Providing Arguments in Discussions on the Basis of the Prediction of Human Argumentative Behavior

ARIEL ROSENFELD and SARIT KRAUS, Bar-Ilan University

Argumentative discussion is a highly demanding task. In order to help people in such discussions, this article provides an innovative methodology for developing agents that can support people in argumentative discussions by proposing possible arguments. By gathering and analyzing human argumentative behavior from more than 1000 human study participants, we show that the prediction of human argumentative behavior using Machine Learning (ML) is possible and useful in designing argument provision agents. This paper first demonstrates that ML techniques can achieve up to 76% accuracy when predicting people's top three argument choices given a partial discussion. We further show that well-established Argumentation Theory is not a good predictor of people's choice of arguments. Then, we present 9 argument provision agents, which we empirically evaluate using hundreds of human study participants. We show that the Predictive and Relevance-Based Heuristic agent (PRH), which uses ML prediction with a heuristic that estimates the relevance of possible arguments to the current state of the discussion, results in significantly higher levels of satisfaction among study participants compared with the other evaluated agents. These other agents propose arguments based on Argumentation Theory; propose predicted arguments without the heuristics or with only the heuristics; or use Transfer Learning methods. Our findings also show that people use the PRH agents proposed arguments significantly more often than those proposed by the other agents.

Categories and Subject Descriptors: I.2.m [**Computing Methodologies**]: Artificial Intelligence—*Miscellaneous*

General Terms: Human Factors, Experimentation, Theory

Additional Key Words and Phrases: Argumentation, human argumentation, automated advice, advising agents, prediction

ACM Reference Format:
Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. ACM Trans. Interact. Intell. Syst. 6, 4, Article 30 (December 2016), 33 pages.
DOI: http://dx.doi.org/10.1145/2983925

1. INTRODUCTION

Argumentative dialogs are part of everyday life. Nevertheless, participating in argumentative dialogs can be a highly demanding task for humans, both mentally and emotionally, as shown in discursive psychology research [Edwards 1997; Krauss 2001]. An intelligent, automated agent can help relieve some of these demands by providing contextual arguments for its human user while the user engages in an argumentative dialog. This can be achieved using the agent's computational advantage and knowledge over humans' argumentative behavior.

The reviewing of this article was managed by special issue associate editors Nava Tintarev, John O'Donovan, and Alexander Felfernig.
Authors' addresses: A. Rosenfeld and S. Kraus, Computer Science Department, Bar-Ilan University Ramat Gan, 5290002 Israel; emails: arielros1@gmail.com, sarit@cs.biu.ac.il.
Permission to make digital or hard copies of part or all 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 show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2016 ACM 2160-6455/2016/12-ART30 $15.00
DOI: http://dx.doi.org/10.1145/2983925

When suggesting an argument to a human user, an agent can consider two possible approaches. First, the agent can suggest an argument that the person has (probably) considered and is prone to use anyway. Second, it can suggest an innovative argument that the person has (probably) not considered.

In order to estimate which arguments a person is likely to use in a given discussion, an agent can use several prediction methods:

(1) Argumentation Theory (see Walton [2009] for an excellent summary) can be employed to logically analyze argumentative discussions. This analysis can be employed as a prediction of which arguments a deliberant might use.
(2) Heuristics can account for the temporal nature of the argumentative dialog and provide a prediction of which arguments are likely to be used next given the current state of the dialog. For example, a simple heuristic might predict that a user's next argument will directly relate to the last argument presented in the dialog.
(3) Machine Learning (ML) techniques, utilizing previous argumentative choices of *other* users on the same topic, can provide a prediction of which argument is expected to be presented next in a discussion. That is, ML can estimate the probability of each argument to be used next in a given dialog based on previous records of human argumentative choices in similar dialogs on the same topic.
(4) Transfer Learning (TL) methods (see Pan and Yang [2010] for a survey) can use previous argumentative discussions by *the same user* and transfer the observed argumentative selections to a new domain, specified as the *target domain*. Given a partial discussion in the target domain, a prediction of the next presented argument in the discussion is generated according to the user's argumentative selections in *other domains*.

In this work, we first tackle the problem of predicting human argumentative behavior. We present a comprehensive empirical evaluation of these four prediction methods. Thus far, the suggested prediction methods were not examined with respect to their predictive abilities in predicting human argumentative behavior. To evaluate the four suggested methods, we conducted a series of four experiments varying in their argumentative complexity: from multiple-choice argumentative questionnaires based on argumentative scenarios taken from the literature, to online chats in natural language in which two deliberants engage in an online discussion over a controversial topic. Next, we present 9 novel argument provision agents, based on the aforementioned prediction methods. These agents are extensively evaluated in online discussions with hundreds of human study participants.

With this extensive human study, with over 1000 human study participants[1] spread across 4 experiments, we make the following contributions:

(1) We establish the lack of predictive abilities of the existing Argumentation Theory in all of the examined experimental settings.
(2) We show that ML can be an extremely useful tool in predicting human argumentative behavior in the real world.
(3) We introduce the heuristics of *Relevance*, which integrates the concept of Bounded Rationality [Gigerenzer and Selten 2002] into Argumentation Theory. We further demonstrate the potential that the Relevance heuristics hold in predicting human argumentative behavior in human discussions.
(4) We show that the combination of ML prediction and the Relevance heuristics significantly outperforms 8 other argument provision agents. These agents propose arguments based on Argumentation Theory, heuristics, predicted arguments

[1]All experiments were authorized by the corresponding institutional review board.

without the heuristics or predicted arguments using a TL method. The agent combining ML prediction with the Relevance heuristics, entitled PRH, achieved significantly superior results on both of the axes that we examined, that is, people's subjective satisfaction from the agent and people's use of its suggested arguments.

The remainder of the article is organized as follows. In Section 2, we survey related work and provide an overview of the models used in the scope of this study. In Section 3, we present a comprehensive empirical evaluation of the four suggested prediction methods of human argumentative behavior. In Section 4, we describe the design and evaluation of 9 novel argument provision agents. In Section 5, we present our conclusions and list recommendations for future work in this area.

This article extends our previous reports [Rosenfeld and Kraus 2014, 2015] in several ways. First, by adding an additional 600 study participants, this article enhances the credibility and validity of our previously reported results. Second, it provides a comprehensive analysis of 4 prediction methods and 9 argument provision agents compared to 2 prediction methods and 6 argument provision agents examined before. Finally, this article covers the investigation of the TL approach, which has not yet been investigated in this argumentation context.

2. RELATED WORK AND BACKGROUND

Since the time of Aristotle, there have been many frameworks for argumentative behavior that have been proposed by philosophers and mathematicians alike. To date, argumentation researchers have extensively studied the concept of a "good" argument and have proposed many models explaining how to identify these arguments [Walton 2009]. Most of the state-of-the-art models, which are known as Argumentation Theory, rely on some fundamental notions from Dung [1995] and expand them in some way. These include the Bipolar Argumentation Framework (BAF) [Cayrol and Lagasquie-Schiex 2005a], the Value Argumentation Framework (VAF) [Bench-Capon 2003] and the Weighted Argumentation Framework (WAF) [Dunne et al. 2011], to name a few (see Brewka et al. [2014] for a recent review). This haystack of theories is based on similar principles and ideas. It is common in Argumentation Theory to define some argumentation framework – a formalized structure in which statements (arguments) can attack or support each other. Using different reasoning rules (semantics) it is possible to build sets of justified arguments (arguments that should be considered correct to some extent), and solve the inherent conflicts. The basic notions suggested in Dung [1995] are available in Appendix A.

Throughout this work, we use the BAF modeling proposed in Cayrol and Lagasquie-Schiex [2005a], to which we will refer as the "argumentation framework".

Definition 2.1. A Bipolar Argumentation Framework (BAF) $< A, R, S >$ consists of a finite set A called arguments and two binary relations on A called attack (R) and support (S).

The BAF modeling assumes 2 types of possible interactions between arguments; attack and support. That is, if argument $a \in A$ relates to argument $b \in A$, then aRb or aSb holds, respective of the relation type. It is argued that the use of both support and attack relations in argumentation frameworks is essential to represent realistic knowledge see Amgoud et al. [2008] for a survey).

The argumentation framework can also be represented as a directed graph with 2 types of edges[2]. A is the set of vertices, and R, S, are the sets of directed edges representing attack and support relations.

[2]Sometimes called a bipolar interaction graph.

In order to be able to perform reasoning about an argumentation framework, it is necessary to define reasoning rules, called *semantics*. Dung [1995] has defined several semantics that have been modified to fit the BAF modeling [Amgoud et al. 2008]. In this work, we examine the 3 classical semantics proposed by Dung: Preferred, Grounded and Stable (see Appendix A). Using the above semantics, a reasoner can identify sets of arguments, called *extensions*, which hold the special properties requested by the semantic. An argument that is a member of some extension is considered acceptable or justified (to some extent).

Given a BAF $<A, R, S>$, a discussion d is a finite sequence of arguments $<a_1, a_2, \ldots, a_n>$, where $a_i \in A$. A discussion d can be split into 2 argument sets A_1 and A_2, where $A_i = \{a_j \| a_j \in d \wedge a_j$was presented by study participant i$\}$. That is, every conversation can be seen as 2 argument sets A_1 and A_2, one per participant in the conversation. When examining whether a set of arguments A_i is a part of some extension, one can consider the calculated extensions of the entire BAF. However, one can also consider the extension derived only on the basis of arguments in d, that is, one can calculate the extension of a BAF consisting only of A_1 and A_2's arguments. We denote this BAF as the *restricted* argumentation framework induced by $A_1 \cup A_2$.

Definition 2.2. Let $W =< A, R, S >$ be a BAF, and $A' \subseteq A$ be an argument set. The *restricted BAF* induced by A' is defined as $W{\downarrow_{A'}} =< A', R \cap A' \times A', S \cap A' \times A' >$.

The Preferred, Grounded, and Stable semantics coincide on a single unique extension if the argumentation framework is *well founded*.

Definition 2.3. Let W be a BAF. W is *well founded* if there exists no infinite sequence $a_0, a_1, \ldots, a_n, \ldots$ such that $\forall i.(a_i, a_{i+1}) \in R \cup S$.

The understanding of the connections between human reasoning and Argumentation Theory is a key requirement for deploying Argumentation-based software and agents in practical applications. To date, very little investigation has been conducted regarding how well the proposed models and semantics describe human reasoning. To the best of our knowledge, only two other papers directly address this topic; Rahwan et al. [2010] studied the reinstatement argumentative principle in questionnaire-based experiments and Cerutti et al. [2014] examined humans' ability to comprehend formal arguments. These works did not examine the possibility of Argumentation Theory predicting people's argumentative behavior nor did they try to use their insights to generate advice or recommendation for a human user. Baroni et al. [2015] recently provided a conceptual analysis and discussion on the incompleteness and undecidedness in Argumentation Theory, which are also common in human reasoning. However, they did not evaluate or consider human argumentative behavior.

Computer-supported argumentation systems have received much attention in the last 20 years [Scheuer et al. 2010]. Such systems are prominent in law, education, formal reasoning, and collaborative discussions. These systems implement a *normative* approach to argumentation, that is, how argumentation should work from a logical standard. For example, ArgTrust [Parsons et al. 2013] provides an argumentation-based software that provides users the means to handle argumentative situations in a coherent and valid manner. MIT's delibrium [Klein 2011] provides an interactive web-based system to allow multiple, distant users to engage in a discussion in a logical manner. To the best of our knowledge, no argumentation-based system deploys a *descriptive* approach, that is, accounts for how argumentation actually works in human reasoning. In this work, we examine the implementation of both the normative model (i.e., Argumentation Theory) and descriptive models (i.e., Heuristics, ML, and TL), first to predict human argumentative behavior, then to provide arguments during a discussion.

Other studies have addressed agents' strategical considerations in argumentative environments involving people. Dsouza et al. [2013] investigate which information an agent should reveal during a deliberation with people. Others have developed policies for the generation of offers in human–agent negotiations [Rosenfeld et al. 2014], or the generation of arguments in automated sales-clerk interactions with human customers [Hiraoka et al. 2014]. None of these works are in the context of Argumentation Theory, nor have any of these researchers considered providing arguments or recommendations to the human user.

In Section 3.3, we introduce the heuristics of *Relevance*, which relies on the concept of *proximity* between arguments. Booth et al. [2012] have provided a formal analysis of proximity between different evaluations of arguments. Both notions use distance and proximity measurements to derive insights mainly based on the argumentation framework's structure. However, our concept of proximity is completely independent of the one presented in Booth et al. [2012], as we do not consider the evaluation of arguments in our distance measurements.

