Argumentation Theory in the Field: An Empirical Study of Fundamental Notions

Ariel Rosenfeld
Bar-Ilan University, Ramat-Gan, Israel
rosenfa5@cs.biu.ac.il

Sarit Kraus
Bar-Ilan University, Ramat-Gan, Israel
sarit@cs.biu.ac.il

Abstract

Argumentation Theory provides a very powerful set of principles, ideas and models. Yet, in this paper we will show that its fundamental principles unsatisfactorily explain real-world human argumentation and should be adapted. We will present an extensive empirical study on the incompatibility of abstract argumentation and human argumentative behavior, followed by practical expansion of existing models.

1 Introduction

Argumentation Theory has developed rapidly since Dung's seminal work (Dung, 1995). There has been extensive work extending Dung's framework and semantics; Value Argumentation Framework (VAF) (Bench-Capon et al., 2002), Bipolar Argumentation Framework (BAF) (Cayrol and Lagasquie-Schiex, 2005) and Weighted Argumentation Framework (WAF) (Dunne et al., 2011) to name a few. All reasonable frameworks and semantics rely on the same fundamental notions: *Conflict Freedom*, *Acceptability*, *Extensions* from (Dung, 1995), and expand upon them in some way. One more notion, which was not addressed in (Dung, 1995), *Support*, has been increasingly gaining attention (Boella et al., 2010). Overall, the same principals and ideas have prevailed for many years.

All of these models and semantics try to provide a *normative* approach to argumentation, i.e, how argumentation should work from a logical standard. From a *descriptive* point of view, the study of (Rahwan et al., 2010), where the authors investigated the reinstatement principle in behavioral experiments, is the only experimental study, as far as we know, that tested argumentation in the field. Nevertheless, many argumentative tools have been developed over time; MIT's delibrium (Klein, 2011), Araucaria (Reed and Rowe, 2004), ArgTrust (Tang et al., 2012) and Web-Based Intelligent Collaborative System (Liu et al., 2007), that try to provide systems where people can handle argumentative situations in a coherent and valid way. We believe that these argumentative tools and others, as efficient and attractive as they might be, have a difficult time attracting users outside the academia due to the gap between the Argumentation Theory and the human argumentative behavior, which, as previously stated, has not been addressed in the context of Argumentation Theory thus far.

In order to further develop argumentative applications and agents, we conducted a novel empirical study, with hundreds of human subjects, showing the incompatibility between some of the fundamental ideas, stated above, and human argumentation. In an attempt to mimic and understand the human argumentative process, these inconsistencies, which appear even in the weakest argumentative requirements as conflict freedom, pose a large concern for theoreticians and practitioners alike. Our findings indicate that the fundamental notions are not good predictive features of people's actions. A possible solution is also presented which provided better results in explaining people's arguments than the existing theory. This solution, which we call *Relevance*, captures a perceptual distance between arguments. That is, how one argument affects another and how this affect is comprehended by a reasoner. Relevance also holds a predicatory value as shown in recent work (Rosenfeld and Kraus, 2014).

This article's main contribution is in showing that the Argumentation Theory has difficulties in explaining a big part of the human argumentative behavior, in an extensive human study. Secondly, the proposed notion of relevance could in turn provide the argumentation community with an additional tool to investigate the existing theory and

semantics.

2 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 language, first order logic...), argument structure or attack/support relations, Dung has designed an abstract argumentation framework (Dung, 1995). 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 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 w.r.t a set of arguments S iff $\forall b.bRa \rightarrow \exists c \in S.cRb$.
Admissible: A set S is considered admissible iff 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 extenstions capture different notions of justification where some are more strict than others.

Definition 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 iff 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 iff 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 iff it is the smallest element (with respect to the inclusion) among the complete extensions of A.

Definition 3. Similar to the attack relation R, one can consider a separate relation S which indicates *Support* (Amgoud et al., 2008). A supporting argument can also be viewed as a part of another argument internal structure. These two options only differ in the AF structure; the reasoning outcome is not influenced. The support relation was introduced in order to better represent realistic knowledge.

