Explanation Versus Meta-Explanation: What Makes a Case More Convincing?

Boris Galitsky[1], Josep Lluis de la Rosa[2] and Boris Kovalerchuk[3]

[1] Univ. Girona Spain
bgalitsky@hotmail.com
[2] EASY Innovation Center, Univ Girona Spain
peplluis@eia.udg.edu
[3] Central Washington University, Ellensburg, WA 98926 7520
borisk@swu.edu

Abstract

Comparative analysis of the roles of explanation and meta explanation is conducted assessing the validity of explanation exchanged between human agents. Meta explanation links the overall structure of a current scenario with that of previously learned scenarios of multi agent interaction. The scenario structure includes communicative actions of involved agents and argumentation attack relations between the subjects of these actions. Object level explanation is based on a traditional machinery to handle argumentative structure of a dialogue, assessing the plausibility of individual claims.

To assess plausibility of customer complaints, we relate them to the classes of valid (consistent, genuine) and invalid (inconsistent, include attempts to get compensation from a company, or expressing a bad mood). Evaluation of contribution of each explanation level shows that both levels of explanation are essential for assessment of whether a multi agent scenario as described by an agent is plausible or not. We demonstrate that meta explanation in the form of machine learning of scenario structure should be augmented by conventional explanation by finding factual based arguments for individual claims.

We also define a ratio between object level and meta explanation as relative accuracy of plausibility assessment based on former and latter sources. We then observe that groups of scenarios can be characterized based on a specific ratio between object level and meta level explanations in a phase space; such ratio is an important parameter of human behavior associated with explaining in a dialogue.

Introduction

Importance of the explanation-aware computing has been demonstrated in multiple studies and systems. Also, it has been argued that the older model of explanations as a chain of inferences with a pragmatic and communicative model that structures an explanation as a dialog exchange (Walton 2007). The field of argumentation is now actively contributing to such areas as legal reasoning, natural language processing and also multi-agent systems (Dunn and Bench-Capon, 2006). It has been shown (Walton 2008) how the argumentation methodology implements the concept of explanation by transforming an example of an explanation into a formal dialog structure. In this study we differentiate between explaining as a chain of inference of facts mentioned in dialogue, and meta-explaining as dealing with formal dialog structure represented as a graph. Both levels of explanations are implemented as argumentation: explanation operates with individual claims communicated in a dialogue, and meta-explanation relies on the overall argumentation structure of scenarios.

When there is a lack of background domain-dependent information to obtain a full object-level explanation, the evolution of *dialogues* where human agents try to explain their decisions should be taken into account in addition to the communicative actions these arguments are attached to. Rather than trying to determine the epistemic status of those arguments involved, in one of our previous studies (Galitsky et al 08) we were concerned with the emerging *structure* of such dialogues in conflict scenarios, based on inter-human interaction. We refer to such structure as *meta-explanation*. Meta-explanation is implemented as a comparison of a given structure with similar structures for other cases to mine for relevant ones for assessing its truthfulness and assessment whether agents provide proper explanations.

In our earlier studies we proposed a concept learning technique for scenario graphs, which encode information on the sequence of communicative actions, the subjects of communicative actions, the causal (Galitsky et al 05), and argumentation attack relationships between these subjects (Galitsky et al 08). Scenario knowledge representation and learning techniques were employed in such problems as predicting an outcome of international conflicts, assessment of an attitude of a security clearance candidate, mining emails for suspicious emotional profiles, and mining wireless location data for suspicious behavior (Galitsky et al 07).

In this study, we perform a comparative analysis of the *two levels* of explanation-related information mentioned above to assess plausibility of scenarios of interaction between agents. The *meta-level* of explanation is expressed via an overall structure of a scenario, which includes communicative actions and argumentation attack relations. This explanation is learned from previous

experience of multi-agent interactions. Scenarios are represented by directed graphs with labeled vertices (for communicative actions) and arcs (for temporal and causal relationships between these actions and their parameters) (Galitsky et al 05). The *object-level* explanation is expressed via argumentative structure of a dialogue, assessing the plausibility of individual claims, which has been a subject of multiple applied and theoretical AI studies.

