Noname manuscript No.
(will be inserted by the editor)

Evaluation of a Trust-modulated Argumentation-based Interactive Decision-making Tool

Elizabeth I. Sklar · Simon Parsons · Zimi Li · Jordan Salvit · Senni Perumal · Holly Wall · Jennifer Mangels

Received: date / Accepted: date

Abstract The interactive *ArgTrust* application is a decision-making tool that is based on an underlying formal system of argumentation in which the evidence that influences a recommendation, or conclusion, is modulated according to values of trust that the user places in that evidence. This paper presents the design and analysis of a user study which was intended to evaluate the effectiveness of ArgTrust in a collaborative human-agent decision-making task. The results show that users' interactions with ArgTrust helped them consider their decisions more carefully than without using the software tool.

Keywords Argumentation · Trust · Human-Agent Interaction

E. I. Sklar
Department of Computer Science, University of Liverpool, Ashton Street, Liverpool, UK E-mail: e.i.sklar@liverpool.ac.uk

S. Parsons
Department of Computer Science, University of Liverpool, Ashton Street, Liverpool, UK E-mail: s.d.parsons@liverpool.ac.uk

Z. Li
Department of Computer Science, Graduate Center, City University of New York, 365 Fifth Avenue, New York, NY, USA E-mail: zimili.sjtu@gmail.com

J. Salvit
Department of Computer Science, Graduate Center, City University of New York, 365 Fifth Avenue, New York, NY, USA E-mail: jordan@jordansalvit.com

S. Perumal
Raytheon BBN Technologies, 10 Moulton Street Cambridge, MA, USA E-mail: senni.peri@gmail.com

H. Wall
Department of Computer Science, Graduate Center, City University of New York, 365 Fifth Avenue, New York, NY, USA E-mail: holly.e.wall@gmail.com

J. Mangels
Department of Psychology, Baruch College, City University of New York, 55 Lexington Avenue, New York, NY, USA E-mail: jennifer.mangels@baruch.cuny.edu

1 Introduction

Argumentation [61] is an approach to reasoning in which the focus is not just on the conclusions that are reached, but also on the data from which the conclusions are inferred, and the inference steps involved. Argumentation has technical advantages over other logic-based approaches to reasoning, particularly in its ability to handle inconsistent information, but also seems to be a mechanism that fits well with the way that human reasoning is carried out. For example, Mercier and Sperber [45] argue that the formation of reasons for and against conclusions is a fundamental part of human reasoning, while Walton and Krabbe [76] cast a large part of human interaction in the form of argumentation-based dialogue.

Argumentation has a long history in *artificial intelligence (AI)*, going back at least as far as 1980 [9], and has a significant, if shorter, history in *agent-based systems*. As early as the 1990's [40, 67], argumentation was suggested as a mechanism to extend *negotiation* between agents from the simple exchange of offers to a process that allows one agent to persuade another to change its position. This led to work on argumentation to underpin joint planning [53][1], and then more general approaches to argumentation-based dialogue that could capture a range of dialogue types [5, 56, 58, 59]. In work on negotiation between autonomous agents, the assumption was that the use of argumentation would lead to agreements that could not be reached by other means, and this was empirically verified by [57]. Subsequent work has shown that argumentation also has advantages in other forms of dialogue between autonomous agents [16, 39]. Given the fact that argumentation appears to be a natural way for humans to reason, and that fact that argumentation is beneficial in agent-agent interactions, an obvious question is: *what is the effect of using argumentation in human-agent interactions?* That is the question on which we focus in this paper.

This is not the first paper to examine matters related to this question. For example see the papers in [38]. However, this paper looks at the use of argumentation in human-agent interaction in a novel context: one in which the human and agent reason collaboratively, using argumentation as the medium through which they represent their beliefs and thought processes. The *"agent"* in the system described here employs formal argumentation for reasoning about a scenario, and the human is given the same scenario. The agent presents its conclusions and relevant evidence (i.e., its arguments) to the user, who can indicate his/her level of agreement with the agent's beliefs. In the version of the system presented here, there is no dialogue about their beliefs; though future work will explore such further levels of interactivity.

In systems of argumentation in which arguments are constructed from logical statements [4, 8, 19, 46], an important feature is the way in which elements of the arguments—the premises and rules from which they are constructed—have a bearing on the quality of the arguments. Premises may be undermined

[1] Called "negotiation" in [53], but much closer to what [76] calls "deliberation".

and hence defeated. Conclusions may be rebutted, and rules themselves may be undercut. This relationship between the parts and the whole, combined with the relationship between the trust individuals place in information and the provenance of that information [21], led us to suggest the use of argumentation in situations where trust in information is critical [52, 55]. The key idea is that since argumentation tracks the data used in deriving conclusions, if that data could be related to the sources from which it comes, information about those sources could be used in reasoning about the conclusions.