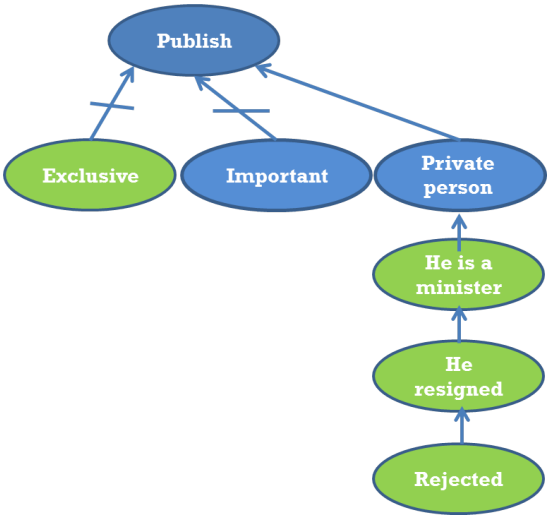

Figure 1: An example of a Bipolar Argumentation Framework; nodes are arguments, arrows indicate attacks and arrows with diagonal lines indicate support.

Let us consider the following example;
Example.
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 his consent.
If you were R_1, what would you say next?
A. X is a minister, so X is a public person, not a private person.
B. X has resigned, so 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.
See Figure 1 for a graphical representation.

In this example, all mentioned semantics agree on a single (unique) extension which consists of all arguments except "Resigned" (option B) and "Private Person" (R_2's argument). Thus, all ar-

guments except "Resigned" and "Private person" should be considered *Justified*, regardless of the choice of semantics.

Argumentation Theory consists of many more ideas and notions, yet the very fundamental ones stated above are the focus of this work.

3 Real Dialogs Experiment

To get a deeper understanding of the relations between people's behaviour in argumentation and the stated notions, we used real argumentative conversations from Penn Treebank Corpus (1995) (Marcus et al., 1993) of transcribed telephone calls and a large number of chats collected toward this aim. The Penn Treebank Corpus consists of transcribed phone calls on various topics, among them some controversial topics such as "Should the death penalty be implemented?" and "Should a trial be decided by a judge or jury?", with which we chose to begin. We went through all 33 dialogs on "Capital Punishment" and 31 dialogs on "Trial by Jury" to identify the arguments used in them and cleared all irrelevant sentences (i.e, greetings, unrelated talk etc.). The shortest deliberation consisted of 3 arguments and the longest one comprised of 15 arguments (a mean of 7). To these dialogs we added another 157 online chats on *"Would you get an influenza vaccination this winter?"* collected from Israeli students, ages ranging from 19 to 32 (mean=24), using a chat interface we implemented. We constructed 3 BAFs, similar to the one in Figure 1, using the arguments extracted from 5 randomly selected conversations. Each conversation which was not selected for the BAF construction was then annotated using the arguments in the BAFs. All in all, we had 64 phone conversations and 157 online chats, totaling 221, all of which are of argumentative nature.

Every conversation provided us with 2 argument sets A_1 and A_2, both subsets of A. We tested every A_i $(i = 1, 2)$ such that $|A_i| \geq 3$ in order to avoid almost completely trivial sets.

Participants were not expected to be aware of *all* arguments in the BAF, as they were not presented to them. Thus, in testing the *Admissibility* of A_i and whether A_i is a part of some *Extension*, we examined both the original BAF and the *restricted* BAF induced by $A_1 \cup A_2$. That is, the argumentation framework in which $A = A_1 \cup A_2$ and the attack and support relations are defined over $A_1 \cup A_2 \times A_1 \cup A_2$, denoted as $AF{\downarrow}_{A_1 \cup A_2}$.