Explaining *why* vs explaining *how* (example 1)

We start with a non-dialogue example of explanations and demonstrate that they refer to object-level and meta-level of explanation in various degrees.

Consider a situation where a teacher explains to the student how to build a house. Let us assume that we have a 2nd grade student. The satisfactory explanation will be that first, somebody builds a foundation, then walls and finishes with building a roof. This is a natural reasoning chain with the relationship 'to be on top'.

The next case is when the student is the 7nd grade. The previous explanation will not be satisfactory. The explanation will have the same structure but more objects such as windows, doors, etc.

Now let us assume that the student is a college student who studies construction. Previous explanations are not satisfactory now either, because the construction student needs to learn how he can actually build a house himself.

These simple examples show that the concept of explanation is context specific. In these examples, it was easy for us to understand that first two explanations are not satisfactory for the construction student. Often students complain that they did not get hand-on experience. This is another was to say that the explanation is not satisfactory, or not an explanation at all. In this study, we attempt to represent such context as two-level explanation system.

The explanation has two aspects: (1) explain -- how to build, and (2) explain -- why to build in a particular way. In many cases, a student cannot judge whether the explanation is satisfactory if he still did not try to build a house using obtained explanations. The explanation may miss critical details ("know-how"). However if he is asked if he has got a satisfactory explanation, he may say "yes" because he simply does not know that the explanation is not complete. We call this situation *illusion of explanation*. Even mathematical proof can be faulty. History tells us that it may take time to discover that the proof was incomplete. The issue is that the person who understands that the constriction explanation is not satisfactory often does not need the explanation. He poses the construction knowledge already; there is no need for reasoning chains here.

Explaining *why* usually requires object-level explanation, whereas *how* may rely on meta-explanation level only. We will further consider *why* -kind of explanation, analyzing how agents in a conflict scenario attempt to explain *why* they are right and their opponent are not.

If one takes a fresh look at the above examples, she can see that various cases are specific combination of object-level and meta-explanation. In the further sections, we define each level formally and outline the means to calculate the *ratio* between the 'roles' of object-level and meta-explanation in meeting the objective of an agent trying to communicate his explanation..

Attempting to explain a scenario: which levels are used? (example 2)

We first introduce a scenario as described in a blog (Rase-dezert.com 08, Fig.1) and give the reader a chance to reconstruct what might have happened. This is a simple example, where one can observe how where both sources of argumentation data help to make sense of the scenario, classifying it with respect of the roles of involved agent. Our first example does not include an explicit dialogue/conflict between agents.

Tragedy on the Baja 1000

A helicopter that was flying over the Baja 1000 race route came down today, leaving a death toll of two. Apparently, the craft came in contact with some high voltage cables. The helicopter was rented in the city of Tijuana with the intention of filming the race from the air and reporting its progress.

However it seems that its mission wasn't as innocent as it seemed at first sight…Armed men burst into Ensenada's morgue Wednesday night and took the body of one of the men who died in the helicopter crash. The commando took two state social workers hostage as they grabbed González's body. The commando then left the morgue carrying their macabre cargo. As they retreated, a group of law enforcement officers pursued the assailants, and in the ensuing shootout two Mexican policemen were killed.

The dead man taken from the morgue was identified as Pablo González. It was unclear whether that is his real name and why such measures were taken to recover the body. Mexican media reported he was believed to be a member of the region's Arellano Félix drug cartel.

Fig. 1: introductory scenario where both sources of argumentation are required to find plausible interpretation

Trying to understand whether González is a prominent criminal, a witness of a crime, or a well known auto race enthusiast, the reader employs two levels to explain her interpretation of this scenario, as outlined in the Introduction:

1. Meta-explain: observe typical most familiar scenarios which are similar to our scenarios with respect to expected logic of events (which we call argumentation patterns in this paper). These include: *helicopter crash*

and accident handling, *attack by criminals* to eliminate witnesses or help gang members escape, and criminal run-away. A number of features of our scenario are hard to match by common / typical scenarios, such as *attack of a morgue*, *appearance of a criminal at a public event*, *attack of a police car* by a large group of criminals.