We developed a formal argumentation system [68] that allows information about sources—represented in the form of the "trust networks" that are standard in the literature of reasoning about trust—to be combined with arguments. This formal system was initially implemented in an inference engine called *ArgTrust* [54]. The version of ArgTrust evaluated here is intended as a prototype for an intelligent interactive agent that can collaborate with a user in making decisions that involve complex situations involving the analysis of data from a range of sources, not all of which can be fully trusted, and which change continuously. In the work presented here, we report on a user study designed to explore how effective *ArgTrust* is in supporting human decision-making and, since *ArgTrust* interacts with the human by providing arguments, how effective argumentation is for human-agent communication. In particular, the aim of the user study was to gather information about how people reason, how they make decisions in uncertain situations, and how they explain their decisions. Participants (i.e., human subjects) used *ArgTrust* to help them visualise a scenario and make sense of information presented that describes elements of the scenario in different ways.

The remainder of this paper is structured as follows. We start in Section 2 with a brief description of the *ArgTrust* system, and pointers to papers in which the reader can obtain more detail. Then, Section 3 describes the design of the user study, giving full details of the materials given to participants in the study. Section 4 gives the results of the study. Related work is highlighted in Section 5, and then Section 6 concludes.

2 ArgTrust

This section briefly describes ArgTrust and the underlying formal model.

2.1 Theoretical basis

The formal argumentation system [68] that underpins ArgTrust starts with the idea that we want to represent the beliefs of a set of individuals, Ags, where each $Ag_i \in Ags$ has access to a knowledge base, Δ_i, containing formulae in some language $\mathcal{L}$. An *argument* is then:

Definition 1 (Argument) An *argument* A from a knowledge base $\Delta_i \subseteq \mathcal{L}$ is a pair (G, p) where p is a formula of $\mathcal{L}$ and $G \subseteq \Delta_i$ such that:

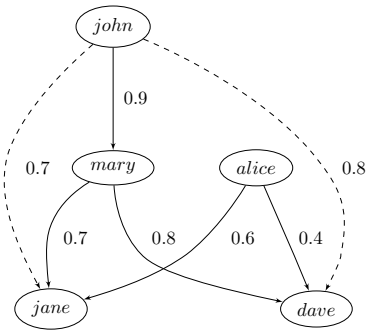


Fig. 1 Social network. Trust is propagated using *TidalTrust* (see text).

1. G is consistent;
2. $G \vdash p$; and
3. G is minimal, so there is no proper subset of G that satisfies the previous conditions.

G is called the *grounds* of A, written $G = Grounds(A)$ and p is the *conclusion* of A, written $p = Conclusion(A)$. Any $g \in G$ is called a *premise* of A. The key aspect of argumentation is the association of the grounds with the conclusion, in particular the fact that we can trace conclusions to the source of the grounds.

The particular language $\mathcal{L}$ we use is the language of defeasible Horn clauses—that is, a language in which formulae are either atomic propositions p_i or formulae of the form $p_i \wedge \ldots \wedge p_n \Rightarrow c$, where $\Rightarrow$ is a defeasible rule rather than material implication. Inference in this system is by a defeasible form of generalised *modus ponens* (DGMP):

$$\frac{p_1, \ldots, p_n \qquad p_i \wedge \ldots \wedge p_n \Rightarrow c}{c} \tag{1}$$

and if p follows from a set of formulae G using this inference rule alone, we denote this by $G \vdash p$. Given its use of defeasible Horn cluases, this argumentation system is related to that of [19].

The set of individuals, Ags, are related to each other by a social network that includes estimates of how much individual agents trust their acquaintances, as illustrated in Figure 1. Nodes represent individuals and links between them are annotated with the degree to which one individual trusts another, represented as values between 0 and 1. The input to the network (i.e., information known *a priori*) consists of the nodes and the solid edges. The output of the network (dashed edges) is the degree of trust inferred between any two nodes in the network. We can, for example, apply *TidalTrust* [20] to propagate trust values through the network and relate agents that are not directly connected in the social network.

In decision-making situations, argumentation can help in two ways. First, it is typical that from the data a given individual Ag_i has about a situation, we can construct a set of arguments that may conflict with each other. We