3.1 Results

The first property we tested was *Conflict-Freedom*, which is probably the weakest requirement of a set of arguments. We had anticipated that all A_i would have this property, yet only 78% of the deliberants used a conflict-free set A_i. Namely, that 22% of the deliberants used at least 2 conflicting arguments, i.e, one attacks the other. From a purely logical point of view, the use of conflicting arguments is very grating. Yet, we know that some people try to portray themselves as balanced and unbiased, and as such use contradictory arguments to show that they can consider both ends of the argument and can act as good arbitrators. When we examined *Acceptability*, we tested if every argument $a \in A_i$ is acceptable w.r.t $A_i \setminus \{a\}$. We found that 58% of the deliberants followed this rule. *Admissibility* was tested according to both the original framework and the restricted framework. Merely 28% of the A_is used are considered admissible w.r.t the original framework, while more than 49% qualify when considering the restricted BAF. We can see that people usually do not make the extra effort to ensure that their argument-set is admissible. A possible explanation can be values (norms and morals), as described in (Bench-Capon et al., 2002). Given a set of values, a reasoner may not recognize the attacking arguments as defeating arguments as they advocate a weaker value. As such, the reasoner considers his set admissible. A similar explanation is provided in (Dunne et al., 2011), where a reasoner can assign a small weight to the attacking arguments and as such still consider his set admissible. These explanations can also partially account for the disheartening results in the test of *Extensions*. When examining the original framework, less than 30% of A_is used were a part of some extension, with Preferred, Grounded and Stable performing very similarly (28%, 30%, 25%). When considering the restricted framework, 49%, 50% and 37% of the deliberants used A_is that were part of some extension prescribed by Preferred, Grounded and Stable (respectively) under the restricted BAF. As for *Support*, 27% of the arguments selected were supporting arguments, i.e, arguments which do not attack any other argument in the framework. Although they cannot change the reasoning outcomes, people naturally consider the supporting arguments, which traditionally are not considered "powerful".

To strengthen our findings we performed yet another experiment. We tested the notions in a controlled and structured environment, where the participant is aware of all arguments in the framework.

4 Structured Argumentative Scenarios

We collected 6 fictional scenarios, based on known argumentative examples from the literature (Walton, 2005; Liu et al., 2007; Cayrol and Lagasquie-Schiex, 2005; Amgoud et al., 2008; Tang et al., 2012).

Two groups of subjects took part in this study; the first consisted of 64 US citizens, all of whom are workers of Amazon Mechanical Turk, ages ranging from 19 to 69 (mean=38, s.d=13.7) with varying demographics. The second consisted of 78 computer science B.Sc. students from Bar-Ilan University (Israel), ages ranging from 18 to 37 (mean=25, s.d=3.7) with similar demographics.

Each subject was presented with the 6 scenarios. Each scenario was presented in a short textual dialog between 2 participants, similar to the journalists' example above. The subject was instructed to place himself in one of the deliberants' roles, given the partial conversation, and to choose the next argument he would use from the four available arguments. We instructed the subject to consider only the arguments in the dialog and the proposed ones, and refrain from assuming any other information or possible arguments in the dialog's context.

The following example, based on (Liu et al., 2007), was presented to the subjects;

Example.

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 participant was then asked to put himself in S_1's shoes and choose the next argument to use in the conversation. The options were: 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, safety is very important to us", D. "There are high taxes on SUVs".

See Figure 2 for a graphical representation of the aforementioned framework.

The distribution of selections in the above example was as follows; A.35%, B.24%, C.8%, D. 33%. There is only one (unique) extension in this scenario which includes "High interest" and "high taxes". Especially when considering "Taking out a loan", it should be considered overruled (unjustified/invalid), or at least very weak, as it is attacked by an undisputed argument. As we can see, only slightly over half of the subjects choose an argument from the extension, i.e, a somewhat *Justified* argument.