2. Explain: take into account relevant commonsense knowledge, assess whether the main involved agent is a *criminal* or a *crime witness*, because his body was hijacked by a criminal gang, and/or an *auto enthusiast*, because he rented a helicopter to report the auto race.

We now show the 'official' explanation in the press:

> "Maybe it was sentimental reasons," said David A. Shirk, director of the Trans-Border Institute at the University of San Diego. The attackers, said Shirk and others, may have wanted to ensure that the man's funeral was attended by his friends. "If he was buried by authorities, they would expose themselves by coming out for any kind of public funeral," Shirk said.
>
> Federal authorities had initially pointed in an extra-official manner, that the son of Alicia Arellano was participating in a vehicle registered with the number 113, but later the authorities presented a new version where it is presumed that a member of the Arellano Felix family was actually aboard the helicopter that crashed.

On one hand, the most plausible explanation of events comes from the scenario *funeral of a criminal, where friends attend without exposure to public*, which is not very frequent. On the other hand, such explanation might be derived in an attempt to find an argument which attacks the statement 'hijack a crime witness' and supports the argument 'release a member of criminal gang' without attacking the assertion that this member is dead at the time of release. Hence the reader observes that both meta-explaining by learning links between events from familiar scenarios (1), and finding plausible explanation (as we illustrated by attack relationships for individual statements (2) contribute to understanding scenarios and assessing its truthfulness. Hence in this example both levels of explanation are required to come up with a plausible scenario interpretation.

The goal of this paper is to estimate relative importance of these levels for the overall assessment of scenario plausibility. To do that, we build both representations, classify scenarios based on these representations, and evaluate which representation improves the classification accuracy in a higher degree.

Then we will explore the correlation between overall semantic characteristics of scenarios (such as level of competence, truthfulness, motivation of the agent being explaining, and possible attitudes of agents being explained to) and the *ratio* between the above degrees of how each level contributes to classification accuracy.

Meta-explaining agents' behavior in dialog

We approximate an *inter-human interaction scenario* as a sequence of communicative actions (such as *inform, agree, disagree, threaten, request*), ordered in time, with *argumentation attack* relation between some of the subjects of these communicative language.

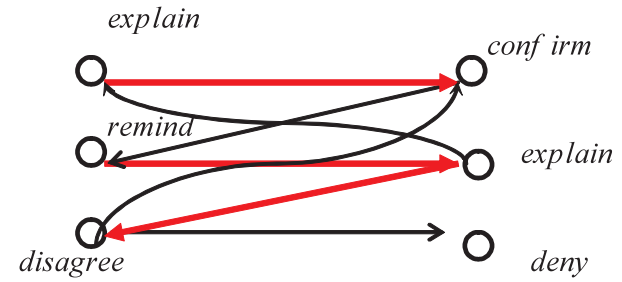
Scenarios are simplified to allow for effective matching by means of *graphs*. In such graphs, communicative actions and attack relations are the most important component to capture similarities between scenarios. Each vertex in the graph will correspond to a communicative action, which is performed by an (artificial) agent. An arc (oriented edge) denotes a sequence of two actions.

In our simplified model of communication semantics (Galitsky 2006) communicative actions will be characterized by three parameters: (1) *agent name*, (2) *subject* (information transmitted, an object described, etc.), and (3) *cause* (motivation, explanation, etc.) for this subject. When representing scenarios as graphs, we take into account all these parameters. Different arc types bear information whether the subject stays the same or not. **Thick arcs** link vertices that correspond to communicative actions with the **same subject**, whereas thin arcs link vertices that correspond to communicative actions with different subjects. We will make explicit conflict situations in which the cause of one communicative action M1 "attacks" the cause or subject of another communicative action M2 via an argumentation arc A (or *argumentation link)* between the vertices for these communicative actions. This attack relationship expresses that the cause of first communicative action ("from") *defeats* the subject or cause of the second communicative action ("to").