Conversational Agents[3] (CAs) have been developed over the years to converse with humans and provide information or assistance to the users' satisfaction [Cassell 2000; Berg 2015]. In the CA framework, a human user directly interacts with the CA and explicitly conveys one's wishes. The CA can also ask the user questions that will help it understand the user's goals or requests. In this work, we deal with a different setting in which 2 people converse while the advising agent cannot take an active part in the conversation. Specifically, the agent can only observe the dialog, and its sole means of communication with its user is by providing arguments for the user to implement.

At first glance, our proposed approach can be viewed as part of the Case-Based Reasoning (CBR) approach [Watson 1999]. In CBR, a reasoner *retrieves* relevant past cases from memory, *reuses and revises* the solutions from these previous cases to generate a solution for the target problem, and *retains* the generated solution for future use. In our setting, the advising agent can use rules learned by ML algorithms to find suitable arguments to propose (similar to the retrieval phase). These arguments can be used by the agent (depending on its policy) to generate an argument list to propose to its user (similar to the reusing and revising phase), and the user's choices can be stored for future use (similar to the retaining phase). However, our proposed approach is quite different from CBR. First, our approach does not require past *advising cases* to derive an advising policy, but needs only past dialogs on the topic. Note that the CBR approach necessitates the identification of a great deal of tasks that were already solved successfully in the past. In our setting, the task is to provide beneficial advice during an argumentation discussion. Attaining an abundant collection of successful advising cases can be extremely expensive. Furthermore, our approach forms its generalizations of the given dialogs by identifying commonalities between the training examples in an offline fashion before the actual dialog takes place. In the CBR approach, the commonalities between the retrieved examples and the target problem are carried out in an online setting that can pose difficulty in fast-changing dialogs.

The two argument provision approaches that we examine in this article hold different rational-psychological explanations for why people would benefit from the suggested arguments. First, people search for *validation* for their existing opinions and beliefs [Linehan 1997]. Thus, receiving consonant (supportive) arguments for their views from an intelligent agent can help validate the person's beliefs. Second, Rational Choice Theory [Coleman and Fararo 1992] suggests that when an individual considers an action (e.g., argument to use), the individual needs to weigh all information that can or will affect that argument. An agent can help a user in this task by revealing additional

[3]Also known as dialog systems.

arguments, that is, arguments of which the user was unaware, or by assisting the user in weighing the different arguments in an analytic manner.

It is common in literature to distinguish between different types of argumentation structures [Walton et al. 2010]. Throughout this work, we focus on *deliberations*, in which the discussion process is aimed at exchanging opinions, beliefs, and information and is trying to reach some consensus on a controversial topic. In the scope of this study, the agent's goal is to provide arguments that its user will find satisfactory. Future work will expand the suggested methodology to account for more complex argumentation structures such as persuasion and negotiation, in which an argument provision agent will be required to assist the user in constructing more convincing and compelling arguments.

The development of automated argumentation-based agents, such as the ones presented in this study, necessitates the assumption that natural-language statements can be automatically mapped into arguments. Despite recent advancements in Natural-Language Processing (NLP) and Information Retrieval (IR) and their studied connections to argumentation [Modgil et al. 2013; Cabrio et al. 2014; Moens 2014], this assumption is not completely met by existing automated tools. Thus, throughout this work, we use a human expert annotator whom we hired as a research assistant. We hope that this work will inspire other researchers in NLP and IR to tackle the problem of automatically mapping natural-language statements into arguments as well as other open problems of great importance in argumentation-based systems. These include the automatic extraction of arguments from texts [Fan et al. 2012] and the automatic identification of relations between natural-language arguments [Slonim et al. 2014].

3. PREDICTING PEOPLE'S ARGUMENTATIVE BEHAVIOR

In this section, we investigate the predictive abilities of the four proposed prediction methods: Argumentation Theory, Relevance Heuristics, ML and TL. To that aim, we first collect extensive data in several experimental settings, varying in complexity, in which human study participants were asked to use arguments. We provide a full description of three experimental settings, followed by an analysis of the gathered data using three of the proposed prediction methods: Argumentation Theory, Relevance Heuristics and ML. Then, we describe an additional experiment (Experiment 4) followed by an analysis of the gathered data using the TL prediction method.

The suggested prediction methods examined in this section were also used in the design of the 9 argument provision agents developed in the scope of this study (see Section 4).

3.1. Experimental Design

3.1.1. Experiment 1—Questionnaire-Based Argumentation. Two groups took part in this experiment. The first group consisted of 64 US residents, all of whom work for Amazon Mechanical Turk (AMT). These study participants, denoted the US-Group, ranged in age from 19 to 69 (mean = 38, s.d. = 13.7), with 38 females and 28 males. The second group consisted of 78 Israeli Computer Science bachelor-degree students, denoted the IL-Group, ranging in age from 18 to 37 (mean = 25, s.d. = 3.7), with 27 females and 51 males. The study participants were presented with 6 fictional scenarios based on scenarios from Walton [2005], Arvapally and Liu [2012], Cayrol and Lagasquie-Schiex [2005a], Amgoud et al. [2008], and Parsons et al. [2013]. The scenarios are available in Appendix B. Small changes were made in the original formulation of the scenarios in order to keep the argumentation frameworks small (6 arguments) and simply phrased, yet the scenarios were kept as close as possible to the original. Each scenario was presented as a short conversation between the 2 deliberants, and the study participant

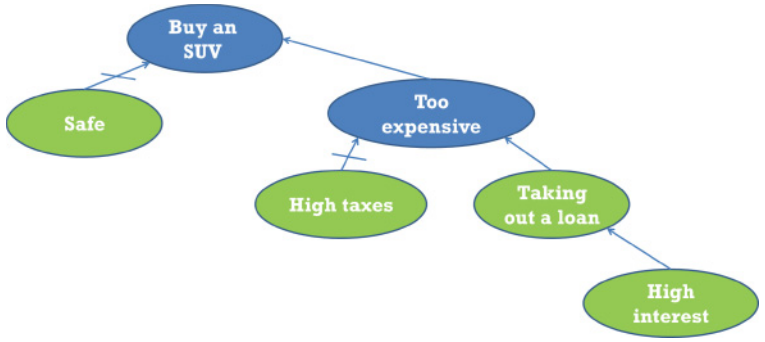

Fig. 1. BAF: nodes are arguments. Arrows indicate attacks and arrows with diagonal lines indicate support.

had to choose which of the 4 possible arguments to use next if the participant was one of the deliberants in the conversation. The following example is one of the 6 scenarios we presented to the study participants:

Example 3.1. A couple is discussing whether or not to buy an SUV.
Spouse number 1 (S_1): "We should buy an SUV; it's the right choice for us".
Spouse number 2 (S_2): "But we can't afford an SUV, it's too expensive".
The study participant was then asked to put oneself in S_1's place and choose the next argument to use in the deliberation[4].
A. Good car loan programs are available from a bank.
B. The interest rates on car loans will be high.
C. SUVs are very safe, and safety is very important to us.
D. There are high taxes on SUVs.

The 6 scenarios were similar in the way in which they were presented: a short conversation of 2 statements and 4 possible arguments from which the study participant was asked to select a preferred argument. However, the argumentative frameworks that they induced were different in order to simulate different argumentative complexity levels. Figure 1 presents a graphical representation of the example. This graphical representation was *not* presented to the study participants.

3.1.2. Experiment 2—Real Phone Conversations (Secondary Data). For this experiment, we used real argumentative conversations from the Penn Treebank Corpus (1995) [Marcus et al. 1993]. Thus, Experiment 2 is an analysis of *secondary data*. The Penn Treebank Corpus includes hundreds of transcribed telephone conversations on controversial topics such as "Should Capital Punishment be implemented?" and "Should trial sentencing be decided by a judge or jury?", on which we chose to focus. We reviewed all 33 deliberations on "Capital Punishment" and 31 deliberations on "Trial by Jury" in order to map the presented utterances into arguments. This process required the clearing of irrelevant sentences (i.e., greetings, unrelated talk, and so forth) and the construction of an argumentation framework comprising all of the arguments presented in the conversations.

The annotation process was performed manually by a human expert; first, the annotator read 5 conversations that were selected at random from the corpus to get an initial idea of which arguments people use. Then, the annotator went over all conversations on a topic, one at a time, and classified each statement presented in the conversation into one of the following categories: (1) Arguments, (2) Opinions, and (3) Other. The

[4]The options were shuffled to avoid biases.

Arguments class consists of all relevant arguments that were presented on a topic in the corpus. The Opinions class consists of all statements that reflect the speaker's opinion on the discussed topic such as "I'm pro X" or "I oppose the Y idea". All other statements, such as greetings or unrelated talk, were classified to the Other class. Once all statements were classified, the annotator went over all presented arguments and constructed an argumentation framework by identifying support and attacks between arguments. In cases in which the human expert was unsure of one's annotation or the relation between arguments, a second human expert was asked to provide a final decision. Then, the presented opinions were mapped into "pro" and "con" subclasses. Finally, the annotator translated each conversation into a series of arguments (with respect to the constructed framework) and opinions (according to their subclassification to "pro" and "con"), providing us with 33 sequences on "Capital Punishment" and 31 sequences on "Trial by Jury".

The shortest sequence is of size 4 and the longest is of size 15 (a mean of 7).

Unfortunately, the participants' demographic data is unavailable.

3.1.3. Experiment 3—Semistructured Online Chats. For Experiment 3, we developed a special chat environment. In our chat environment, deliberants communicate only by using arguments from the predefined argument list. We chose the topic of *"Would you get an influenza vaccination this winter?"* and constructed a predefined argument list consisting of Pro (20) and Con (20) arguments. These arguments were extracted from debate sites[5] and medical columns[6]. In order to bolster the natural flow of the dialog, study participants were also provided with a "bank of discourse statements", comprised of a set of discourse markers and short statements such as "I agree", "I think that", "However", and others. The "bank of discourse statements" allows users to express themselves more naturally. Unlike Experiment 2, the collected chats in Experiment 3 are not considered to be natural chats but rather semistructured chats.

We recruited 144 Israeli college students to participate in this experiment, ranging in age from 20 to 36 (mean of 27), with 64 females and 80 males. Participants were coupled at random and were asked to deliberate over the question *"Would you get an influenza vaccination this winter?"* for a minimum of 5min. Deliberations ranged in length from 5 arguments to 30 (mean 14). Each deliberation ended when one of the deliberants chose to exit the chat environment.

3.2. Analysis using Argumentation Theory to Predict People's Arguments

People's decision-making processes are known to be affected by a multitude of social and psychological factors, and they often do not maximize expected utilities or use equilibrium strategies [Camerer 2003; Rosenfeld et al. 2012]. The question that we investigate in this section is whether, in the context of argumentative discussions, people would choose justified arguments according to some semantic choice. That is, would Argumentation Theory provide predictive tools to predict human argumentative behavior. Appendix A provides an overview of the concepts used in the following analysis.

Recall that each of the tested scenarios in Experiment 1 and each of the tested domains in Experiments 2 and 3 was mapped into an argumentation framework as described in Section 3.1. In order to analyze the data, we first calculated the Grounded, Preferred, and Stable extension for each of the resulting argumentation frameworks.

In Experiment 1, each of the 6 tested scenarios (Appendix B) was mapped into a *well founded* (Definition 2.3) BAF. As such, the three tested semantics coincide—only one

[5]such as http://www.debate.org/, http://idebate.org/.

[6]such as http://healthresearchfunding.org/pros-cons-flu-shots/.

Table I. Semantics' Accuracy in Describing Study Participants' Arguments with Respect to the Corresponding Original (Orig.) Argumentation Framework and the Restricted (Rest.) Argumentation Framework

	Preferred		Grounded		Stable	
	orig.	rest.	orig.	rest.	orig.	rest.
"Capital Punishment"	20%	47%	25%	40%	21%	33%
"Trial by Jury"	36%	50%	34%	59%	30%	40%
"Influenza Vaccination"	38%	44%	38%	49%	30%	41%

extension is Grounded, Stable, and Preferred. On average, in the 6 tested scenarios of Experiment 1, a justified argument was selected only 67.3% of the time (under the tested semantics, i.e., Preferred, Grounded, and Stable). Moreover, only 8% of the participants chose justified arguments in all 6 scenarios. Note that, in all 6 scenarios, more than a single justified argument was available. If we were to predict one of the justified arguments in each of the 6 examined scenarios, the expected accuracy of our prediction model would be 32%. This result is only slightly better than random selection (selecting 1 argument out of 4–25%) and worse than using majority prediction (43%), which is predicting the argument that the majority of participants chose in the examined scenario.