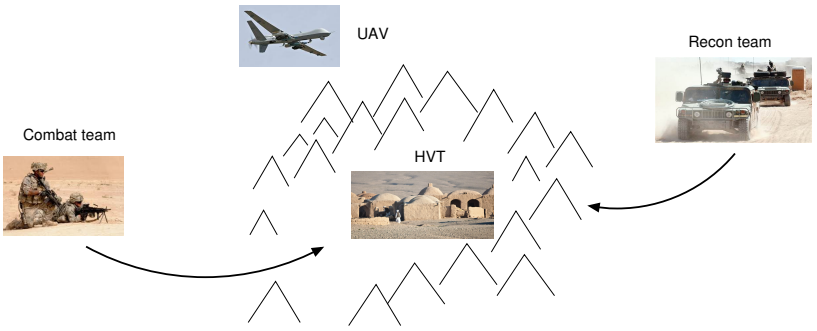


Fig. 2 The example scenario

might have an argument (G, p) in favour of some decision option, and another argument $(G', \neg p)$ against it (in this case, we say that the arguments *rebut* each other). We might also have a third argument $(G'', \neg g)$ where $g \in G$ is one of the grounds of the first argument (in this case we say that $(G'', \neg g)$ *undermines* (G, p)). Finally, we might have a fourth argument $(G''', \neg i)$ where i is one of the conclusions to one of the defeasible rules in (G, p). (This is another form of rebut, rebuttal of a sub-argument.) Argumentation provides a principled way—or rather a number of alternative ways—for Ag_i to establish which of a conflicting set of arguments it is most reasonable to *accept* [7].

Second, the grounds of an argument G, can be related back to the sources of the information that constitutes the grounds. If that information comes from some individual Ag_j that Ag_i knows, then Ag_i will *believe* the information according to how much they trust Ag_j (an extension of Liau's [43] principle that you believe information from individuals that you trust). The same principle can be applied to other sources of information[2]. This weight can be used to resolve conflicts between arguments. It is possible to provide the decision maker with links between information that feeds into a decision and the source of that information, allowing them to explore the effect of trusting particular sources.

To see more concretely how this can be useful, let's look at a simple decision-making example, loosely based on Operation Anaconda [48] and depicted in Figure 2. In this example, a decision is being made about whether to carry out an operation in which a combat team will move into a mountainous region to try to apprehend a high value target (HVT) believed to be in a village in the mountains.

We have the following information:

1. If there are enemy fighters in the area, then an HVT is likely to be in the area.

[2] For example, military intelligence traditionally separates information into that which comes from human sources, that which comes from signals intercepts, and that which comes from imagery. All of these sources can be rated with some measure of trustworthiness.

2. If there is an HVT in the area, and the mission will be safe, then the mission should go ahead.
3. If the number of enemy fighters in the area is too large, the mission will not be safe.
4. UAVs that have flown over the area have provided images that appear to show the presence of a significant number of camp fires, indicating the presence of enemy fighters.
5. The quality of the images from the UAVs is not very good, so they are not very trusted.
6. A reconnaissance ("recon") team that infiltrated the area saw a large number of vehicles in the village that the HVT is thought to be inhabiting.
7. Since enemy fighters invariably use vehicles to move around, this is evidence for the presence of many enemy fighters.
8. Informants near the combat team base claim that they have been to the area in question and that a large number of fighters are present.
9. In addition, we have the default assumption that missions will be safe, because in the absence of information to the contrary we believe that the combat team will be safe.

Thus there is evidence from UAV imaging that sufficient enemy are in the right location to suggest the presence of an HVT. There is also some evidence from informants that there are too many enemy fighters in the area for the mission to be safe.

We might represent this information as follows (the numbers in parentheses indicate the correspondence between the logic representations, below, and the relevant piece(s) of information, above)[3]:

$$
\begin{aligned}
(1)\ & InArea(enemy) \Rightarrow HVT \\
(2)\ & HVT \wedge Safe(mission) \Rightarrow Proceed(mission) \\
(3)\ & InArea(enemy) \wedge Many(enemy) \Rightarrow \neg Safe(mission) \\
(4,5)\ & InArea(campfires) \\
(4)\ & InArea(campfires) \Rightarrow InArea(enemy) \\
(6)\ & InArea(vehicles) \\
(7)\ & InArea(vehicles) \Rightarrow Many(enemy) \\
(7,8)\ & Many(enemy) \\
(9)\ & Safe(mission)
\end{aligned}
$$

[3] While stressing that this is purely illustrative — a real model of this example would be considerably more detailed.

From this information, we can construct arguments such as:

$$\left(\left\{ \begin{array}{l} InArea(campfires), \\ InArea(campfires) \Rightarrow InArea(enemy), \\ InArea(enemy) \wedge Safe(mission) \Rightarrow HVT, \\ Safe(mission), \\ HVT \Rightarrow Proceed(mission) \end{array} \right\}, Proceed(mission) \right)$$

which is an argument for the mission proceeding, based on the fact that there are campfires in the area, which suggest enemy fighters, that enemy fighters suggest the presence of an HVT, and that the presence of an HVT (along with the default assumption that the mission will be safe) suggests that the mission should go ahead. The level of belief in this argument will depend on the trust in the source of the information from which the argument is constructed. Since the crucial information in the argument, the presence of the campfires, is derived from the UAV imaging, trust in the UAV imaging will determine the belief in the argument.