For the sake of example, consider the text given below representing a complaint scenario in which a client is presenting a complaint against a company because he was charged with an overdraft fee which he considers unfair (Fig1). We denote both parties in this complaint scenario as ***Pro*** and ***Con*** (proponent and opponent), to make clear the dialectical setting. In this text communicative actions are shown in **bold**. Some expressions appear underline, indicating that they are defeating earlier statements. Fig. 2 shows the associated graph, where straight thick and thin arcs represent temporal sequence, and curve arcs denote defeat relationships.

Note that first two sentences (and the respective subgraph comprising two vertices) are about the current transaction (*deposit*), three sentences after (and the respective sub-graph comprising three vertices) address the *unfair charge,* and the last sentence is probably related to both issues above. Hence the vertices of two respective subgraphs are linked with thick arcs: *explain-confirm* and *remind-explain-disagree*. It must be remarked that the underlined expressions help identify where conflict among arguments arise. Thus, the company's claim *as disclosed in my account information* defeats the client's assertion *due to a bank error*. Similarly, the expression *I made a deposit well in advance* defeats *that it usually takes a day to*

process the deposit (makes it non-applicable). The former defeat has the intuitive meaning "*existence of a rule or criterion of procedure attacks an associated claim of an error*", and the latter defeat has the meaning "*the rule of procedure is not applicable to this particular case*".



- *(Pro) I **explained** that I made a deposit, and then wrote a check which bounced due to a bank error.*
- *(Con) A customer service representative **confirmed** that it usually takes a day to process the deposit.*
- *(Pro) I **reminded** that I was unfairly charged an overdraft fee a month ago in a similar situation.*
- *(Con) They **explained** that the overdraft fee was due to insufficient funds as disclosed in my account information.*
- *(Pro) I **disagreed** with their fee because I made a deposit well in advance and wanted this fee back*
- *(Con) They **denied** responsibility saying that nothing can be done at this point and that I need to look into the account rules closer.*

Fig. 2: A conflict scenario with attack relations.

Our task is to classify (for example, by determining its plausibility) a new complaint scenario without background knowledge, having a dataset of scenarios for each class. We intend to automate the above analysis given the formal representation of the graph (obtained from a user-company interaction in the real world, filled in by the user via a special form where communicative actions and argumentation links are specified).

Explaining individual claims

In this section, we briefly outline our approach to computationally treat object-level explanations in the domain of customer complaints.

To verify the truthfulness of a complainant's claim, we use the special form called Interactive Argumentation Form, which assists in structuring a complaint. Use of this form enforces a user to explicitly indicate all causal and argumentation links between statements which are included in a complaint. The form is used at the object-level argumentation to assess whether a particular scenario has plausible argumentation pattern: does it contain self-attacks (explicit for the complainant).

The role of the Interactive Argumentation Form is a visual representation of argumentation, as well as its intuitive preliminary analysis. To specify supporting and defeating links for a number of statements for each section, multiple instances of these forms may be required for a given complaint. Since even for a typical complaint manual consideration of all argumentation links is rather hard, automated analysis of inter-connections between the complaint components is desired. We use the defeasible logic programming approach to verify whether the complainant's claims are plausible (cannot be defeated given the available data).

In our previous study (Galitsky et al 2008) we provided the definition and algorithm for building dialectic trees to discover implicit self attack in a defeasible logic program, specified by the Interactive Argumentation Form.

Evaluation of contribution

To observe the comparative contribution of explanation in object-level and meta-level to the judgment of scenario plausibility, we used the database of textual complaints which were downloaded from the public website PlanetFeedback.com. For the purpose of this evaluation, each complaint was manually assigned a plausibility assessment: plausible (valid, consistent) or implausible (includes faulty explanations of agents' positions).