For example, in the SUV scenario (see Figure 1), most people (72%) chose the "Taking out a loan" or "High taxes" arguments that directly relate to the last argument presented in the discussion. The "Taking out a loan" argument was the most popular one (37%). However, the "Taking out a loan" argument is supposed to be considered weaker than the other 3 possible arguments as it is not a part of the Grounded, Preferred, and Stable extension. The other 3 arguments should be considered justified (as they are unattacked and part of the Grounded, Preferred, and Stable extension), whereas "Taking out a loan" is attacked by a justified argument ("High interest") and is not part of the Grounded, Preferred, and Stable extension. Such phenomena were encountered in all other scenarios as well.

In Experiments 2 and 3, the resulting 3 argumentation frameworks on "Capital Punishment", "Trial by Jury" (Experiment 2), and "Influenza Vaccination" (Experiment 3) were mapped into BAFs. In the resulting BAFs, the Grounded, Preferred, and Stable semantics do not coincide on a single extension. The argumentation framework for "Capital Punishment" consisted of 30 arguments, the argumentation framework for "Trial by Jury" consisted of 20 arguments, and the argumentation framework for "Influenza Vaccination" consisted of 40 arguments.

Experiments 2 and 3 differ from Experiment 1 in that they are comprised of discussions between study participants. Each discussion was split into 2 argument sets, A_1 and A_2, comprising the arguments used by each of the participating parties of the discussion. Each discussion was analyzed twice: First, each of the argument sets comprising the discussion was examined with respect to the argumentation framework consisting of all arguments on the topic. Second, each of the argument sets was examined with respect to the *restricted* argumentation framework induced by the argument sets of the discussion (see Definition 2.2).

On average, across the 3 domains, when examining the original framework, less than 35% of A_is used by the study participants were a part of some extension, with Preferred, Grounded, and Stable performing very similarly (34%, 35%, 27%). When considering the restricted argumentation framework, 47%, 50%, and 39% of the deliberants used A_is that were part of some extension prescribed by Preferred, Grounded, and Stable (respectively) under the restricted framework. See Table I for a summary.

More surprising was the fact that even the mildest requirement suggested by Argumentation Theory was not upheld by many study participants. We tested the arguments

that the study participants used in the context of *Conflict-Freedom* (CF). In particular, we checked whether the study participants refrained from using contradictory arguments in different stages of the discussion. CF is probably the weakest requirement from a set of arguments. We anticipated that all study participants would adhere to this requirement, yet only 78% of the deliberants used a conflict-free argument set. Namely, 22% of the deliberants used at least 2 conflicting arguments, that is, one argument that contradicts the other, during their discussions.

3.3. Analysis Using Relevance Heuristics

At a given point in a deliberation, not all arguments are necessarily relevant to the context of the deliberation (i.e., the current focus of the deliberation)[7]. For instance, the argument "Safe" in our example (Figure 1) seems to be irrelevant to the current focus of the discussion, since the focus is on economic concerns. First, in order to identify "relevant" arguments, we propose several distance measurements that heavily rely on the current state of the deliberation and the structure of the argumentation framework. These distance measurements will help us investigate how the *proximity* between arguments, as portrayed by the edge-distance in the argumentation framework, truly affects the course of a deliberation.

We defined 15 relevance measurements, each of which captures different aspects of proximity. In the definitions, a denotes a possible argument, a_l is the last argument presented in the discussion, a_c is the "closest" argument to a that was previously presented in the dialog (using edge-distance metric) and Ω denotes a designated argument that represents the discussed issue (in Figure 1, it is whether or not to "Buy an SUV"). The relevance measurements of a possible argument a can be summed up in the following 4 points:

(1) Minimum un/directed paths' length from a to a_l.
(2) Minimum un/directed paths' length from a to a_c.
(3) Minimum directed paths' length from a to Ω.
(4) Minimum of all/some of these features.

When omitting redundant calculations in the fourth criterion (e.g., the minimum of the shortest directed and undirected paths from a to a_l), 15 distinct measurements remain, denoted $d^1, \ldots, d^{15}$.

In the SUV scenario, S_2's argument is considered to be a_l, and Ω is S_1's argument ("Buy an SUV"). When we consider a as "Safe", its distance to a_c or Ω (in this case, they are the same) is 1, while its directed distance to a_l is undefined and the undirected distance is 3. If a is "Taking out a loan", then its distance to a_l and a_c is 1, whereas its distance to Ω is 2.

Note that as arguments are presented during a discussion, some of the arguments' relevance heuristic values may change.

Given the current state of deliberation, each of the proposed distance metrics induces a partial order ranking over all arguments in the argumentation framework. To perform a prediction using d^i, we compute and predict $\text{argmin}_a\, d^i(a, a_l, a_c, \Omega)$, where ties are broken randomly. That is, after an argument has been put forward, we compute $d^i(a, a_l, a_c, \Omega)$ for every argument $a \neq a_l$ given a_l, a_c, and Ω (as observed in the partial conversation and the argumentation framework). Then, we rank the arguments accordingly and predict the top-ranking argument. The process was repeated for all 15 proposed measurements, $d^1, \ldots, d^{15}$, resulting in 15 predictions for every argument presented in every discussion.

[7]Not to be confused with the concept introduced in Liao and Huang [2013], which states that it might not be necessary to discover the status of all arguments in order to evaluate a specific argument/set of arguments.

Table II. Relevance Prediction Accuracy Across Experiments 1, 2, and 3

Distance Measurement	Experiment 1	Experiment 2	Experiment 3
Directed paths' length from a to a_l (d_1)	35%	17%	15%
Directed paths' length from a to Ω (d_2)	40%	15%	17%
Minimum$\{d^1, d^2\}$ (d_3)	37%	14%	16%
Others (average)	25%	5%	5%

In all 3 experiments, the directed paths' length from a to a_l (denoted d^1), the directed paths' length from a to Ω (denoted d^2) and their combination (the minimum between d^1 and d^2, denoted d^3), yield the highest average prediction accuracy averaging 38%, 15%, and 16% in Experiments 1, 2, and 3, respectively. No statistically significant difference was found between d^1, d^2, and d^3. The other 12 measurements performed significantly worse, averaging 25%, 5%, and 5% in Experiments 1, 2, and 3, respectively. See Table II for a summary.

Specifically, in the SUV scenario, the prediction using d_1 would predict either "Taking out a loan" or "High taxes" (randomly, as they both have the minimal distance value of 1). The prediction using d_2 would predict "Safe", and the prediction using d_3 would predict "Taking out a loan", "High taxes", or "Safe" (randomly).

In Experiments 2 and 3, there are many more arguments to consider in the prediction compared to Experiment 1. Naturally, the prediction accuracy declined. However, we can use the ranking induced by d^i and predict more than 1 argument: that is, we can predict the top k ranked arguments with regard to their relevance values. When predicting the top 3 ranking arguments, d^1, d^2, and d^3 average 57% prediction accuracy across Experiments 2 and 3. Again, no statistically significant difference was found between them. The other 12 measurements performed significantly worse, averaging 35% across Experiments 2 and 3.

3.4. Analysis Using Machine Learning to Predict People's Arguments

The use of ML in predicting human behavior has shown much promise in developing automated human-interacting agents; a few recent examples are Rosenfeld [2015], Rosenfeld et al. [2015a, 2015b], Azaria et al. [2015], and Peled et al. [2013]. However, the use of ML for the prediction of human argumentative behavior has not been investigated thus far.

The task of predicting human argumentative choices in a discussion can be defined as a multiclass prediction problem. We seek to construct a prediction function $P : \chi \rightarrow A$, where $\vec{x} \in \chi$ is a feature vector (sampled from the feature space χ) representing both the deliberant characteristics and the discussion's state, and $a \in A$ is an argument that is predicted to be presented next in a deliberation given $\vec{x}$.

For this purpose, we first suggest the characterization of *arguments* in the argumentation framework. For every argument a in the argumentation framework, we suggest a calculation of a measurement vector m_a. m_a describes a in the context in which it is judged (the context in which a reasoner evaluates the argument). That is, in a given state of the discussion, m_a represents the characteristics of a with respect to the current state of the discussion. Hence, m_a might require an update after each presented argument in the discussion by either of the deliberants. We divide m_a into 3 categories; *Justification* measurements, *Relevance Heuristic* values and *Confirmation Factor*.

Given the characterization of *arguments*, we then present the procedure by which we compute $\vec{x}$—the *feature vector* used in our prediction model P. $\vec{x}$ relies on the arguments presented in the discussion and their characteristics. We divide these features into 2 categories: the *Deliberant features* and the *Deliberation context features*, together

comprising $\vec{x}$. That is, $\vec{x}$ represents both the deliberant and the deliberation using a vector of feature values.

3.4.1. The Characterization of Arguments.

Justification: There have been a number of proposals for more sophisticated analysis of argumentation frameworks. These proposals mainly consider the relative strength of the arguments or the authority of the party who presented the argument (e.g., Pazienza et al. [2015]). One commonly used proposal is the gradual valuation [Cayrol and Lagasquie-Schiex 2005b] in BAFs. The idea is to evaluate the *relative strength* of argument a using some aggregation function that conciliates between its attacking arguments' strength and its supporting arguments' strength. This recursive calculation allows us to aggregate the number of supporters and attackers, as well as their strength, through the argumentation framework and reach a strength value in a (possibly bounded) interval (e.g., [−1,1]) for each argument. Note that, given an argumentation framework $< A, R, S >$ and an argument $a \in A$, the identification of a's attackers ($R(a)$) and a's supporters ($S(a)$) is straightforward. The technical definition of the gradual valuation functions, as well as its most popular instantiation that is used in this article, is provided in Appendix D. We denoted this gradual valuation function as "Cayrol's calculation".

The strength value returned by the valuation function represents the deliberant's ability to support that argument and defend it against potential attacks. The higher the strength level, the easier it is to support and defend the argument, and the harder it is to attack it.

In our SUV example in Figure 1, Cayrol's calculation J (provided in detail in Appendix D) provides J("Safe") = J("High Taxes") = J("High interest") = 0 and J("Taking out a loan") = −0.33. The intuition behind this example is that the "Safe", "High Taxes", and "High interest" arguments cannot be attacked or supported, thus have a strength level of 0. The strength value of 0 means that a logical reasoner is capable of defending the argument to the same extent that one is capable of attacking it. On the other hand, the "Taking out a loan" argument is considered weaker, as it is attacked by another argument.

In the empirical study [Bonnefon et al. 2008], the authors examined the problem of predicting people's choice between 2 options (e.g., going to movie A or movie B) based on supportive and attacking arguments (pieces of information) relevant to the options at hand. The main and most relevant insight from their work is that a favorable prediction method should not ignore the number of supporting and attacking arguments when predicting people's choices. In order to integrate this insight into the arguments' characteristics, we first identify the relation between every pair of arguments using the four General Argumentation Heuristic Rules [Klein et al. 2003] (see Definition 3.2). That is, we calculate the relation between every pair of arguments in the argumentation framework using simple heuristics and a simple graph traversal.

Definition 3.2. Let $a, b, c \in A$ be arguments in an argumentation framework $< A, R, S >$. a is said to be a *direct* supporter (attacker) of b if aSb (aRb) holds.

The General Argumentation Heuristic Rules [Klein et al. 2003] are defined as follows:

(1) If a supports b and b supports c, then a (indirectly) supports c.
(2) If a attacks b and b supports c, then a (indirectly) attacks c.
(3) If a supports b and b attacks c, then a (indirectly) attacks c.
(4) If a attacks b and b attacks c, then a (indirectly) supports c.

In our SUV example, the "Safe" argument is a *direct* supporter of the "Buy an SUV" argument and "Too expensive" is a *direct* attacker of it. The "Taking out a loan"

argument is a direct attacker of the "Too expensive" argument and as such acts as an *indirect* supporter of the "Buy an SUV" argument according to rule 4 of the General Argumentation Heuristic Rules. Following rule 3, "High Taxes" is considered an *indirect* attacker of the "Buy an SUV" argument (as it directly supports the "Too expensive" arguments), and following rule 2, the "High interest" argument is considered an *indirect* attacker of the "SUV" argument (as it directly attacks the "Taking out a loan" argument).

For each argument a, we calculate the number of supporters (direct and indirect), denoted $|Sup(a)|$ and the number of attackers (direct and indirect), denoted $Att(a)$. Then, we compute the supporters' portion $\frac{|Sup(a)|}{|Att(a)|+|Sup(a)|}$ as a member of m_a (the argument's characteristics). If $|Att(a)|+|Sup(a)| = 0$, then we define the support portion as 0.5. The proposed "Support portion" captures another aspect of the argument's strength, providing each argument a strength value in the [0,1] interval. In our SUV example, the "'Safe", "High Taxes", and "High interest" arguments have 0.5 support portion values and the "Taking out a loan" argument has a 0 support portion value.

It is important to state in this context that the suggested justification measurements rely solely on the argumentation framework and, as such, require only a single, offline calculation of their values for each argumentation framework (regardless of the current deliberation).

