY

We have described the *ArgHRI* framework and a user study which was conducted previously. Our contribution here is to analyse the data we collected through the lens of *Explainable AI* in an attempt to discover whether the ability of a user to ask “why” when interacting with a robot is reflected in improved performance metrics and user satisfaction. Our results are inconclusive with respect to this question, but encouraging and warrant further investigation.

Finally, we re-examine the questions posed in Section 1. Some of these sample questions are contained in the example transcripts we analysed in Section 4. Question Q1 (*Why did you do that?*) is inherent in Figure 4, line 4 and answered in line 5 (*Because I had insider information from the Game Master*). Question Q2 (*Why didn't you do something else (that I would have done)?*) is inherent in Figure 6, line 10 and answered in line 11 (*Because my colour sensor detected “blue” in the imagery*) and then refuted in line 12 (*I don't see anything “blue”*), implying that the human's answer is “something else” and indeed does something different by rejecting the robot's belief. Question Q3 (*Why is what you propose to do more efficient/safe/cheap than something else*) is inherent in Figure 5, line 4 and answered in line 5 (*Because I computed the path costs and found the optimal solution*).

In our current work, we are designing and planning to conduct follow-on experiments in which a human-robot team employs explanation more explicitly to support collaboration. Our analysis has shown that attention needs to be paid to the well-known *credit assignment* problem so that performance scores and user satisfaction metrics can be unequivocally attributed to the use of explanation, or not. Our earlier results demonstrated clearly that the use of argumentation-based dialogue improves both of these measures, and as shown here, the incorporation of explanation is a feature inherent in argumentation-based dialogue. How much individual explanations contribute to end results remains an open question and one that we will investigate.

We will also consider definition of performance metrics that relate outcomes to individual dialogues. In situations which are complex—precisely where the ability to have dialogue becomes important and interesting—the assessment of explanation clarity and associated impact is complicated. Here, it is likely that multiple actions will (need to) be performed before a task is complete enough that any meaningful performance metrics could be assigned to that task; and if multiple dialogues take place during task execution, it will be difficult—perhaps impossible—to assign a performance outcome to a particular dialogue. These challenges, and many others, will ensure that this line of investigation will continue in our future work.

ACKNOWLEDGMENTS

This work was partially funded initially by the US National Science Foundation (NSF) under grants #IIS-1116843, #IIS-1338884 and #CNS-1117761, by the US Army Research Laboratory under the Network Science Collaborative Technology Agreement #W911NF-09-2-0053 and more recently in the UK by the Engineering and Physical Sciences Research Council (EPSRC) under grant #EP/P010105/1. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders.

REFERENCES

- [1] Katie Atkinson, Trevor J. M. Bench-Capon, and Sanjay Modgil. 2006. Argumentation for Decision Support. In *Proceedings of DEXA*. 822–831.
- [2] Mohammad Q. Azhar. 2015. *Toward an argumentation-based dialogue framework for human-robot collaboration*. Ph.D. Dissertation. City University of New York.
- [3] Mohammad Q. Azhar and Elizabeth I. Sklar. 2016. Analysis of empirical results on argumentation-based dialogue to support shared decision making in a human-robot team. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 861–866.
- [4] Mohammad Q. Azhar and Elizabeth I. Sklar. 2017. A study measuring the impact of shared decision making in a human-robot team. *International Journal of Robotics Research (IJRR)* 36 (2017), 461–482. Issue 5–7.
- [5] Alexandros Belesiotis, Michael Rovatsos, and Iyad Rahwan. 2010. Agreeing on Plans Through Iterated Disputes. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Vol. 1. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 765–772.
- [6] Trevor J. M. Bench-Capon and Paul E. Dunne. 2007. Argumentation in artificial intelligence. *Artificial intelligence* 171, 10–15 (2007), 619–641.
- [7] Martin Caminada, Roman Kutlak, Nir Oren, and Wamberto W. Vasconcelos. 2014. Scrutable Plan Enactment via Argumentation and Natural Language Generation (Demonstration). In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. International Foundation for Autonomous Agents and Multiagent Systems.
- [8] Eva Cogan, Simon Parsons, and Peter McBurney. 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)*. New York, NY, USA: ACM Press, Utrecht, The Netherlands, 544–551.