For the purpose of this

1) manually represented at a meta-level for machine learning evaluation
2) manually represented as an object level for finding self-defeating explanation claims

This complaint preprocessing resulted in 560 complaints, divided in fourteen banks (or datasets), each of them involving 40 complaints. In each bank 20 complaints were used for training and 20 complaints for evaluation.

We performed the comparative analysis of relating scenarios to the classes of plausible/implausible taking into account 1), 2), and combined (1+2) . Such an analysis sheds a light on the possibility to recognize a scenario (1) without factual knowledge or individual claims, but taking into account similar plausible and implausible dialogues, and (2) with partial background knowledge, expressed as a set of attack relations between claims. We evaluate a cautious approach combining 1) and 2), where scenario is *plausible* if a) it is similar to a plausible one **or** b) it does not contain self-defeated claims, and *implausible* otherwise.

Plausibility assessment results for combined evaluation (1+2) are shown in Table1. On the left, the first three columns contain bank number, and the numbers of plausible/implausible complaints as manually assessed by human experts. The middle set of columns show the classifications results based on 1) & 2) together.

Classification based on the combination of levels (Table1) gives substantial increase in recognition accuracy: F(1)= 63%, F(2) = 77%, and F(1+2)= 89%, which is a 26% of increase of accuracy for (1) and 12% increase of the accuracy for (2).

Obviously, each bank has its own policy in handling customer complaints. In the table above, assuming we processed a statistically significant set of complaints,

peculiarity of each bank is reflected as different contribution of object- and meta-level for scenario classification. Explaining their decisions, some banks rely more on individual facts and their policy rules, and other banks prefer references to "common practice", communicating their explanations. Hence for every group of scenarios involving a fixed set of agents (e.g. representatives of the same bank), one can observe a characteristic ratio between the levels of explanation.

Bank	As assigned by experts		Results of classification: both levels								
	plausible	implausible	plausible	Classified as	Classified as plausible but	Not Classified as implausible but	Precision plausible	Precision implausible	Recall plausible	Recall implausible	F-measure
Bank 1	**8**	**12**	8	11	0	1	100%	92%	100%	92%	96%
Bank 2	**6**	**14**	7	8	3	6	70%	57%	88%	57%	69%
Bank 3	**7**	**13**	9	11	2	2	82%	85%	82%	85%	83%
Bank 4	**5**	**15**	6	14	2	1	75%	93%	86%	93%	89%
Bank 5	**8**	**12**	8	10	0	2	100%	83%	100%	83%	91%
Bank 6	**8**	**12**	7	10	2	2	78%	83%	117%	83%	97%
Bank 7	**11**	**9**	11	9	0	0	100%	100%	100%	100%	100%
Bank 8	**8**	**12**	9	9	1	3	90%	75%	90%	75%	82%
Bank 9	**7**	**13**	7	11	0	2	100%	85%	100%	85%	92%
Bank 10	**9**	**11**	10	10	3	1	77%	91%	91%	91%	91%
Bank 11	**10**	**10**	10	8	0	2	100%	80%	100%	80%	89%
Bank 12	**5**	**15**	6	13	1	2	86%	87%	86%	87%	86%
Bank 13	**10**	**10**	10	9	0	1	100%	90%	100%	90%	95%
Bank 14	**8**	**12**	9	11	1	1	90%	92%	90%	92%	91%
Average	**7.9**	**12**	**8.4**	**10**	**1.07**	**1.86**	**89%**	**85%**	**95%**	**85%**	**89%**

Table 1: Results of the combined classification.

Explanation phase space

Having discussed two levels of explanation, we now intend to explore how the explanation style of individual agent and multi-agent system can be characterized in terms of **degrees** each of these two levels are used. Our intention here is to characterize explanation behavior by a numerical parameter. Since one can 'measure' contribution of object-level and meta-explanation to scenario plausibility as relative accuracy, we believe this measure can serve as explanation behavior parameters, which is invariant with respect to subjects of dialogue and even individual attitudes of particular scenario agents. Hence, we depict a scenario with explanation behavior as a point of two-dimensional space (which we call *explanation phase space*).