Relevance Heuristics: The Relevance values of an argument a capture the proximity of a to the a_l, a_c, and Ω arguments. As Relevance values, we used the metrics d^1, d^2, and d^3 defined in Section 3.3, as they provided the highest average prediction accuracy on the gathered data in Experiments 1, 2, and 3. Given the current state of the deliberation, we compute the d^1, d^2, and d^3 values for every argument a in the argumentation framework as a member of each argument's characterization m_a. Unlike the Justification values of an argument, the Relevance values may change as more arguments are presented in the discussion.

Confirmation Factor: Confirmation bias is a phenomenon in psychology in which people have been shown to actively seek out and assign more weight to evidence that confirms their beliefs, and ignore or underweigh evidence that could disprove their beliefs [Nickerson 1998]. In argumentative situations, people may present a confirmation bias by selectively considering arguments that reinforce their expectations and disregard arguments that support alternative possibilities or attack their own.

We use these insights and assign each argument a a value named *confirmation factor* in its characterization m_a. This value depends on the effect that a has on the deliberant's previously stated arguments. The confirmation factor can be *positive* if a supports (directly or indirectly) previously presented arguments by the deliberant or it can be *negative* if a attacks (directly or indirectly) previously presented arguments by the deliberant. If the relation is ambiguous (both positive and negative) or unknown (the argument does not affect previous arguments presented by the deliberant), then a has a *neutral* confirmation factor.

Given the current state of the deliberation, we compute the confirmation factor for every argument a in the argumentation framework and save it as a member of m_a. This is carried out by iterating over the arguments previously used by the deliberant and using a's relation with them as defined by the General Argumentation Heuristic Rules (Definition 3.2) to determine a's confirmation factor.

Similar to the Relevance values of an argument, the confirmation factor of an argument may also change as more arguments are presented in the discussion.

In our SUV example, the "Safe" and "Taking out a loan" arguments have a positive confirmation factor, as both support S_1's previous argument (the "SUV" argument). Symmetrically, the "High Taxes" and "High interest" arguments have a negative confirmation factor.

3.4.2. Feature Vectors. Earlier, we described the characterization of each argument a that we denoted as m_a. In order to perform a prediction given a partial discussion, we use the presented arguments in that discussion to compute a feature vector $\vec{x} \in \chi$. These features represent the current state of the deliberation and the deliberant's preferences in arguments, denoted the *deliberation context features* and the *deliberant features*, respectively. Next, we describe how we computed these features.

Deliberation context features: During a deliberation, we account for the last 2 arguments presented by each of the deliberants and indicate which of the deliberants presented the last argument. These 5 features, that is, the last 2 arguments presented by each deliberant (represented by their labels) and a binary feature representing which deliberant presented the last argument in the discussion, are denoted as the *deliberation context features*. These features are recorded as part of $\vec{x}$—the features used by the ML model.

Features of the Deliberant: In order to capture the deliberant's preferences in arguments, we aggregated the characteristics of the deliberant's presented arguments in the discussion. Namely, we analyzed the arguments that the deliberant presented and calculated the average justification value (both the average of J values and the support portion values), the average relevance heuristic values and the percentage of times that a confirmatory argument was used (of the number of times at least one was available).

In addition, we hold a *proneness feature* in the [0,1] interval, which indicates the person's inclination toward accepting a specific position on the discussed issue. For example, in a deliberation on "Capital punishment" a value of 1 means that the deliberant supports capital punishment and 0 means that the deliberant opposes it. The higher the proneness feature value, the stronger the deliberant's inclination to agree with the discussed issue. This feature stems from the Dissonance Theory [Festinger 1962], which suggests that once committed to an alternative (knowingly or unknowingly), people prefer supportive (consonant) arguments compared to opposing (dissonant) arguments to avoid or reduce post decision-making conflicts. In order to calculate the proneness feature, we distinguished between 2 cases. In cases in which the deliberant explicitly expressed an opinion (e.g., "I'm pro Capital Punishment"), the proneness value is simply 1 or 0 (depending on the opinion expressed). In cases in which the deliberant's opinion was not explicitly declared, we assessed the deliberant's position using the deliberant's previously stated arguments. We calculated this estimation using the portion of supportive arguments to the discussed issue that the deliberant used during the conversation. Namely, using only arguments supportive of the discussed issue is the same as explicitly stating your opinion.

These values, denoted as the *deliberant features*, are part of $\vec{x}$—the input to the ML prediction model.

3.4.3. Analysis of Experiment 1. Each study participant provided 6 argumentative selections, one per each presented scenario. Given a learning period of k scenarios, where $k = 1, 2, \ldots, 5$, we took $6 - k$ scenarios from each study participant's answer set and used the remaining scenarios as training data. For example, for $k = 5$, we removed 1 scenario at a time from all study participants' selections and used the 5 remaining scenarios for training.

In order to predict the argumentative choice made by study participant i in scenario j given a learning period k, we first calculated the deliberant's features according to k scenarios (not including scenario j) for all study participants other than i. Then, we labeled each calculated deliberant's features with the actual argument selection made by each study participant in scenario j. The resulting vectors and labels are used to train the prediction model.

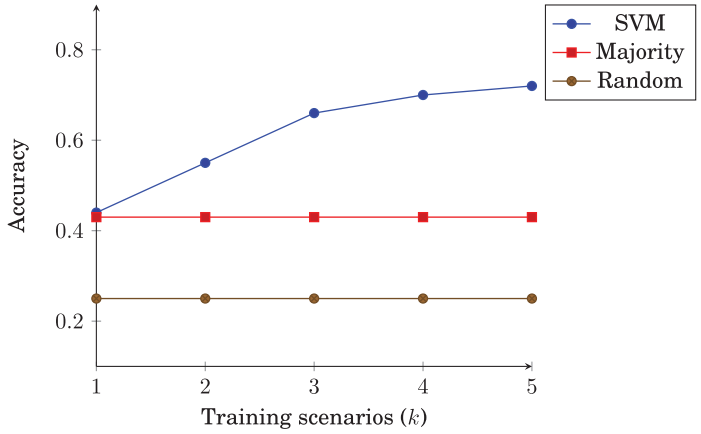

Fig. 2. The SVM's learning curve in Experiment 1. The more argumentative choices available for training the SVM, the better its prediction accuracy on unseen scenarios.

We used 3 ML algorithms to test the prediction accuracy of our methodology: the Support Vector Machine (SVM), Decision Tree Learning (DTL), and the Multilayered Neural Network (MLNN). We trained and tested our model on the US-Group and the IL-Group separately (see the groups' descriptions in Section 3.1.1). For both groups, SVM was found to be the most accurate learning model of our observed data, as it provided 72% and 78% accuracy in predicting the study participant's 6th selection when learning from the first 5 selections (US-Group and IL-Group, respectively). DTL and MLNN both yielded less than 68% accuracy for both groups. For the US-Group, as the learning period (k) increased from 1 to 5, the SVM's accuracy increased from 42% to 72%. That is, the more observations the prediction model had on the study participants' selections, the higher its prediction accuracy. Random selection naturally provides 25% (in every scenario, the study participant was requested to choose 1 of 4 suggested arguments), and predicting the majority's selection (predicting the most popular selection among the other study participants) provides 41% accuracy. We remind the reader that the Argumentation Theory prediction and the Relevance Heuristics prediction provided less than 41% accuracy on the data collected in Experiment 1. See the learning curve of the SVM model in Figure 2.

Similar results were obtained for the IL-Group, in which the prediction accuracy ranged from 45% (when $k = 1$) to 78% (when $k = 5$). The accuracy in predicting the IL-Group's selection was slightly higher than the accuracy in predicting the US-Group's selections, probably due to the more homogeneous nature of the IL-Group.

In order to check cultural differences, we examined the use of the US-Group as a training set and the IL-Group as the test set, and vice versa. In the first setting, in which the model was trained using the data from the US-Group and evaluated using the data from the IL-Group, the model achieved 76% accuracy. The second setting, in which the data from the IL-Group was used as training data and the data from the US-Group was used as a test set, demonstrated 69% accuracy.

The features contributing to the prediction (using an entropy measurement [Gray 2011]) were (in the following order of importance):

(1) Relevance (edge-distance from a to a_l)
(2) Cayrol's justification value
(3) Support portion among the influential arguments
(4) Proneness

Most surprising was the fact that the 4 most influential features in the prediction (using an entropy measurement) were the same for both groups, in the exact same order of importance.

3.4.4. Experiments 2 and 3. Each conversation collected on "Capital Punishment", "Trial by Jury" (Experiment 2), and "Influenza Vaccination" (Experiment 3) was then analyzed argument by argument (each argument is considered a step in the deliberation). For each step in the deliberation, we computed the deliberation context features and deliberant features, and labeled the resulting vector with the argument that was presented next in the discussion.

The evaluation of the model was carried out using the 1-left-out methodology. That is, we learned from $n - 1$ conversations and predicted the arguments presented in the different steps of the left-out conversation.

Recall that the features used by the prediction model rely on the previously presented arguments in the discussion. Therefore, in early stages of the discussion, the model may present inadequate predictions. We tested how the prediction quality changes given the steps in which the model is in "learning mode". That is, we tested how the prediction accuracy changes according to the time period in which the model is required to present predictions.

As a baseline model, we used the best model of 8 (simple) statistical models that do not model the deliberant but treat the argument selection process as a stochastic process. The *Bigram model* [Jelinek 1989] of the participant was found to be the best of the 8 models, using perplexity measurements. It outperformed the Trigram model of the participant as well as the Bigram and Trigram statistical models of the other party in the deliberation. It also outperformed the combinations of the these models[8]. Bigram modeling of the deliberant calculates the probability $P(a_2|a_1)$ for every pair of arguments a_1, a_2, that is, the probability that a_2 follows a_1. These probabilities were estimated using a Maximum Likelihood Estimator on the data that we collected. Given a_1 as the last presented argument in the discussion, the model predicts $\text{argmax}_{a_2 \in A} P(a_2|a_1)$.

We again trained and tested the SVM, DTL, and MLNN models and found that the DTL was the leading method, accuracy-wise. Unlike in Experiment 1, in Experiments 2 and 3, there were many more arguments to consider in the prediction. Naturally, the prediction accuracy declined. However, if we use the probability measurements provided by the learning algorithm, we can predict more than 1 argument – that is, we can predict the top k ranked arguments with regard to their probability. On the topic of "Capital Punishment", in Figure 3, we can see how the prediction accuracy increased over the number of predicted arguments (X axis) and the stages from which we began our prediction (the different curves). When predicting the top 3 ranked arguments on the issue of "Capital Punishment", we achieved a prediction accuracy of 71% to 76%, depending on the starting phase of the prediction. Very similar results were obtained for the "Trial by Jury" and "Influenza Vaccination" deliberations as well.

Regardless of the number of predictions, our model's predictions reached better accuracy than the baseline model. To quantify this difference, we used the Mean Reciprocal Rank (MRR) measure [Craswell 2009], which evaluates any process that produces a list of options ordered by their probability of correctness. Our model's MRR was 0.48 for "Capital Punishment", 0.58 for "Trial by Jury", and 0.51 for "Influenza Vaccination", whereas the baseline's MRR was 0.36 for both "Capital Punishment" and "Trial by Jury" and 0.34 for "Influenza Vaccination" (the higher the better).

When comparing the influential attributes found in Experiment 1 to the ones found in Experiments 2 and 3, we can see that the very same features were found to be

[8]All models used a simple smoothing method to avoid the assignment of 0's.

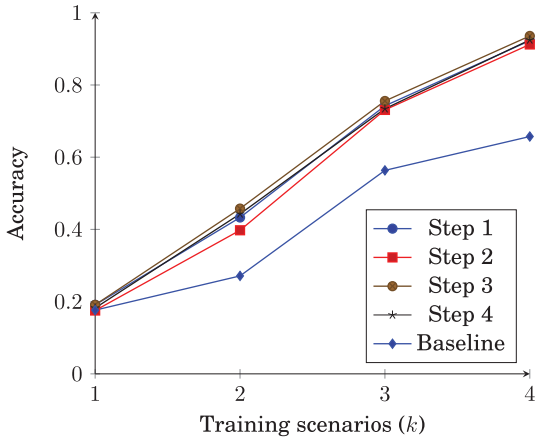

Fig. 3. Prediction curve for Capital Punishment.

influential except for Cayrol's justification calculation, which was ranked much lower in Experiments 2 and 3. The feature that indicated which deliberant presented the last argument (part of the deliberation context features) took its place. Note that this feature was not applicable in Experiment 1. The prediction accuracy skyrocketed to 91.2% ("Capital Punishment"), 88.6% ("Trial by Jury"), and 87.7% ("Influenza Vaccination") in cases in which the deliberant used more than one argument sequentially without interruption. In our study, 100% of the time when a deliberant used more than one argument in a row, the second one was supportive and directly affects the first one. The indication of which deliberant used the last argument was found to be very influential.