- [9] Susan L. Epstein, Anoop Aror, Matthew Evanusa, Elizabeth I. Sklar, and Simon Parsons. 2015. Learning Spatial Models for Navigation. In *Spatial Information Theory (COSIT). Lecture Notes in Computer Science*, Vol. 9368. Springer, 403–425.
- [10] Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable Planning. *arXiv preprint arXiv:1709.10256*.
- [11] GDPR 2016. Regulation (EU) 2016 of the European Parliament and of the Council of the European Union on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). <http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf>.
- [12] Roderic A. Girdle. 1996. Commands in Dialogue Logic. In *Practical Reasoning: Proceedings of the First International Conference on Formal and Applied Practical Reasoning (FAPR), Lecture Notes in Artificial Intelligence*, D. M. Gabbay and H. J. Ohlbach (Eds.). Vol. 1085. Berlin: Springer, 246–260.
- [13] David Gunning. 2017. Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)* (2017).
- [14] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimal cost paths. *IEEE Transactions on Systems Science and Cybernetics* 4, 2 (1968), 100–107.
- [15] Joris Hulstijn. 2000. *Dialogue Models for Inquiry and Transaction*. Ph.D. Dissertation. University of Twente, Enschede, Netherlands.
- [16] Peter McBurney and Simon Parsons. 2001. Chance discovery using dialectical argumentation. *New Frontiers in Artificial Intelligence* (2001), 414–424.
- [17] Peter McBurney and Simon Parsons. 2001. Representing epistemic uncertainty by means of dialectical argumentation. *Annals of Mathematics and Artificial Intelligence* 32, 1–4 (2001), 125–169.
- [18] Peter McBurney and Simon Parsons. 2002. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language, and Information* 11, 3 (2002), 315–334.
- [19] Peter McBurney and Simon Parsons. 2004. A Denotational Semantics for Deliberation Dialogues. In *Proceedings of the 3rd International Joint Conference of Autonomous Agents and Multiagent Systems (AAMAS)*, 86–93.
- [20] Tim Miller, Piers Howe, and Liz Sonenberg. [n. d.]. Explainable AI: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*, Vol. 36.
- [21] Soren Holbech Nielsen and Simon Parsons. 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)*.
- [22] Henry Prakken. 2006. Formal systems for persuasion dialogue. *Knowledge Engineering Review* 21, 2 (2006), 163–188.
- [23] Iyad Rahwan, Sarvapali D. Ramchurn, Nicholas R. Jennings, Peter McBurney, Simon Parsons, and Liz Sonenberg. 2003. Argumentation-based negotiation. *The Knowledge Engineering Review* 18, 4 (2003), 343–375.
- [24] Iyad Rahwan and Guillermo R. Simari (Eds.). 2009. *Argumentation in Artificial Intelligence*. Berlin: Springer-Verlag.
- [25] Ariel Rosenfeld and Sarit Kraus. 2014. Argumentation Theory in the Field: An Empirical Study of Fundamental Notions. In *Workshop on Frontiers & Connections between Argumentation Theory and Natural Language Processing*.
- [26] Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green, and Stanley N. Cohen. 1975. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and biomedical research* 8, 4 (1975), 303–320.
- [27] Elizabeth I. Sklar and Mohammad Q. Azhar. 2015. Argumentation-based Dialogue Games for Shared Control in Human-Robot Systems. *Journal of Human-Robot Interaction* 4, 3 (2015), 120–148. <https://doi.org/10.5898/JHRI.4.3.Sklar>
- [28] Elizabeth I. Sklar, A. Tuna Özgelen, J. Pablo Muñoz, Joel Gonzalez, Mark Manashirov, Susan L. Epstein, and Simon Parsons. 2011. Designing the HRTTeam 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)*. Taipei, Taiwan.
- [29] Elizabeth I. Sklar and Simon Parsons. 2004. Towards the Application of Argumentation-based Dialogues for Education. In *Proceedings of the Third International Conference of Autonomous Agents and Multi Agent Systems (AAMAS)*, 14207–1421.
- [30] Elizabeth I. Sklar, Simon Parsons, Zimi Li, Jordan Salvit, Senni Perumal, Holly Wall, and Jennifer Mangels. 2016. Evaluation of a trust-modulated argumentation-based interactive decision-making tool. *Journal of Autonomous Agents and Multi-Agent Systems* 30, 1 (2016), 136–173.
- [31] Elizabeth I. Sklar, Simon Parsons, A. Tuna Özgelen, Eric Schneider, Michael Costantino, and Susan L. Epstein. 2013. HRTTeam: a framework to support research on human/multi-robot interaction (Demonstration). In *Proceedings of Autonomous Agents and Multiagent Systems (AAMAS)*. St Paul, MN, USA.
- [32] Yuqing Tang, Kai Cai, Peter McBurney, Elizabeth I. Sklar, and Simon Parsons. 2012. Using argumentation to reason about trust and belief. *Journal of Logic and Computation, Special Issue on Agreement Technologies* 22, 5 (2012), 979–1018.
- [33] Yuqing Tang, Elizabeth I. Sklar, and Simon Parsons. 2012. An Argumentation Engine: ArgTrust. In *Proceedings of the Workshop on Argumentation in Multiagent Systems (ArgMAS) at Autonomous Agents and MultiAgent Systems (AAMAS)*. Valencia, Spain.
- [34] Alice Toniolo, Hengfei Li, Timothy J. Norman, Nir Oren, W. R. Ouyang, Mani B. Srivastava, Timothy Dropps, John A. Allen, and Paul Sullivan. 2015. Enabling Intelligence Analysis through Agent-Support: The CISpaces Toolkit. In *Proceedings of AAMAS*.
- [35] Michael van Lent, William Fisher, and Michael Mancuso. 2004. An Explainable Artificial Intelligence System for Small-unit Tactical Behavior.
- [36] Bart Verheij. 2003. Dialectical argumentation with argumentation schemes: An approach to legal logic. *Artificial intelligence and Law* 11 (2003), 167–195.
- [37] Douglas Walton and Erik C. W. Krabbe. 1995. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. Albany, NY, USA: State University of New York Press.
- [38] Wayne Wobcke, Van Ho, Anh Nguyen, and Alfred Krzywicki. 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 (MMUI)*, Vol. 57. Darlinghurst: Australian Computer Society, Inc., Sydney, Australia, 61–66.