We demonstrate that using explanation phase space (Fig. 3), one can visualize the phenomenology of various forms of multi-agent behavior associated with explanation. Less plausible explanation scenarios are shown in the left-bottom corner, and fully valid ones are shown in the top-right corner. A number of epistemic states are shown in the phase space and their object-level and meta-explanations are described. When the trust is high, detailed causal links in explanation do not have to be provided. When object level explanation is very incomplete and meta-explanation is somewhat complete, illusion of explanation (discussed above in example 1) may occur. An adult can say a child about somebody: "He is happy because he is always friendly. Be friendly too". This explanation may not be true but if the child trusts this adult, it will be accepted. This explanation has a structure close to the explanations by a politician: "These people do good things because they are friendly, Let's be friendly people". The incompleteness at object-level can be augmented by the agent by accepting meta-explanation. In English precedent-based legal system meta-level prevails over continental statute-based. Change in a behavior of agent system (demonstrated as a set of scenarios) can be shown as a trajectory in this space.

Results and discussions

We suggested how to split explanation-related behavior presented in a human language into two levels, using reasoning chains (deduction) and similarity between explanation structures (inductive learning). Relative to the former level (explanation), the latter level is an explanation of explanation structure, which we refer to as meta-explanation (explanation of explanation). Hence we split the explanation in multiagent behavior into a deductive object-level and an inductive meta-level. For the first level, we use an interactive form to obtain a defeasible logic program to verify plausibility of individual claims used in explanation. For the meta-level, we represented scenarios as graphs and used graph-based nearest neighbor technique to determine whether given scenario is similar to plausible or implausible scenarios.

We then observed how two levels of explanation, overall argumentation pattern of a scenario and explanations for individual claims, compliment each other. Comparative computational analysis of scenario classification with respect to plausibility showed that assessment of both levels of explanation is essential to determine whether a scenario is plausible or not (contains misrepresentation or self-contradiction). Hence, we believe a practical explanation management system where explanation is implemented via argumentation should include scenario-oriented machine learning capability in addition to handling argumentation for individual claims.

In our previous studies of argumentation in complaint scenarios (Galitsky et al 2007, Galitsky et al 2008) we verified that using attack relationship in addition to communicative actions as a way to express dialogue discourse indeed increases the accuracy of scenario plausibility assessment in a similar setting to the current

study. In the current study, having showed the importance of both explanation levels, we proceeded to defining such characteristic parameter of scenarios with explanation behavior as *ratio* between contributions of each level to overall scenario assessment. We then demonstrated that using such measure a phase space can visualize scenarios with various forms of explanation activities by agents. We also showed that a number of various behaviors can be represented at a explanation phase space.