3.5. Analysis Using Transfer Learning

In this section, we first describe Experiment 4. Experiment 4 differs from Experiments 1, 2, and 3 in that it investigates a *repeated scenario* in which we have previous argumentative discussions on the same study participant. Then, we analyze Experiment 4 using the TL approach, which is appropriate for handling such settings[9].

3.5.1. Experiment 4—Transferring Argumentative Choices from One Domain to Another. In this experiment, 150 study participants were recruited, 110 of whom were bachelor degree students studying Computer Science and 40 of whom were Intel employees. The participants were asked to participate in 3 discussions, each 1 month apart, on the following topics (in the presented order[10]): "Should voting be made obligatory?", "Should gambling be legalized?" and "Would you get an influenza vaccination this winter?". In each chat, study participants were coupled randomly such that students were coupled with peer students and Intel employees were coupled with their peers. The coupling was carried out manually by our research assistant, who asked the study participants for their preferred time slots and matched every couple accordingly.

[9]Due to the complex logistics of this experiment, in the conversations on "Would you get an influenza vaccination this winter?" one of each paired participants was equipped with an argument provision agent, as described in Section 4. Thus, the prediction analysis presented here reflects all study participants in the "Voting" and "Gambling" topics, but only half of the study participants in the "Influenza Vaccinations" topic (i.e., study participants who were not equipped with an advising agent).

[10]Due to the high costs and logistics of recruiting study participants for such a long experiment (over 2 months) we used half of the group to test the TLA in Section 4 with the topic of "influenza vaccinations"; therefore, the topic had to be the last of the three. Note that the discussions took place a month apart from each other to decrease possible biases.

During each chat, participants did not know the identity of their deliberation partner and were instructed to refrain from revealing identifying details, such as their name, age, and so forth. Thus, participants were represented as participant A and participant B during the chat. In order to keep track of the study participants' argumentative choices across the 3 domains, each participant was assigned an experiment identification number $id \in Identifiers$. When logging in to the chat system, both participants were requested to type in their experiment ids, which were saved for later analysis. Participants were informed of the deliberation topic a week before the chat and were instructed to deliberate over the topic as they would in a face-to-face conversation. Note that, for each of the 3 topics, participants were coupled randomly with different deliberation partners.

The chats on "Should voting be made obligatory?" and "Should gambling be legalized?" were annotated by a human expert using the argument corpus provided by Watson, The Debater© research team at IBM. Chats on "Would you get an influenza vaccination this winter?" were annotated using the argument corpus constructed in Experiment 3 (Section 3.1.3).

ML methods, such as the ones suggested in Section 3.4, work only under the assumption that the training and test data are drawn from the same feature space and label space. That is, given a training set of argumentative choices in an argumentative domain α, an ML method generates a prediction model $P_\alpha : \chi_\alpha \rightarrow A_\alpha$, where χ_α is the feature space for the α domain and A_α is the argument set available for the α domain.

Unfortunately, given a training set of argumentative choices from domain α, constructing a prediction model suitable for domain β raises 2 major problems:

(1) *Labels:* The argument set A_α from which the training data's labels are sampled differs from the argument set A_β on which the model is evaluated.
(2) *Features:* α and β induce different argumentation frameworks; thus, $\chi_\alpha \neq \chi_\beta$. Namely, the *context-based features* described in Section 3.4.2 are domain-dependent features. The context-based features hold the last arguments presented in a discussion, which differ between argumentative domains.

To bridge the two issues, we use a feature-based approach for Inductive Transfer Learning [Pan and Yang 2010].

To address the first issue, we used the characterization of each argument (m_a), suggested in Section 3.4, and mapped all arguments, from all topics and conversations, to a joint space called *ArgSpace*.

Definition 3.3. $ArgSpace = J \times D_1 \times D_2 \times D_3 \times C$, where $j \in J$ is the justification value of an argument calculated by Cayrol's calculation, $d_i \in D_i$ is a relevance measurement of an argument, and $c \in \{1, 0, -1\}$ is the confirmation value of an argument (as described in Section 3.4.1).

We deployed the function $\varphi : A \rightarrow ArgSpace$ on all arguments from all conversations such that $\varphi(a) = m_a$. *ArgSpace* is the arguments' characterization space, where $m_a \in ArgSpace$ describes argument a in the context in which it is judged (the context in which a reasoner evaluates the argument). That is, in a given state of the discussion, m_a represents the characteristics of a with respect to the current state of the discussion. Note that *ArgSpace* ignores the topic of the argumentative discussion (Ω) and represents all arguments of all topics and conversations in a single, unified space.

We define *ArgSpace* to be the target space of our TL prediction model. That is, unlike the prediction model described in Section 3.4, which defines A (the argument set of the learned domain) as being the target space of the prediction, the TL prediction model uses *ArgSpace* regardless of the target domain (which, in our setting, is the "Influenza Vaccinations" domain) to allow it to make predictions across different, and

possibly unknown, domains (specifically, across different argument sets). This change has its drawbacks; for example, the proposed prediction model would not predict which argument would be used next in a discussion but rather the *characteristics* of that argument ($m_a =< j, d_1, d_2, d_3, c >\in ArgSpace$).

In the SUV example, the four arguments presented to the study participant are mapped to *ArgSpace* in the following fashion;

φ(Taking out a loan) $=< -0.33, 1, 2, 1, 1 >$
φ(High taxes) $=< 0, 1, 2, 1, -1 >$
φ(High interest) $=< 0, 2, 3, 2, -1 >$
φ(Safe) $=< 0, N/A, 1, 1, 1 >$

Regarding the second issue mentioned earlier, we cannot use the feature space χ as described in Section 3.4. χ includes the context-based features as described in Section 3.4.2, which are domain dependent. The context-based features hold the last arguments presented in a discussion, which cannot be used when learning from domain α and predicting on domain β. Therefore, we used a different feature space definition for the TL approach by replacing the last presented arguments in the discussion (i.e., the context-based features) with their characteristics in *ArgSpace*. Note that each argument a, regardless of the topic, can be mapped to a tuple $< j, d_1, d_2, d_3, c >\in ArgSpace$ by using a mapping function φ that implements the calculations of j, d_1, d_2, d_3 and c as described in Section 3.4. Namely, we change the context-based features from representing the *labels* of the last presented arguments (e.g., a and b) to the *characteristics* of the last presented arguments (e.g., $\varphi(a)$ and $\varphi(b)$). Furthermore, in contrast with previous experiments, Experiment 4 provides us with 3 conversations *per study participant*. This provides an additional dimension to consider when transferring argumentative choices from one domain to another. Specifically, when transferring an argumentative choice from domain α to domain β, we ought to provide special attention to transferred argumentative choices made by the same study participant on whom we perform the prediction. The rationale is to use all provided argumentative choices from all available domains and study participants in order to generate a prediction. However, when transferring previous argumentative choices made by the same study participant on whom we perform the prediction, we will consider these argumentative choices as more influential than choices made by other study participants. In order to distinguish between argumentative choices made by different study participants, we change the feature space to also include the user's identifier (a unique number representing the user in the experiment) as part of the feature vector.

Overall, we change the features used in Section 3.4 by changing the context-based features from the last argument labels to their characterizations and by adding the study participant experiment id. We refer to this new feature space as χ^*.

As a result of these two solutions, the choice of argument a in a discussion is mapped into a pair $< \vec{x}^*, \varphi(a) >$, where $\vec{x}^* \in \chi^*$ and $\varphi(a) = m_a \in ArgSpace$. We emphasize that this representation of argumentative choices uses a single feature space χ^* and a single label space *ArgSpace* across all domains, making TL methods applicable.

In order to predict the next argument to be presented in a discussion, we first calculate the feature vector $\vec{x}^* \in \chi^*$ as described earlier, to represent the deliberant and the deliberation context. Then, we predict the characteristics of the next presented argument, that is, we predict $< \hat{j}, \hat{d_1}, \hat{d_2}, \hat{d_3}, \hat{c} >= \hat{\hat{a}} \in ArgSpace$. To that aim, we trained a Multidimensional Regression model[11] that receives $\vec{x}^*$ as input and predicts

[11]A meta algorithm that allows several one-dimensional regression algorithms to be combined to allow an M-dimensional input to be mapped to an N-dimensional output.

$\hat{\vec{a}} \in ArgSpace$. The model uses 5 separate SVM regression models [Smola and Schölkopf 2004], each predicting a different value characterizing the next predicted argument in the discussion, that is, SVM_1 predicts $\hat{j}$, SVM_2 predicts $\hat{d_1}$ and so on.

For the evaluation of the Multidimensional Regression model, we removed all conversations over each topic, one at a time, and used the remaining conversations (from the two remaining topics) as training data. The conversations from the topic removed are used to evaluate the model. Interestingly, the Multidimensional Regression model achieved a relatively high accuracy with respect to the different examined dimensions of *ArgSpace*; the Mean Absolute Errors (MAE) in predicting j (i.e., the mean value of $|\hat{j} - j|$), d_1, d_2, d_3, and c are 0.15, 0.5, 0.4, 0.45, and 0.07, respectively.

Note that the prediction of $\hat{\vec{a}} \in ArgSpace$ does not naturally translate into an argument in the target domain's argument set. The idea is to search for the argument in the target domain's argument set whose *characteristics* are the most *similar* to the predicted ones, that is, $\hat{\vec{a}}$. For that purpose, we need to define a distance measurement between arguments in *ArgSpace*. Given such a distance measurement w, we can translate any $\hat{\vec{a}} \in ArgSpace$ into the target domain's argument set A using $argmin_{a \in A} distance_w |\hat{\vec{a}} - a|$.

In order to define a distance measurement over *ArgSpace*, we used a Genetic Algorithm (GA)–based method successfully deployed in different applications (a recent example is the painter classification problem [Levy 2014]), which uses the Weighted Nearest Neighbor (WNN) method. In the WNN approach, one seeks to find a weight vector (i.e., chromosome) that will define the distance between every pair of arguments in *ArgSpace* using a weighted sum; let $a, b \in ArgSpace$, then $distance_w(a, b) \Sigma_{i=0}^{5} w_i (a_i - b_i)^2$. This distance measurement (in fact, this is a metric) will be used to identify the argument in the target domain's argument set A that is most similar to the predicted $\hat{\vec{a}}$ using $argmin_{a \in A} \Sigma_{i=0}^{5} w_i (a_i - \hat{\vec{a}}_i)^2$.

We seek to find a chromosome that will maximize the correct classifications using this classification approach. To evolve this chromosome, we removed all conversations over each topic, one at a time, and used the remaining conversations (from the two remaining topics) as training data. The conversations from the topic removed are used for the fitness calculation of the chromosome, that is, the number of correctly classified argumentative choices in the conversation from the topic removed.

Given the training data, we randomly generated a population of 500 weight vectors (i.e., chromosomes), with each vector of 5 nonnegative numbers representing the weights associated with the different dimensions of *ArgSpace*. We then implemented a stochastic universal sampling module, with double-allele-mixing crossover operators (mating two genotypes by randomizing the parents' alleles) with an 80% occurrence, a Gaussian additive mutation operator with 40% occurrence, and 5% elitism for 250 generations. The fitness function for a chromosome is the number of argumentative choices correctly classified in the left-out domain (the higher the better) given the Multidimensional Regression model's prediction $\hat{\vec{a}}$. The process was repeated 3 times, arriving at 3 distinct chromosomes, one per topic. In simple terms, we allowed the weight vector population to produce the best weights per target domain, that is, to evolve the weight vector that produces the highest number of correctly classified argumentative choices in the target domain per the Multidimensional Regression model's prediction.

We then evaluated the 3 calculated chromosomes. For the evaluation, we used the left-out topic as the evaluation set. Surprisingly, despite evolving with overfitting fitness function (the weight vector population evolves with respect to the evaluation set), the accuracy of the best weight vectors was rather poor. When using the Multidimensional Regression model and predicting the 3 closest WNN arguments in the target domain's argument set, the approach averaged 12% accuracy across the 3 domains.

These results are worse than relevance-based prediction (see Section 3.3), which does not require *any* training data, and averages 57% accuracy across the 3 domains.

3.6. Discussion on the Prediction of Human Argumentative Behavior

These results, based on structured argumentation (Experiment 1), free-form human deliberations (Experiment 2) and semistructured chats (Experiment 3), show that the fundamental principles of Argumentation Theory cannot explain or predict a large part of human argumentative behavior. Thus, Argumentation Theory, as it stands, should not be assumed to have descriptive qualities when it is implemented with people.