We can build other arguments from the available information, and, since these will conflict, then compute a subset that are *acceptable*. (Approaches to this computation are discussed in [7].) In this case, we can use information from the informants to build an argument that there are many enemies in the area and hence the mission will not be safe:

$$\left(\left\{ \begin{array}{l} InArea(vehicles), \\ InArea(enemy), \\ InArea(vehicles) \Rightarrow Many(enemy) \\ InArea(enemy) \\ \quad \wedge Many(enemy) \Rightarrow \neg Safe(mission) \end{array} \right\}, \neg Safe(mission) \right)$$

This conflicts with the previous argument by undermining the assumption about the mission being safe. The belief in this second argument will depend, again, on the trust in the sources of the information from which the argument is constructed. In this case, the crucial information is that from the informants. Since information from informants is trusted less than that from the UAV[4], the level of belief in this second argument will be less than that in the first argument. Following [3], we use the degree of belief in an argument to determine which arguments are *defeated*, and in this case the first argument is not defeated by the second. This, in turn means that the first argument is acceptable.

The relation between trust in the source of an argument, defeat between arguments and the computation of acceptability is explored in more detail in in [55].

[4] In this scenario, because informants are paid for useful information, they are widely considered to simply make up plausible information with the result that it is considered to be untrustworthy, and certainly less trustworthy than information derived from the high-resolution imaging from a UAV.

2.2 Implementation

An initial version ($v1.0$) of ArgTrust was described in [69]. Here we present some aspects of a more recent version, $v2.0$, which was used for the user study. Like $v1.0$, this current version takes as input an XML file in a format which we sketch here. First, we have a specification of how much sources of information are trusted, for example:

```
<trustnet>
 <agent> recon </agent>
    ...
  <trust>
   <truster> me </truster>
   <trustee> recon </trustee>
   <level> 0.95 </level>
 </trust>
    ...
</trustnet>
```

which specifies the individuals involved (including "me", the decision maker) and the trust relationships between them, including the level of trust (specified as a number between 0 (no trust) and 1 (completely trustworthy)). The current implementation uses these values to compute the trust that one agent places on another using a choice of TidalTrust [20] or the mechanism described in [78].

The XML file also contains the specification of each individual's knowledge, for example:

```
<beliefbase>
 <belief>
   <agent> recon </agent>
   <fact> enemy_in_area </fact>
   <level> 0.9 </level>
 </belief>
     ...
 <belief>
  <agent> me </agent>
    <rule>
     <premise> many_enemy </premise>
     <conclusion> not safe </conclusion>
    </rule>
    <level> 1.0 </level>
 </belief>
    ...
</beliefbase>
```

Here the numbers reflect the belief each individual has in its information about the world.

From this data, and a query about a particular proposition, ArgTrust constructs arguments for that proposition by backward chaining. Once these arguments have been constructed, ArgTrust examines each formula used in the derivation of these arguments to identify if there are arguments with conclusions that attack these formulae. Each formula in those attacking arguments are then examined in turn. (And so on.) Once the full set of arguments is constructed, trust and belief are used to establish which arguments are defeated (as sketched above), and the grounded semantics [12][5] are applied to establish acceptability, and the conclusions labelled IN, OUT or UNDEC [7].

ArgTrust $v2.0$ extends the previous version [70, 68] by implementing a more robust and flexible data model. ArgTrust $v2.0$ uses a MySQL database and the Python programming language (for reasons outlined below), in place of Java (which was employed for ArgTrust $v1.0$). The language choice was largely made in order to simplify the recursive methods for storing the data and traversing it in different ways and because it was desirable to develop an interactive front-end that could be executed in a standard web browser. In a MySQL[6] database, we maintain arguments as a set of trees that represent the logical steps needed to arrive at the argument's conclusion. Thus, to return to our Operation Anaconda example, the combination of premise

$$InArea(campfires)$$

with rule

$$InArea(campfires) \Rightarrow InArea(enemy)$$

to infer conclusion

$$InArea(enemy)$$

would be represented as a tree in which each of the above formulae was a node, and arcs led from premise to rule to conclusion. The representation allows us to easily overlap arguments that share predicates or rules. Thus, if we had another argument with conclusion $InArea(enemy)$, we would represent the two arguments together as a tree with a single conclusion node.

Another important piece of the data model is its flexibility to receive new attributes and easily facilitate reconstructing arguments for the conclusions at hand. For example, our experience is that users each have different senses of what "very trustworthy" means. Therefore, we built the system in such a way that changing values to belief levels or trust levels does not require completely reloading the scenario, instead entails just changing a parameter value.

The underlying ArgTrust inference engine can be invoked in four different modes: (1) as a command-line tool; (2) as a visualisation tool; (3) as an interactive decision support agent (the mode evaluated here); and (4) as a back-end reasoning engine. In command-line tool mode, a user can load an XML file, modify its contents on the ArgTrust command line, and pose queries to the

[5] The latest version of ArgTrust at the time of writing implements all the common semantics.