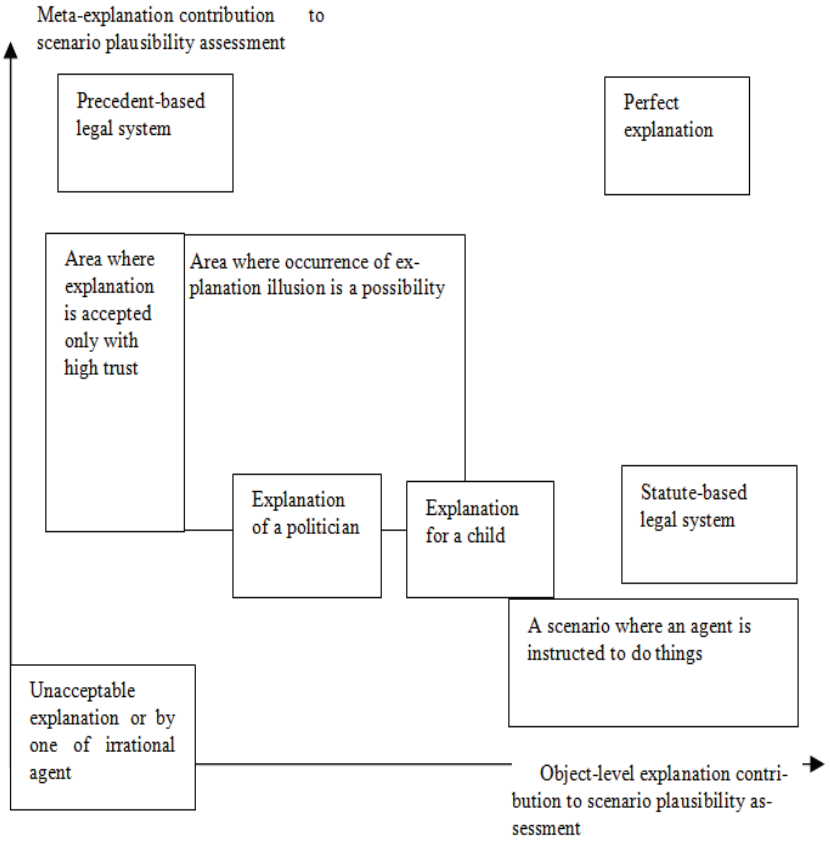


Fig. 3: Areas for specific forms of explanation behavior at the explanation phase space

Acknowledgements

This research is funded by the European Union project Num. 216746 PReservation Organizations using Tools in AGent Environments (PROTAGE), FP7-2007- ICT Challenge 4: Digital libraries and content, European Union project Num. 238887, *a unique European citizens' attention service* (iSAC6+) IST-PSP, SIE 2007 *Motor intelligent de recuperació i processament semàntic d'informació en base a criteris de personalització*, as well as RDiSAC *Recerca en incentivació de la participació 2.0 i llenguatge natural per contexts per a la creació de serveis no presencials d'informació i atenció ciutadana SAC*, a grant from ACC1O Catalan Government, as well as the CSI-ref.2009SGR-1202 consolidate research group.

References

Chesñevar, C., Maguitman, A. & Loui, R. Logical Models of Argument. ACM Computing Surveys 32(4) 337 383 (2000).

Cox, M., Raja, A. Metareasoning: a manifesto. BBN Technical Memo BBN TM 2028. http://www.mcox.org/Metareasoning/Manifesto/manifesto.pdf (2008).

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning and logic programming and n person games. Artificial intelligence, 77, 321 357.

Resconi, R., Kovalerchuk, B., Agents' model of uncertainty, Knowledge and Information Systems Journal, Springer London, vol. 18, no. 2, pp. 213 229, Feb 2009, DOI 10.1007/s10115 008 0164 0.

Fum, D., Missiera, F. D. and Stoccob, A., The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. Cognitive Systems Research v8 3, September 2007, 135 142.

Galitsky, B., Kovalerchuk, B. Kuznetsov, S.O. Learning Common Outcomes of Communicative Actions Represented by Labeled Graphs. ICCS , 387 400 (2007)

Galitsky, B. Reasoning about mental attitudes of complaining customers. Knowledge Based Systems Elsevier. Volume 19, Issue 7, 592 615, 2006.

Galitsky, B. Merging deductive and inductive reasoning for processing textual descriptions of inter human conflicts. J Intelligent Info Systems v27 N1 21 48 (2006b).

Galitsky, B., Gonzalez M.P., and Chesnevar C., Processing Customer Complaints Scenarios through Argument Based Decision Making. Decision Support Systems (2008).

García, A., Simari, G. Defeasible Logic Programming: an argumentative approach. Theory and Practice of Logic Programming 4(1), 95 138 (2004).

Parsons, S., Wooldridge, M., and Amgoud, L. An analysis of formal inter agent dialogues, Proceedings of the International Conference on Autonomous Agents and Multi Agent Systems, Bologna (2002).

Walton, D. 'Dialogical Models of Explanation' Explanation Aware Computing: Papers from the 2007 AAAI Workshop, Association for the Advancement of Artificial Intelligence, Technical Report WS 07 06, AAAI Press, 2007,1 9.

Walton, D. 'Can Argumentation Help AI to Understand Explanation?', Kunstliche Intelligenz, 22(2), 2008, 8 12.