Despite its simple implementation and promising results in predicting human argumentative behavior, the Relevance heuristics have not received any attention in the existing literature on human argumentative behavior.

The results from using ML techniques in predicting human argumentation suggest that the prediction of human argumentative behavior is possible in structured, semistructured, and free-form argumentation *as long as training data on the desired topic is available*. The results also suggest that ML can be useful in investigating argumentation in the real world.

On the other hand, the use of TL did not perform satisfactorily. Apparently, given conversations on the target topic, conversations on different topics (even from the same deliberant) do not enhance prediction accuracy. Moreover, when no conversations over the desired domain are available, it is better to use the relevance heuristics rather than deploy a sophisticated TL and GA-based WNN approach. There are several possible explanations for these results, two of which arethe following. (1) People do not use a cross-topic deliberation style—people might deliberate differently over different topics, depending on varying factors such as their knowledge of the topic, their attitude toward the discussed issue, and so on. (2) The topics are too different—even if people follow some argumentative patterns, these patterns are hard to detect, as they manifest themselves differently in unrelated topics. Some study participants claimed that one or two of the selected topics were not interesting; thus, the conversations were rather absent-minded, making it extremely difficult to predict the study participants' arguments. We hope that these results will inspire researchers in other fields to take on the challenge of investigating cross-topic human argumentative behavior.

Identifying cultural differences has been shown to have a vast impact on automated negotiations [Haim et al. 2012]. In the scope of this work, on two occasions, the argumentative cultural difference between Israeli and American study participants was investigated: first, Experiment 1 was performed twice, once with an American group (from AMT) and once with an Israeli group (students). Despite the age and potential cultural differences between the groups, the ML model was able to learn from one group and predict for the other without enduring a significant loss in prediction accuracy. Second, and perhaps the most surprising result, is the fact that the results of Experiment 2 and Experiment 3 were shown to be extremely similar both in reference to influential features of the prediction model and in the model's accuracy itself. There are several major differences between Experiment 2 (annotated phone conversations between American residents[12], in English, recorded in the late 1980s and early 1990s) and Experiment 3 (semistructured chats between Israeli students, in Hebrew, collected at the end of 2014). However, extremely similar results were recorded. The results suggest that using a cross-cultural (US–Israeli) model is possible, though further investigation of this topic is needed.

Additional argumentative, psychological, and social issues should be investigated in accordance with the gathered data.

[12]Unfortunately, no demographic data is available.

4. AGENTS FOR PROVIDING ARGUMENTS

Given the encouraging results in predicting human argumentative behavior (Section 3), we now direct our attention to the task of utilizing the suggested prediction models in developing argument provision agents.

4.1. Agents' Policies

There are two main approaches when suggesting an argument to a deliberant: suggest an argument that the deliberant has considered and would (probably) use anyway or suggest innovative arguments, those that the deliberant has (probably) not considered. We designed 9 argument provision policies that implement the two approaches, separately and combined. The agents used our four suggested prediction methods (Section 3) to identify which arguments people are prone to use in a given deliberation state.

—*Predictive agent* (PRD) offers the top 3 ranked arguments using the ML prediction model that were not already mentioned in the discussion, that is, the arguments that best fit the discussion's situation and the deliberant as learned from the training-set with the exception of arguments that were already presented in the discussion.
—*Predictive agent with repeated arguments* (REP) offers the top 3 ranked arguments in the prediction model while enabling the provision of arguments that were already used in the conversation, that is, the agent provides the best-fitting arguments to the situation and the deliberant as learned from the training set without any restrictions on the provided arguments[13].
—*Relevance-based heuristic agent* (REL) offers the 3 "closest" arguments to the last given argument (using edge-distance). We tested whether the relevance notion could act as a good policy. Note that the REL agent requires no complex modeling or training.
—*Weak Relevance-based heuristic agent* (WRL) offers the 3 least related arguments to the last argument (using edge-distance). The idea behind this policy is to offer the user arguments that the user would not naturally contemplate or put forth.
—*Predictive and Relevance-based Heuristic agent* (PRH) offers the top 2 predicted arguments and the most relevant argument (using edge-metrics) which was not part of the predicted arguments. This agent attempts to enjoy the better of the two policies – PRD and REL.
—*Theory-based agent* (TRY) calculates the extension of the argumentation framework using the Grounded semantics and offers 3 arguments that are part of that extension. Because the extension is usually larger than 3, we offer the 3 "closest" arguments to the last given one (using edge-distance). That is, among the "justified" arguments (according to the Grounded semantics), the agent offers the top 3 relevant arguments at the moment.
—*Transfer Learning agent* (TLA) uses the TL prediction methodology described in Section 3.5. The agent suggests the 3 weighted nearest arguments in the target domain's argument set to the predicted values in *ArgSpace*. The agent was tested only when no previous data on the target domain was available.
—*Transfer Learning and Relevance agent* (TLR) suggests the 2 weighted arguments in the target domain's argument set nearest the predicted values in *ArgSpace* and the most relevant argument (using edge-distance) which was not part of the predicted arguments. Similar to the TLA agent, this agent was tested only when no previous data on the target domain was available.

[13]In several conversations in Experiments 2 and 3, we encountered study participants that repeated certain arguments more than once during the conversation, possibly in an attempt to stress the importance of those arguments.

—*Random agent* (RND) offers 3 arguments in a random fashion while avoiding previously used arguments. This policy served as a baseline.

4.2. Experimental Design

In order to evaluate the proposed policies, we used the "Influenza Vaccinations" topic that was shown to spark long and quality conversations in Experiment 3.

First, we used Experiment 3's conversations on "Influenza Vaccinations" to train the prediction model for the PRD and REP agents and Experiment 4's conversations on "Voting" and "Gambling" to train the TLA and TLR agents[14].

Second, we implemented the 9 different agents; each was tested in 17 chats, totaling 153 deliberations with 306 human study participants. Similar to Experiment 3, in each chat we coupled 2 study participants who were asked to deliberate over the same topic of "Influenza Vaccinations", but in a *free -form chat*. Note that only one participant in each chat was assigned a personal agent in order to maintain the scientific integrity of the results. From the 306 study participants who took part in this experiment, 286 were Israeli students who were recruited from classrooms, libraries, and the like and 20 were Intel employees. The study participants ranged in age from 18 to 65, with about 60% male and 40% female participants[15].

The identification of the arguments used by the deliberants was done in a *Wizard of Oz* fashion, in which during the chat a human expert[16] mapped the given sentences into the known arguments in the previously built argumentation framework (consisting of 40 arguments; see Section 3.1.3). The deliberant who was assigned an agent received 3 suggestions on the right side of the screen in a textual form following each presented argument in the discussion (by either of the deliberants). Suggestions started to appear after encountering 2 arguments in the deliberation to enable a short learning period for the agent. We emphasize that, excluding the TLA and TRL agents, other agents had no prior knowledge of the deliberant and required no information from the study participant prior or during the deliberation. Study participants could not select a suggested argument by clicking on it, but had to type their arguments in a designated message box. This restriction was implemented to avoid "lazy" selections.

All obtained deliberations consisted of 4 to 20 arguments (mean 9), and took between 5min to 21min (mean 12min). Deliberations ended when one of the deliberants chose to end them, just as in real life. Yet, in order to receive the 15 NIS payment (the price of a cup of coffee and a large pastry in the University cafeteria), the deliberants had to deliberate for a minimum of 5min.

At the end of each session, the study participant who was equipped with an agent was asked to provide a subjective benefit from the agent on the following scale: Very positive, Positive, Neutral (neither positive nor negative), Negative, Very Negative.

4.3. Evaluation

We evaluated the agents' quality using the study participants' subjective *Reported Benefit* and the *Normalized Acceptance Rate*, which is defined as follows: For each conversation, we calculated the percentage of arguments that the study participant used from the agent's suggestion. Then, we averaged those percentages to calculate the

[14]As part of Experiment 4, in the conversations about "Influenza Vaccinations", one of each paired study participants was equipped with an argument provision agent, either the TLA or the TLR agent. Thus, the analysis presented here reflects only *half* of the study participants who participated in Experiment 4.

[15]Per Intel's request, Intel employees were not asked to provide their demographics; thus, the presented numbers are based on the student group and our estimation concerning the Intel group.

[16]In order to prevent the expert from being biased toward one of the agents, the expert was not involved in any other part of the research and, in particular, in building the agents.

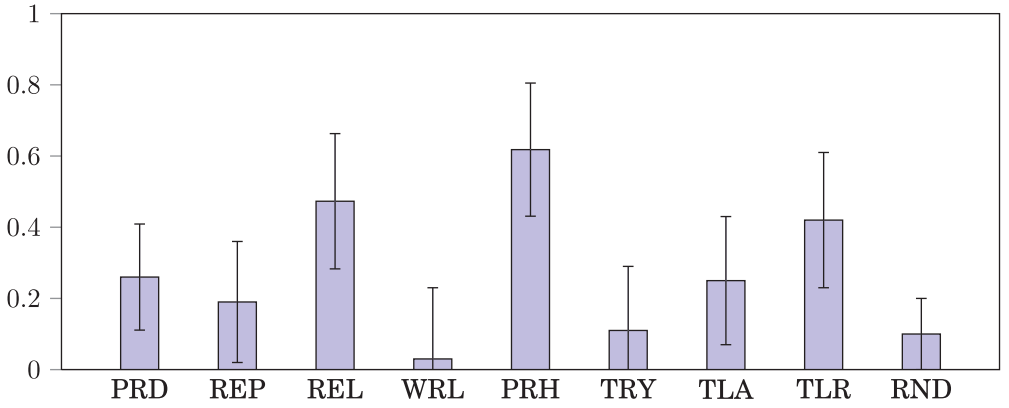

Fig. 4. Normalized Acceptance Rate for the agents. Error bars indicate standard errors.

Normalized Acceptance Rate for each agent. The Normalized Acceptance Rate of the PRH agent was significantly higher than the other agents, averaging 62% acceptance (the study participant's acceptance rate ranged between 20% and 100%), whereas the PRD agent averaged 26% (0%–50%), the REP agent averaged 19% (0%–50%) and the REL agent averaged 47% (10%–100%). The comparison of the TL agents, TLA and TLR, with the PRH agent is biased. The TL agents do not have training data on the target topic ("Influenza Vaccinations"), thus cannot compete with the PRH agent that specializes in the target topic. Nevertheless, the TL agents (TLR achieved 42% (0%–80%) and TLA achieved 25% (0%–50%)), were outperformed[17] by the REL agent (47%), which also does not use training data on the target domain. Note that the REL agent does not require any training data; thus, it uses a much simpler modeling than the TL agents. The WRL, RND, and TRY agents performed very poorly, achieving 3%, 10%, and 11%, respectively. The PRH agent outperformed the other 8 agents in a statistically significant manner ($p < 0.05$), using post-hoc univariate ANOVA (see Graph 4).

As for the study participants' subjective *Reported Benefit*, again, the PRH agent outperformed the others in a convincing manner. All 17 study participants equipped with the PRH agent reported a positive benefit (5 reported very positive, 12 reported positive), which is significantly better than the contending agents in the $p < 0.05$ range using Fisher's Exact test. For comparison, the PRD agent achieved 2 very positive benefits, 10 positive benefits, and 5 neutral benefits, whereas no one reported very positive benefits, 12 reported positive benefits, and 5 reported neutral benefits from the REL agent. Also, only 10 study participants reported positive benefits from the TLA and TLR agents (combined), with the other 24 study participants reporting neutral benefits. RND, WRL, and TRY agents again performed very poorly with very few study participants reporting positive benefits.

4.4. Discussion

A strong positive correlation of 0.78 was demonstrated between the study participants' Normalized Acceptance Rate and their subjective Reported Benefit. However, it is hard to claim that the results suggest a causal relation between the two, as both are probably correlated with the quality of the suggested arguments, which was not explicitly evaluated.

[17]This result is not statistically significant.

The PRH agent stood out on both examined axes, the Normalized Acceptance Rate and the subjective Reported Benefit, surpassing the other 8 agents, including the PRD and REL agents that provided the inspiration for its design. A close examination of these agents' results can provide a possible explanation; while the average Reported Benefit from the PRD agent was higher than the one reported from the REL agent, the Normalized Acceptance Rate was lower. We believe that providing predicted arguments is beneficial to the user, as it strengthens the user's beliefs and opinions, yet it does not provide any novel insights. Consequently, using the suggested (predicted) arguments might seem trite, and may cause the user to feel unoriginal or like a conventional deliberant. On the other hand, the REL agent provided novel, yet closely related, arguments. However, some of these arguments did not fit the beliefs or desires of the user, resulting in a lower subjective benefit. Apparently, the combination of the two policies, captured by the PRH agent, takes advantage of the better of the two policies, resulting in a dominating policy.