[6] `http://www.mysql.com`

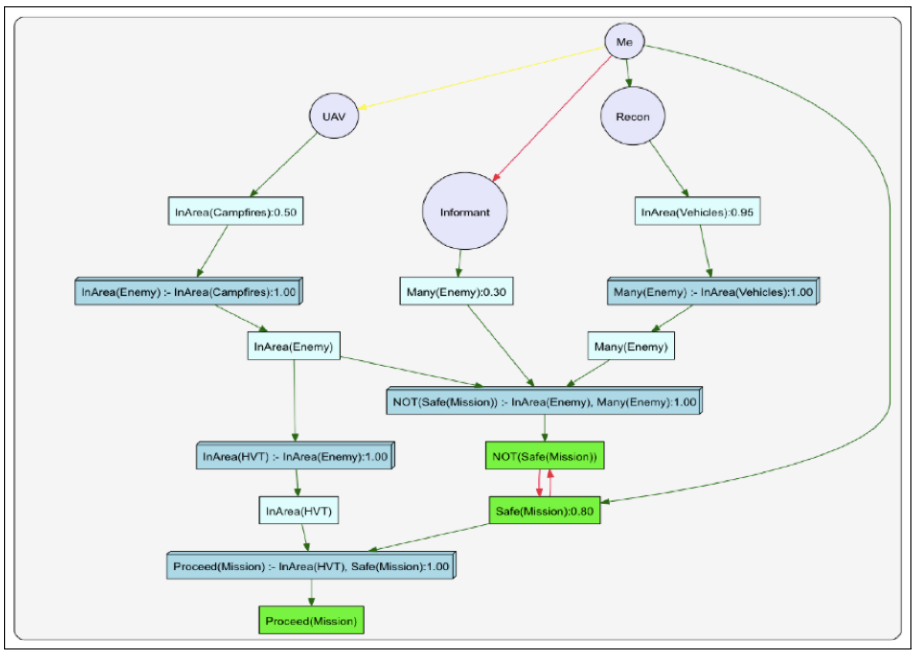


Fig. 3 *ArgTrust* screen: A high-level view of the argument graph

inference engine. The system responds by outputting text that reports the status of arguments supporting the query. In visualisation tool mode, the system produces output in a graphical display of the resulting arguments—here the result of inference is an *argument graph* (see below) like that in Figure 3. In interactive agent mode, users collaboratively step through a *decision scenario* and analyse it interactively, with help from the agent. In back-end reasoning engine mode, ArgTrust is called by another program—which might itself have an interactive front-end. Input is in the form of an XML file, as with the previous three modes; and output is also presented in the form of an XML file, where the burden of communicating the content of the output to a human user becomes the responsibility of the calling program. An example of this mode has been implemented and tested in related work involving a human-robot environment [6].

In visualisation and interactive agent modes, ArgTrust makes use of *argument graphs* to visualise complex scenarios and assign probabilities to all the possible outcomes. These graphs, which are distinct from the attack graphs common in the literature (also often called "argument graphs"[7]), represent the relationship between the facts and rules that make up the arguments, and the relationships between the arguments themselves. A full explanation of the

[7] See, for example, Figure 1 in [75].

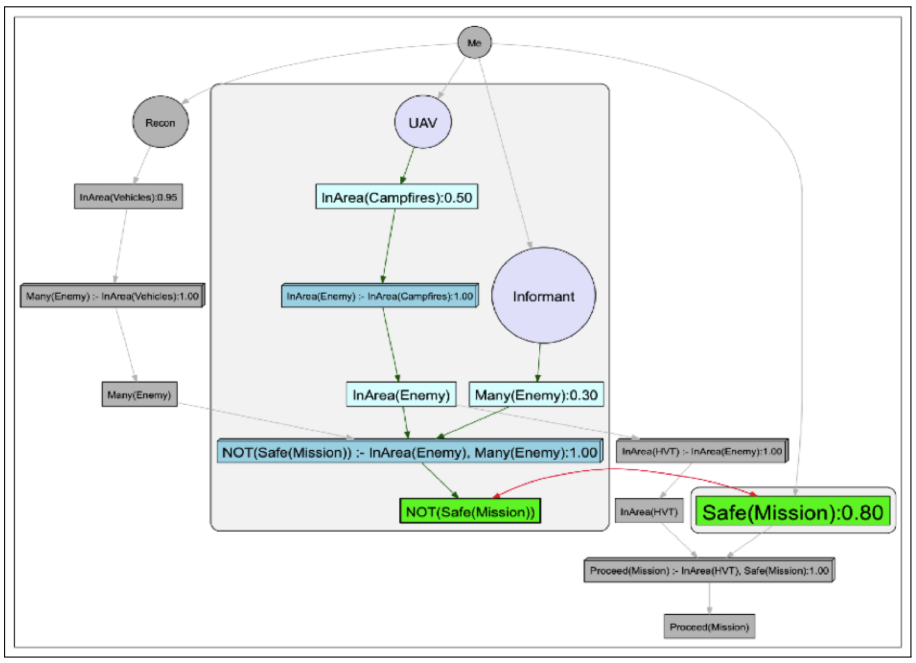


Fig. 4 *ArgTrust* screen: A more focussed view of one argument for the mission being unsafe

graphs can be found in [68], along with the translation into graph-theoretic terms of the usual ideas of *extension* and the *acceptability* of arguments.