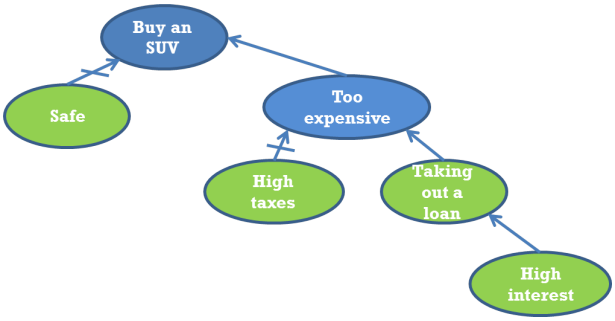

Figure 2: SUV example of BAF

4.1 Results

The distribution of selections, in all scenarios, suggests that there could be different factors in play, which differ from one subject to another. Thus, there is no decisive answer to what a person would say next. Unfortunately, testing *Conflict Freedom* and *Admissibility* is inapplicable here. None of the subjects was offered an argument that conflicts with its previous one and could not choose more than one argument to construct an admissible set. When examining *Extensions*, all scenarios which were presented to the subject are *Well Founded* (that is to say, there exists no infinite sequence $a_0, a_1, \ldots, a_n, \ldots$ such that $\forall i.(a_i, a_{i+1}) \in R$). As such, all mentioned semantics coincide - only one extension is Grounded, Stable and Preferred. Of the 6 scenarios, 5 had suggested 2 justified arguments and 2 overruled arguments (arguments which are not part of any extension) to the subject. In these 5 scenarios, 67.3% of the time a justified argument was selected (on average). This result is disappointing since 50% is achieved by randomly selecting arguments. As for *Support*, 49.4% of the arguments selected were supporting arguments, i.e, arguments which do not attack any other argument in the framework. Even more interesting is that 80% of the time people chose (directly or indirectly) an argument supporting their

first argument. This phenomenon can be regarded as a *Confirmation Bias*, which is recorded in many fields (Nickerson, 1998). Confirmation bias is a phenomenon wherein people have been shown to actively seek and assign more weight to evidence that confirms their beliefs, and ignore or underweigh evidence that could disconfirm their beliefs. Confirmation Bias can also explain the persistence of discredited beliefs, i.e, why people continue to consider an argument valid/invalid despite its logical argumentative status. Here it is extremely interesting since the subjects only played a role and it was not really their original argument. There is a strong tension between the *Confirmation Bias* and *Extensions*. In some scenarios the subject is given a situation in which he "already used" an overruled argument, and therefore had a problem advocating it by using a supporting argument.

We had anticipated that in finite and simple argumentative frameworks people would naturally choose the "right" arguments, yet we again see that the argumentative principals unsatisfactorily explain people's argumentative selections. This is not a complete surprise, since we have many examples in the literature where people do not adhere to the optimal, monolithic strategies that can be derived analytically (Camerer, 2003).

We have shown here, in two separate experiments, that a similar phenomenon occurs in the context of argumentation - people do not choose "ideal" arguments according to the Argumentation Theory.

5 Relevance

It is well known that human cognition is limited, as seen in many examples in (Faust, 1984) and others. In chess for example, it is common to think that a beginner can consider about 3 moves ahead and a master about 6. If we consider the argumentation process as a *game* (McBurney and Parsons, 2009), a player (an arguer) cannot fully comprehend *all* possible moves (arguments) and their utility (justification status) before selecting a move (argument to use) when the game (framework) is complex. The depth and branching factor limitations of the search algorithms are of course personal. For example, we would expect an educated adult to be able to better consider her arguments than a small child.

Definition 4. Let a,b be arguments in some AF. $Rel : A \rightarrow P(A)$ is a personal relevance function which given argument $a \in A$ (for evaluation) returns a set of arguments $A' \subseteq A$ which are, given the reasoner's cognitive limitations and knowledge, relevant to a. Using Rel, we can distinguish between relevant and irrelevant arguments w.r.t a given argument, yet we gain additional strength in incorporating the reasoner's limitation and biases.

We denote the restriction of AF to arguments relevant to a as $AF{\downarrow}_{Rel(a)} \equiv < A', R' >$ where $A' = Rel(a)$ and $R' = A' \times A' \cap R$.
On $AF{\downarrow}_{Rel(a)}$ one can deploy any semantic of choice.