To our surprise, the lowest-ranking agent on both of the axes that we examined was WRL. As the only agent that does not attempt to predict its users' argumentative behavior, we believed that the users would receive novel arguments that would provide an additional point of view that would enrich the deliberants' knowledge and perspective, resulting in a good subjective benefit. However, only 1 of the 17 study participants equipped with WRL reported a positive benefit, whereas the rest reported neutral benefits.

The TRY and RND agents performed poorly as well. Despite the very limited predictive abilities of Argumentation Theory (Section 3.2), we hoped that we would observe study participants using TRY's suggested justified arguments. Nonetheless, study participants reported low benefits from the agent and showed a low acceptance rate for its arguments.

The TL approach, which failed to predict people's argumentative behavior, has been shown to provide reasonable policies. However, using the TL approach requires prior knowledge about the deliberant (in the form of prior discussions) and rather complex modeling. Thus, the TL agents were dominated by the REL agent, which not only achieved better acceptance rates[18] and higher subjective benefits, but also required much simpler modeling.

Overall, the results demonstrate that, given prior discussions on the desired domain, the PRH is the dominant approach. However, in the absence of such data—for example, when discussing a new or relatively unexplored topic—the REL policy can provide an easy-to-implement methodology that has been shown to outperform theory-based and TL-based policies.

Interestingly, none of the study participants reported negative or very negative benefits from any of the agents. The fact that no one reported a negative or a very negative benefit is very encouraging; even when the advice was not used by the study participants, the agent did not "bother" them. This finding indicates that argument-provision agents, regardless of their algorithms or policies, hold much potential in real-world implementation.

5. CONCLUSIONS AND FUTURE WORK

We performed a pioneer, large-scale empirical study with over 1000 human study participants on the prediction of human argumentative behavior and its implications in the designing and deployment of automated argument-provision agents.

[18]This result is not statistically significant.

Four prediction methods—Argumentation Theory, Heuristics, Machine Learning and Transfer Learning—as well as nine argument-provision agents were implemented and extensively tested with hundreds of human study participants.

We first conclude that the prediction of human argumentative behavior is possible and that its use in the designing of argument-provision policies is beneficial.

We show that Argumentation Theory, despite its appealing properties, does not provide a useful prediction method, nor does it provide a favorable argument-provision policy in deliberations. This finding suggests that other aspects of argumentation in addition to justification should be explored to better bridge the differences between human argumentative behavior and Argumentation Theory.

We further show that the use of the *Relevance* notion, which was first introduced in this study, provides a simple, yet beneficial, prediction method and argument-provision policy. Note that only an argumentation framework is needed for the implementation of this approach. No training phase is needed, nor does it require the collection and annotation of argumentative dialogs prior to deployment.

The use of ML in argumentation has been shown to provide a valuable prediction method. However, this prediction did not translate into favorable argument provision methods without the inclusion of the relevance heuristics. We claim that ML is needed to further investigate argumentative behavior in the real world, and its combination with simple heuristics can enhance its attractiveness as an argument-provision policy in deliberations.

The TL approach provided poor results as a prediction method and unfavorable results as an argument provision policy. We believe that a more fruitful use of TL in argumentation can be achieved in simpler and more restricted domains. For example, when predicting the argumentative choices of a salesman's overtime, it is reasonable to think that similar persuasive arguments will be used when pitching similar goods—for instance, a toaster and a microwave oven. However, this study considered deliberations on varying, unrelated topics, which resulted in disadvantageous results of the TL approach.

Regardless of policy, none of the study participants reported a negative or a very negative benefit from the agent's suggestions, with many study participants reporting positive and very positive benefits. This finding emphasizes the potential held by automated agents in the context of argument provision and the promising possibilities that the prediction of human argumentative behavior holds in designing such agents.

During the research process, we constructed a rather large annotated corpus, in both English and Hebrew, which we would be pleased to share for future research[19].

We intend to expand the suggested methodology and explore how automated argument-provision agents could be used to help people in different argumentative structures, such as negotiations and persuasion (see Rosenfeld and Kraus [2016] for a perliminary report). Note that these argumentative structures are remarkably different from the deliberation structures that we considered in this study. For example, in negotiations, both parties try to maximize some personal utility in the face of partially conflicting interests, while in deliberations, the deliberants merely exchange opinions and beliefs and do not strive to maximize any explicit utility function.

APPENDICES

A. DUNG'S FUNDAMENTAL NOTIONS

Argumentation is the process of supporting claims with grounds and defending them against attacks. Without explicitly specifying the underlying language (natural

[19]Some of our data are available at http://u.cs.biu.ac.il/~rosenfa5/TiiS_experiments.zip.

language, first-order logic...), argument structure or attack/support relations, Dung [1995] has designed an abstract argumentation framework. This framework, combined with proposed semantics (reasoning rules), enables a reasoner to cope and reach conclusions in an environment of arguments that may conflict, support, and interact with each other. These arguments may vary in their grounds and validity.

Definition A.1. A Dungian Argumentation Framework (AF) is a pair $< A, R >$, where A is a set of arguments and R is an attack relation over $A \times A$.

Conflict-Free: A set of arguments S is conflict-free if there are no arguments a and b in S such that aRb holds.

Acceptable: An argument $a \in A$ is considered acceptable with regard to a set of arguments S if and only if $\forall b.bRa \rightarrow \exists c \in S.cRb$.

Admissible: A set S is considered admissible if and only if it is conflict-free, and each argument in S is acceptable with respect to S.

Dung also defined several semantics by which, given an AF, one can derive the sets of arguments that should be considered *Justified* (to some extent). These sets are called *Extensions*. The different extensions capture different notions of justification, some of which are stricter than others.

Definition A.2. An extension $S \subseteq A$ is a set of arguments that satisfies some rules of reasoning.

Complete Extension: E is a complete extension of A if and only if it is an admissible set and every acceptable argument with respect to E belongs to E.

Preferred Extension: E is a preferred extension in A iff it is a maximal (with respect to set inclusion) admissible set of arguments.

Stable Extension: E is a stable-extension in A if and only if it is a conflict-free set that attacks every argument that does not belong in E. Formally, $\forall a \in A \backslash E, \exists b \in S$ such that bRa.

Grounded Extension: E is the (unique) grounded extension of A if and only if it is the smallest element (with respect to the inclusion) among the complete extensions of A.

These semantics have been modified to fit the BAF modeling as described in Amgoud et al. [2008]. This modification is done without losing the semantics' theoretical underpinnings.

B. SCENARIOS USED IN EXPERIMENT 1

The following 6 scenarios were presented to each study participant in a random order. Also, the 4 arguments from which the study participant was asked to select an argument were shuffled as well to avoid biases.

B.1. Scenario 1

This scenario is based on Cayrol and Lagasquie-Schiex [2005a].

During a discussion between reporters R_1 and R_2 about the publication of information I concerning person X, the following arguments are presented:

R_1: I is important information; thus, we must publish it.

R_2: I concerns the person X, where X is a private person and we cannot publish information about a private person without the person's consent.

If you were R_1, what would you say next?

A. X is a minister; thus, X is a public person, not a private person.

B. X has resigned; thus, X is no longer a minister.

C. His resignation has been refused by the chief of the government.

D. This piece is exclusive to us. If we publish it, we can attain a great deal of appreciation from our readers.

In this example, all mentioned semantics agree on a single (unique) extension that consists of all arguments except "Resigned" (option B) and "Private Person" (R_2's argument). Thus, all arguments except "Resigned" and "Private person" should be considered *Justified*, regardless of the choice of semantics.

B.2. Scenario 2

This scenario is based on Cayrol and Lagasquie-Schiex [2005a].

A murder has been committed and the suspects are *Liz* and *Peter*. Two investigators (I_1 and I_2) try to decide who the main suspect is: Liz or Peter. The following pieces of information are available to both investigators:

—The CSI-team analysis suggests that the killer is a female.
—The CSI-team analysis suggests that the killer is small.
—Peter is small.
—A witness claims that he saw the killer, who was tall.

During the discussion between the investigators, the following arguments are presented:
I_1: "Liz should be our primary suspect, as the murder type suggests that the killer is female".
I_2: "Also, the witness testimony indicates Liz".
If you were I_1, what would you say next?
A. The witness is short-sighted; he is not reliable.
B. The crime scene analysis suggests that the killer is small; that does not fit Liz's physical description.
C. Liz is tall; that fits.
D. The killer has long hair and uses lipstick; those are female characteristics.

B.3. Scenario 3

This scenario is based on Amgoud et al. [2008].

Two doctors (D_1, D_2) discuss whether or not to install a prosthesis on the patient X.

During a discussion between the doctors, the following arguments are presented:
D_1: "The patient cannot walk without a prosthesis; we should install it."
D_2: "The installation of a prosthesis requires surgery."
If you were I_1, what would you say next?
A. We can use local anesthesia to lower the risks of the operation.
B. The patient is a tour guide; he needs to walk in order to keep his job.
C. Surgery carries the risk of a post-op infection.
D. Post-op infections are difficult to cure; we should take this into consideration.

B.4. Scenario 4

This scenario is based on Walton [2005].

During a discussion between two judges (J_1, J_2) about whether Alice can be accused of breach of contract, the following arguments are presented:
J_1: "Alice admitted to signing a contract and failed to meet her obligation."
J_2: "Alice was forced to sign a contract; therefore, it is not valid."
If you were J_1, what would you say next?
A. A witness said that Alice was coerced by a known criminal into signing the agreement.

B. The witness in Alice's favor is not objective; she is a close friend of Alice.
C. The witness in Alice's favor works in the same shop; thus, she is a valid witness.
D. A well-known criminal is known for threatening local businesses in Alice's area.

B.5. Scenario 5

This scenario is based on Parsons et al. [2013].

Two military men (M_1, M_2) discuss whether or not to attack an enemy post. During a discussion between the military men, the following arguments are presented:
M_1: "A high-value target is likely to be on the enemy's post; we should consider attacking it."
M_2: "The presence of enemy troops does indicate it."
If you were M_1, what would you say next?
A. An Informant reported seeing a large number of vehicles in the area.
B. We should also consider that the mission may not be very safe.
C. Our UAV reported that there are a small number of fighters in the area.
D. Our UAV is not reliable; its picture quality is low.

B.6. Scenario 6

This scenario is presented in Example 3.1 and is based on Arvapally and Liu [2012].

C. ARGUMENT LIST USED IN EXPERIMENT 3

The argument list used in Experiment 3 was presented to the study participants in Hebrew. The following is a translation of that list. Note that the arguments were presented in a random order to avoid biases, and study participants were given time to go over the arguments before the chat commenced.

(1) Vaccination is the best protection against influenza and can help prevent it.
(2) The flu vaccine increases some people's risk of getting sick.
(3) There is a 20% chance that you could get vaccinated and still end up with the seasonal flu.
(4) Immunity develops if the body's immune system fights a disease on its own.
(5) For young healthy adults, the shot is less effective, as their immune system is strong.
(6) Not all types of the flu have vaccines, and many are more dangerous than those with a vaccine.
(7) Each year, the flu shot is specially formulated to protect against the standard flu as well as a couple of other strains.
(8) The Centers for Disease Control and Prevention recommends that everyone over the age of 6 months get vaccinated against influenza.
(9) It reduces the risk of getting the flu by 60%.
(10) Flu shots can be life-saving.
(11) Some types of flu can cause death in some cases.
(12) Less than half of the population chooses to receive the flu shot each year.
(13) It's the responsible thing to do if you care about your grand-parents/parents/children.
(14) The flu is very rare in our region; the chance of contracting it is low.
(15) The public is unaware of the severity of the flu.
(16) Overuse of the flu vaccine can actually alter flu viruses and cause them to mutate into a more deadly strain.
(17) Getting the shot will not cause you to get the flu.
(18) You might experience some flu symptoms in the days immediately following receipt of the vaccine.

(19) Flu shots may not be safe for some people (due to allergic reactions, for example).
(20) The shot can cause soreness, redness, or swelling in your arm.
(21) Who can guarantee that a future study will not prove that the shot is harmful in the long run?
(22) Each person responds differently to the vaccine, depending on his or her age, immune system, and underlying medical conditions.
(23) Most doctors recommend getting the vaccination.
(24) Some doctors do not recommend getting the vaccination.
(25) Flu shots are easy to get; they are available in almost every clinic.
(26) The flu shot is administered for free.
(27) Getting the shot can take time; you have to wait in line.
(28) More than 200,000 people in the United States are hospitalized every year due to flu-related symptoms.
(29) Severe cases of the flu usually occur among the old and ill.
(30) You can spare yourself the aches and pains of the flu and loss of work days.
(31) You may have the flu virus without showing symptoms and infect others.
(32) Serious complications are much more likely to occur among the elderly and the ill.
(33) The flu has potentially serious complications.
(34) I'm not old and am in good general health.
(35) I prefer to avoid needles as much as I can.
(36) I have never had a flu shot and I have never had the flu.
(37) I'm against vaccinations altogether.
(38) I may be / I am allergic to the vaccine.
(39) There is no reason to believe that you are allergic to the shot.
(40) The vaccination is just another way to get money from people.