The next section, below, describes the user interface developed for the interactive decision support mode. Then, Sections 3 and 4 describe a user study designed to evaluate the effectiveness of this mode.

2.3 User interface

The interactive decision support mode of ArgTrust $v2.0$ includes an interface which allows users to manipulate the argument graphs at different levels of detail, and to focus on individual components of an argument, in order to better understand a scenario in its entirety and reach an informed conclusion. This mode was developed for the user study, and is intended to be used in a context where both the text of a narrative describing an scenario and the arguments describing a scenario are both to be presented to a test subject.

When used in this mode, there are four main components to the ArgTrust interface, each accessed by clicking on the following *tabs* in the user interface:

– **Review Scenario**: This component shows the text-based narrative describing a scenario and allows users to read through it before progressing. This tab is persistent, allowing users to revisit the scenario text at any time.

- **Review Trust/Beliefs/Rules**: These three tabs allow a user to change the belief level for any belief or rule, and the levels of trust in individual agents. The corresponding sentence in the scenario is displayed as well, facilitating the user's ability to set the level more accurately. After setting a belief level, the user can navigate to one of the last two tabs to see how it impacts the argument. This process can be iterative, as the user understands and learns their own process for defining belief and trust levels.
- **Trust+Beliefs**: In this part of the system, the combination of an agent's trust values and belief values are displayed to illustrate the belief value for the decision maker in a particular proposition. The goal is to fill the logical gap between setting trust values and belief values by displaying the combination of both.
- **Argument Graph**: This component displays the argument graph that corresponds to the scenario. Scenarios are broken down into arguments which are built from bits of knowledge (e.g., facts or evidence), rules (or beliefs), and the resulting conclusions. Facts and rules are linked to individuals (sources of information or agents). Arrows connect facts, rules and conclusions together to form a chain. A chain is referred to as an argument. (A chain can have only one arrow linking two nodes, or multiple arrows linking more than two nodes.) Each argument ends in a conclusion, and every conclusion is assigned a belief. The user can control the amount of information displayed in the graph by selecting "zoom level" and "detail" options and "focus".
  —*Zoom-level and Detail-level* controls: Located in the upper right-hand corner of the argument graph panel are the zoom and detail controls. The zoom buttons allow users to visually zoom in and out of the graph (i.e., magnifying the visual display, but not changing the content). The detail slider allows users to adjust the level of detail displayed in the graph (i.e., changing the content to be more refined or more abstract). At the highest level of detail (most abstract), only the conclusions and their corresponding beliefs are displayed. Alternatively, at the lowest level of detail (most refined), all sources, i.e., agents, beliefs, facts, rules are shown.
  —*Focus* controls: The focus feature, located in the sidebar, enables users to focus on individual arguments of a graph. The graph updates to highlight the chosen conclusion or piece of knowledge, allowing users to focus in on that particular piece of the scenario.

3 User study design

We conducted a user study designed to evaluate the effectiveness of the ArgTrust interactive decision support mode. A primary goal of the study was to provide a preliminary assessment of the impact of ArgTrust on users' decision-making processes. Two scenarios were developed for the user study, including narratives and logical representations of information contained in each narrative, such as in the example outlined in Section 2.1. One scenario is short, relatively

simple and was created as a training exercise; the other scenario is longer, more complex and was built as an evaluation exercise.

The user study procedure involved multiple steps. First, participants were asked to provide demographic (e.g., gender and age) and background information (e.g., education, level of experience working with computers and decision-making tools) by filling out a Pre-survey. These data were collected for statistical purposes in order to describe the population of human subjects and to satisfy reporting requirements of funding agencies. Then participants completed a short training exercise, using the short and simple scenario mentioned above and shown in Figure 5, to familiarise them with a formal notion of decision making under uncertainty and to give them a preliminary experience using ArgTrust.

> Your grandparents are coming to visit you in New York City, and they are arriving at the airport shortly. They get anxious when visiting big cities, so you promised to meet them at the airport and escort them to their hotel. You had planned to take the train to the airport straight from work. Right before you planned to leave, your co-worker tells you there was an earlier incident at a station and that train line is experiencing delays. You text a friend, who you know lives near that train line, to confirm. Your friend tells you that she left her house at the usual time and arrived to work on time, without experiencing any problems with the train. **Do you risk taking the train, which may be delayed, or do you take a taxi instead (more expensive, but quicker)?**

Fig. 5 Narrative of short, simple training scenario: "Grandparents Scenario"

Next, participants were presented with a text-based narrative describing a more complex scenario (the longer scenario, mentioned above and shown in Figure 6). They were asked to analyse the details of the scenario and come to a decision about an action to take with respect to the scenario. Some of the questions required users to simply repeat information given in the scenario; this was intentional, to ensure that users had carefully read and understood the text. Once they made their decision, they were then asked to report on why and how they made that decision, via an on-line Mid-survey. Participants were asked to provide as much detail as possible regarding their thought processes.