The simplest way to instantiate the ***Rel*** is $Rel(\cdot) = A$, meaning that all arguments in the AF are relevant to the given argument. This instantiation is the way the classic frameworks address the reasoner's limitations, simply by saying – there are none. As shown in (Liao and Huang, 2013), it is not necessary to discover the status of all arguments in order to evaluate a specific argument/set of arguments. Thus, considering $Rel(a)$ as the maximal set of *affecting arguments* (arguments in which their status affects the status of a) is another natural way to consider relevance, yet without considering cognitive limitations.

We suggest the following instantiation, which we examined empirically.

Definition 5. Let $D(a, b)$ be a distance function, which given arguments a, b returns the directed distance from argument a to b in AF's graph.

Given a distance measurement D we can define an edge-relevance function as follows:

Definition 6. $Rel_D(a) = \{b | D(b, a) \leq k\}$ where k is a non-negative constant.

Naturally, when setting k to 0, every argument a is considered justified in $AF{\downarrow}_{Rel_D(a)}$ (under any semantics). k can be thought of as a depth limitation for the search algorithm used by the reasoner. Of course, if $k = \infty$, $AF{\downarrow}_{Rel_D(a)} = \{$All affecting arguments on $a\}$.

5.1 Empirical Testing

We used several D functions in our work on predicting arguments given a partial conversation (Rosenfeld and Kraus, 2014). When $k = 0$, as stated above all arguments should be considered justified. Analyzing the free-form dialogs using Grounded semantics with $k = 2$ resulted in 72% of the arguments used being part of some exten-

sion, whereas without relevance a little less than 50% was part of some extension.

Relevance provides a way to rationally justify every argument within an AF to some extent. Unlike VAF (Bench-Capon et al., 2002) and WAF (Dunne et al., 2011), which rely on exogenous knowledge about values and weights from the reasoner, relevance can be instantiated without any prior knowledge on the reasoner and still offer a better explanatory analysis of the framework.

6 Conclusions

We presented an empirical study, with over 400 human subjects and 250 annotated dialogs. Our results, based on both free-form human deliberations and structured experiments, show that the fundamental principles of Argumentation Theory cannot explain a large part of the human argumentative behavior. Thus, Argumentation Theory, as it stands, should not be assumed to have descriptive or predicatory qualities when it is implemented with people.

Our relevance notion provides a new way to rationalize arguments without prior knowledge about the reasoner. Relevance, as well as other psychological and social aspects, should be explored to better fit the Argumentation Theory to human behavior. This required step is crucial to the integration of argumentation in different human domains.

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.

Trevor JM Bench-Capon, Sylvie Doutre, and Paul E Dunne. 2002. Value-based argumentation frameworks. In *Artificial Intelligence*.

Guido Boella, Dov M Gabbay, Leendert WN van der Torre, and Serena Villata. 2010. Support in abstract argumentation. In *COMMA*, pages 111–122.

Colin Camerer. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *Symbolic and quantitative approaches to reasoning with uncertainty*, pages 378–389. Springer.

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.

David Faust. 1984. *The limits of scientific reasoning*. U of Minnesota Press.

Mark Klein. 2011. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper*.

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

Xiaoqing Frank Liu, Samir Raorane, and Ming C Leu. 2007. A web-based intelligent collaborative system for engineering design. In *Collaborative product design and manufacturing methodologies and applications*, pages 37–58. Springer.

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.

Peter McBurney and Simon Parsons. 2009. Dialogue games for agent argumentation. In *Argumentation in artificial intelligence*, pages 261–280. Springer.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175.

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.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979.

Ariel Rosenfeld and Sarit Kraus. 2014. Providing arguments in discussions based on the prediction of human argumentative behavior. Unpublished manuscript.

Yuqing Tang, Elizabeth Sklar, and Simon Parsons. 2012. An argumentation engine: Argtrust. In *Ninth International Workshop on Argumentation in Multi-agent Systems*.

Douglas N Walton. 2005. *Argumentation methods for artificial intelligence in law*. Springer.