D. GRADUAL VALUATION AND CAYROL'S CALCULATION

Definition D.1. Let $W =< A, R, S >$ be a BAF. Consider $a \in A$ with $R(a) = \{b_1, \ldots, b_n\}$ and $S(a) = \{c_1, \ldots, c_m\}$. A *gradual valuation* function on W is $v : A \to V$ such that $v(a) = g(f_{sup}(v(b_1), \ldots, v(b_n)), f_{att}(v(c_1), \ldots, v(c_m)))$, where the *summation function* $f_{def} : V^* \to F_{def}$ (resp. $f_{sup} : V^* \to F_{sup}$) valuates the quality of all of the attacking (resp., supporting) arguments together, and $g : F_{att} \times F_{sup} \to V$ is the *consolidation function* that combines the impact of the attacking arguments with the quality of the supporting arguments.

An instantiation f of f_{sup} or f_{att} must adhere to the following rules:

—$x_i > x_i' \to f(x_1, \ldots, x_i, \ldots, x_n) > f(x_1, \ldots, x_i', \ldots, x_n)$
—$f(x_1, \ldots, x_n) > f(x_1, \ldots, x_n, x_{n+1})$
—$f() = \alpha \leq f(x_1, \ldots, x_n) \leq \beta$[20]

Instantiations of $g(x, y)$ must increase in x and decrease in y.

The following instantiation of v is used throughout the article (taken from Cayrol and Lagasquie-Schiex [2005a]). Let $V = [-1, 1]$, $F_{sup} = F_{att} = [0, \infty]$, $f_{sup}(x_1, \ldots, x_n) = f_{att}(x_1, \ldots, x_n) = \Sigma_{i=0}^{n} \frac{x_i+1}{2}$ and $g(x, y) = \frac{1}{1+y} - \frac{1}{1+x}$.

ACKNOWLEDGMENTS

We would like to thank Intel Collaboration Research Institute for Computational Intelligence for their support in this research. We also want to thank Noam Slonim from Watson, The Debater© research team for sharing data with us.

[20] α (β) is the minimal (maximal) value of F_{sup} (resp. F_{att}).

REFERENCES

Leila Amgoud, Claudette Cayrol, Marie-Christine Lagasquie-Schiex, and Pierre Livet. 2008. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems* 23, 10, 1062–1093.

Ravi Santosh Arvapally and Xiaoqing Frank Liu. 2012. Analyzing credibility of arguments in a web-based intelligent argumentation system for collective decision support based on K-means clustering algorithm. *Knowledge Management Research & Practice* 10, 4, 326–341.

Amos Azaria, Ariel Rosenfeld, Sarit Kraus, Claudia V. Goldman, and Omer Tsimhoni. 2015. Advice provision for energy saving in automobile climate-control system. *AI Magazine* 36, 3, 61–72.

Pietro Baroni, Massimiliano Giacomin, and Beishui Liao. 2015. I don't care, I don't know I know too much! On incompleteness and undecidedness in abstract argumentation. In *Advances in Knowledge Representation, Logic Programming, and Abstract Argumentation*. Springer, 265–280.

Trevor J. M. Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13, 3, 429–448.

Markus M. Berg. 2015. *Modelling of Natural Dialogues in the Context of Speech-based Information and Control Systems*. Ph.D. Dissertation. Christian-Albrechts University of Kiel, Kiel, Germany.

Jean-François Bonnefon, Didier Dubois, Hélène Fargier, and Sylvie Leblois. 2008. Qualitative heuristics for balancing the pros and cons. *Theory and Decision* 65, 1, 71–95.

Richard Booth, Martin Caminada, Mikołaj Podlaszewski, and Iyad Rahwan. 2012. Quantifying disagreement in argument-based reasoning. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12), Valencia, Spain, June* 4–8, 2012. 493–500.

Gerhard Brewka, Sylwia Polberg, and Stefan Woltran. 2014. Generalizations of Dung frameworks and their role in formal argumentation. *Intelligent Systems* 29, 1, 30–38.

Elena Cabrio, Serena Villata, and Adam Wyner (Eds.). 2014. *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing,* Forlì-Cesena, Italy, July 21–25, 2014. CEUR Workshop Proceedings, Vol. 1341. CEUR-WS.org. http://ceur-ws.org/Vol-1341

Colin Camerer. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, Princeton, NJ.

Justine Cassell. 2000. *Embodied Conversational Agents*. MIT Publication.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005a. On the acceptability of arguments in bipolar argumentation frameworks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer, 378–389.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005b. Graduality in argumentation. *Journal of Artificial Intelligence Research JAIR* 23, 245–297.

Federico Cerutti, Nava Tintarev, and Nir Oren. 2014. Formal arguments, preferences, and natural language interfaces to humans: An empirical evaluation. In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI'14),* 18–22 August 2014, Prague, Czech Republic. 207–212.

James S. Coleman and Thomas J. Fararo. 1992. *Rational Choice Theory*. Sage, Thousand Oaks, CA.

Nick Craswell. 2009. Mean reciprocal rank. In *Encyclopedia of Database Systems*. Springer, 1703.

Sohan Dsouza, Y. K. Gal, Philippe Pasquier, Sherief Abdallah, and Iyad Rahwan. 2013. Reasoning about goal revelation in human negotiation. *Intelligent Systems, IEEE* 28, 2, 74–80.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77, 2, 321–357.

Paul E. Dunne, Anthony Hunter, Peter McBurney, Simon Parsons, and Michael Wooldridge. 2011. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence* 175, 2, 457–486.

Derek Edwards. 1997. *Discourse and Cognition*. Sage, Thousand Oaks, CA.

James Fan, Aditya Kalyanpur, D. C. Gondek, and David A. Ferrucci. 2012. Automatic knowledge extraction from documents. *IBM Journal of Research and Development* 56, 3.4, 1–5.

Leon Festinger. 1962. *A Theory of Cognitive Dissonance*. Stanford University Press, Stanford, CA.

Gerd Gigerenzer and Reinhard Selten. 2002. *Bounded Rationality: The Adaptive Toolbox*. MIT Press, Cambridge, MA.

Robert M. Gray. 2011. *Entropy and Information Theory*. Springer Science & Business Media, New York, NY.

Galit Haim, Ya'akov Gal, Michele Gelfand, and Sarit Kraus. 2012. A cultural sensitive agent for human-computer negotiation. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12),* Valencia, Spain, June 4–l8, 2012. 451–458.

Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Reinforcement learning of cooperative persuasive dialogue policies using framing. In *Proceedings of the 25th*

International Conference on Computational Linguistics (COLING'14), August 23–29, 2014, Dublin, Ireland. 1706–1717.

Fred Jelinek. 1989. Self-organized language modeling for speech recognition. In *Readings in Speech Recognition*, Alex Waibel and Kai-Fu Lee (Eds.). Morgan Kaufmann.

Mark Klein. 2011. The MIT deliberatorium: Enabling large-scale deliberation about complex systemic problems. In *Proceedings of the International Conference on Collaboration Technologies and Systems (CTS'11),* May 23–27, 2011, Philadelphia, PA. 161.

Mark Klein, Hiroki Sayama, Peyman Faratin, and Yaneer Bar-Yam. 2003. The dynamics of collaborative design: Insights from complex systems and negotiation research. *Concurrent Engineering* 11, 3, 201–209.

Robert M. Krauss. 2001. The psychology of verbal communication. *International Encyclopaedia of the Social and Behavioural Sciences* 16161–16165.

Erez Levy. 2014. *Automatic Painter Classification via Genetic Algorithms and Deep Learning*. Master's thesis. Bar-Ilan University, Ramat Gan, Israel.

Beishui Liao and Huaxin Huang. 2013. Partial semantics of argumentation: Basic properties and empirical. *Journal of Logic and Computation* 23, 3, 541–562.

Marsha M. Linehan. 1997. *Validation and Psychotherapy.* American Psychological Association, Washington, DC.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19, 2, 313–330.

Sanjay Modgil, Francesca Toni, Floris Bex, Ivan Bratko, Carlos I. Chesñevar, Wolfgang Dvořák, Marcelo A. Falappa, Xiuyi Fan, Sarah Alice Gaggl, Alejandro J. García, and others. 2013. The added value of argumentation. In *Agreement Technologies*. Springer, 357–403.

Marie-Francine Moens. 2014. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *Postproceedings of the Forum for Information Retrieval Evaluation (FIRE'13)*.

Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2, 175.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10, 1345–1359.

Simon Parsons, Elizabeth Sklar, Jordan Salvit, Holly Wall, and Zimi Li. 2013. ArgTrust: Decision making with information from sources of varying trustworthiness. In *Proceedings of the 12th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'13),* Saint Paul, MN, May 6–10, 2013. 1395–1396.

Andrea Pazienza, Floriana Esposito, and Stefano Ferilli. 2015. An authority degree-based evaluation strategy for abstract argumentation frameworks. In *Proceedings of the 30th Conferenza Italiana di Logica Computazionale*.

Noam Peled, Moshe Bitan, Joseph Keshet, and Sarit Kraus. 2013. Predicting human strategic decisions using facial expressions. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13),* Beijing, China, August 3–9, 2013.

Iyad Rahwan, Mohammed I. Madakkatel, Jean-François Bonnefon, Ruqiyabi N. Awan, and Sherief Abdallah. 2010. Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science* 34, 8, 1483–1502.

Ariel Rosenfeld. 2015. Automated agents for advice provision. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15),* Buenos Aires, Argentina, July 25–31, 2015. 4391–4392.

Ariel Rosenfeld, Noa Agmon, Oleg Maksimov, Amos Azaria, and Sarit Kraus. 2015a. Intelligent agent supporting human-multi-robot team collaboration. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15),* Buenos Aires, Argentina, July 25–31, 2015. AAAI Press, 1902–1908.

Ariel Rosenfeld, Amos Azaria, Sarit Kraus, Claudia V. Goldman, and Omer Tsimhoni. 2015b. Adaptive advice in automobile climate control systems. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS'15),* Istanbul, Turkey, May 4–8, 2015. 543–551.

Ariel Rosenfeld and Sarit Kraus. 2014. Argumentation theory in the field: An empirical study of fundamental notions. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, Forlì-Cesena, Italy, July 21–25, 2014.

Ariel Rosenfeld and Sarit Kraus. 2015. Providing arguments in discussions based on the prediction of human argumentative behavior. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence,* January 25–30, 2015, Austin, Texas, USA. 1320–1327.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion: A preliminary report. In *22nd European Conference on Artificial Intelligence (ECAI'16)*, The Hague, 29 August-2 September 2016.

Avi Rosenfeld, Inon Zuckerman, Amos Azaria, and Sarit Kraus. 2012. Combining psychological models with machine learning to better predict people's decisions. *Synthese* 189, 1, 81–93.

Avi Rosenfeld, Inon Zuckerman, Erel Segal-Halevi, Osnat Drein, and Sarit Kraus. 2014. Negochat: A chat-based negotiation agent. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'14),* Paris, France, May 5–9, 2014. 525–532.

Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning* 5, 1, 43–102.

Noam Slonim, Ehud Aharoni, Carlos Alzate, Roy Bar-Haim, Yonatan Bilu, Lena Dankin, Iris Eiron, Daniel Hershcovich, Shay Hummel, Mitesh Khapra, Tamar Lavee, Ran Levy, Paul Matchen, Anatoly Polnarov, Vikas Raykar, Ruty Rinott, Amrita Saha, Naama Zwerdling, David Konopnicki, and Dan Gutfreund. 2014. Claims on demand – an initial demonstration of a system for automatic detection and polarity identification of context dependent claims in massive corpora. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING'14): System Demonstrations*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 6–9.

Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14, 3, 199–222.

Douglas Walton. 2009. Argumentation theory: A very short introduction. In *Argumentation in Artificial Intelligence*. Springer, 1–22.

Douglas Walton, Katie Atkinson, Trevor Bench-Capon, Adam Wyner, and Dan Cartwright. 2010. 11 Argumentation in the framework of deliberation dialogue. *Arguing Global Governance: Agency, Lifeworld and Shared Reasoning* (2010), 210.

Douglas N. Walton. 2005. *Argumentation Methods for Artificial Intelligence in Law*. Springer.

Ian Watson. 1999. Case-based reasoning is a methodology not a technology. *Knowledge-based Systems* 12, 5, 303–308.

Received June 2015; revised December 2015; accepted February 2016