Finally, participants were given the same scenario and asked to reconsider their decision, this time with the aid of ArgTrust. The input to ArgTrust was an XML file with contents that we extracted manually from the scenario in Figure 6. Participants were asked to employ the user interface to interact with ArgTrust and explore the data describing the scenario, and then report on how they utilised the software in their decision-making process, by completing an on-line Post-survey. The study took approximately 60 minutes to complete.

We make a few comments on the design of the user study. First, it is likely that a *learning effect* took place between the Mid-survey and Post-survey, because participants answered questions about the same scenario in both surveys. A different study design might have had users explore two different scenarios: one for the Mid-survey (without using ArgTrust) and one for the Post-survey (after using ArgTrust). This might control better for learning effect, with re-

A week ago, a powerful earthquake struck Brax causing widespread devastation to the country's infrastructure and leaving over 10,000 dead and over 50,000 injured. The two cities in Brax that were hit the hardest are Waga and Tapel. The Braxian Government and the UN have requested global assistance to launch the largest humanitarian relief operation ever executed. The Braxian Military, with its extensive and modern military force and airlift capability, is leading the effort and coordinating the international response. You are an Intelligence Analyst at your desk in the Operations Center of the main Forward Operating Base (FOB) in Tapel, monitoring the flow of data and reports coming in related to conditions, casualties and relief requirements. You have direct communications with the other FOB location in Waga, which was likewise affected by the earthquake. There are two rebel insurgent cells operating in the region: Reds and Lions. Each one is vying for power with the population, local and national politicians. Each one is seeking to take advantage of the situation to consolidate their political positions and establish local control with their rebel militia forces. The rebel militia forces have access to only small arms weapons and limited explosives. The rebel militias are stirring up the local population to protest the incompetence of the Braxian government. Braxian military forces are now stretched thin, trying to defend against the rebel militia forces while, at the same time, leading humanitarian rescue efforts in the wake of the earthquake. It has been 6 days since the earthquake hit Brax. Your Army Commander has asked you to answer the following Priority Intelligence Requirement (PIR): Which rebel militia cell is encouraging the most violence against the Braxian government? You have the following information (the order of the items listed is arbitrary):

- The Braxian Military reports that they have encountered many attacks/incidents of violence involving Red rebels and only some incidents of violence involving Lion rebels.
- Many incidents of violence by a rebel group imply that it is creating/encouraging much violence whereas some incidents of violence by a rebel group imply that it is not creating much violence.
- Sources of information include: Braxian Department of State, Braxian First Responders, Braxian Officials, Braxian Civilians, International Civilians and Open Media (like newspapers). Collectively, these sources of information reports only few incidents each, which makes information from them incomplete.
- Twitter feeds are inundated with reports of violence which are often contradictory. Twitter feeds are not considered very reliable.
- The Braxian Military reconnaissance reports that they have seen lots of vehicles outside the Lion Headquarters both in Tapel and in Waga.
- The presence of large number of vehicles outside a rebel militia headquarters can indicate that the rebel militia is planning many attacks/incidents of violence on relief personnel.
- Members of the rebel Lion militia who are paid by the Braxian government to inform on their comrades indicate that they have been directed to increase violence and use small arms against the Braxian military.
- A rebel group that may be planning many attacks as well as directing its members to increase violence could be a group that will create much violence.

You have to decide which rebel militia the Braxian Military efforts should focus on defending against.

Fig. 6 Narrative of longer, more complex evaluation scenario: "Humanitarian Relief Scenario"

spect to participants' knowledge of the scenario (though not with respect to participants' knowledge of the software tool). Second, a further study design might attempt to control for *order effect*: half of the participants work without ArgTrust to answer questions about the first scenario and with ArgTrust to answer questions about a second scenario; and the other half of the partici-

pants use ArgTrust to answer questions about the first scenario and without the software for the second scenario. This might control for participants liking the software tool better because they work with it after struggling without it, or vice versa. Future studies will consider other designs such as these.

4 User study results

The user study was conducted in three sessions, where each session was conducted in a different location and involved a different set of participants. This division occurred purely due to logistics with regard to scheduling multiple sessions in which to accommodate sufficient numbers of participants. However, as mentioned below, the three groupings led to interesting distinctions with respect to the analysis of the results. **Group I** consisted of psychology undergraduate and graduate students, and the session was conducted in a university computer lab setting. **Group II** consisted of computer science undergraduate and graduate students, and the session was conducted in a university computer lab setting. **Group III** consisted of technical employees of an engineering research company, and the session was conducted in a corporate conference room. Each session followed the same procedure (outlined in Section 3), although the first group did not complete the Mid-survey (due to unforeseen logistical problems that occurred during the session).

This section discusses the results obtained by the three surveys (Pre-survey, Mid-survey and Post-survey), followed by comparative analysis across the surveys, especially the relationships between answers on the Pre- and Post-surveys, and on the Mid- and Post-surveys. Questions from the Mid-survey and Post-survey are analysed in four categories: A. questions about facts (i.e., reading comprehension and paying attention to misleading questions); B. questions about applying rules found in the scenario; C. questions about trust of information sources; and D. questions about conclusions drawn from the information provided in the scenario.

4.1 Pre-survey

Twenty-two (22) participants completed the user study. Basic demographics are shown in Table 1. All participants were well-educated: 12 participants had a Masters Degree or above, and 8 had a PhD. Nobody reported previous experience with computer-based decision making tools, but almost everyone (21 of 22) claimed prior experience with data management tools, such as Microsoft Excel (20 out of 22). Only 2 of the 22 participants indicated on the Pre-survey that they had previously encountered the concepts of “logical argumentation” or “argumentation graphs”.

We collected the demographic data with two hypotheses in mind. The first hypothesis was that people whose education and experience was non-technical (versus technical) would find ArgTrust harder to work with. The

			gender				*age*			
group	*count*		*female*		*male*		*18-24*		*25-39*	
everyone	**22**	(100%)	9	(41%)	13	(59%)	6	(27%)	16	(73%)
Group I	6	(27%)	6	(100%)	0	(0%)	2	(33%)	4	(67%)
Group II	7	(32%)	1	(14%)	6	(86%)	4	(57%)	3	(43%)
Group III	9	(41%)	2	(22%)	7	(78%)	0	(0%)	9	(100%)
Tech	12	(55%)	2	(17%)	10	(83%)	2	(17%)	10	(83%)
Non-Tech	10	(45%)	7	(70%)	3	(30%)	4	(40%)	6	(60%)
English	17	(77%)	7	(41%)	10	(59%)	5	(29%)	12	(71%)
Non-English	5	(23%)	2	(40%)	3	(60%)	1	(20%)	4	(80%)

	ethnicity							
group	*Asian*		*Black*		*Latino*		*White*	
everyone	7	(32%)	1	(5%)	2	(9%)	12	(55%)
Group I	2	(33%)	0	(0%)	0	(0%)	4	(67%)
Group II	3	(43%)	1	(14%)	1	(14%)	2	(29%)
Group III	2	(22%)	0	(0%)	1	(11%)	6	(67%)
Tech	4	(33%)	0	(0%)	1	(8%)	7	(58%)
Non-Tech	3	(30%)	1	(10%)	1	(10%)	5	(50%)
English	4	(24%)	1	(6%)	2	(12%)	10	(59%)
Non-English	3	(60%)	0	(0%)	0	(0%)	2	(40%)

	academic major				*native language*			
group	*Tech*		*Non-Tech*		*English*		*Non-English*	
everyone	12	(55%)	10	(45%)	17	(77%)	5	(23%)
Group I	0	(0%)	6	(100%)	4	(67%)	2	(33%)
Group II	3	(43%)	4	(57%)	5	(71%)	2	(29%)
Group III	9	(100%)	0	(0%)	8	(89%)	1	(11%)
Tech	12	(100%)	0	(0%)	9	(75%)	3	(25%)
Non-Tech	0	(0%)	10	(100%)	8	(80%)	2	(20%)
English	9	(53%)	8	(47%)	17	(100%)	0	(0%)
Non-English	3	(60%)	2	(40%)	0	(0%)	5	(100%)

Table 1 User study participants. There were 22 human subjects in total. The table shows participants broken down into study session, undergraduate major subject and native language groups.

second hypothesis was that people who were non-native (versus native) English speakers would find ArgTrust more helpful for understanding the subtleties in the narrative (which was presented in English). Table 1 tallies the number of participants in each session group who majored as *undergraduates* in Technical subjects (Computer Science, Information Technology or Engineering), as well as the number of participants who are native English speakers (at least, who speak English at home). As will be discussed below, our analysis of the results collected in the study indicates that these are relevant groupings of participants for highlighting differences in the impact and effectiveness of interacting with ArgTrust for making decisions.

Our first hypothesis when designing the study was that people who had no particular experience in technical subjects would find ArgTrust harder to learn how to work with, but more useful, as compared to people who had been trained in technical subjects such as Computer Science, Information Technology and/or Engineering. So we asked for participants to include information

on the pre-survey about their favourite subject(s) when they were in high school and the academic subject they majored in as college undergraduates. The results showed that 12 (55%) of all participants in the study majored as undergraduates in Computer Science, Information Technology and/or Engineering; thus, the analysis (below) considers the participants grouped according to **Technical** majors and **Non-Technical** majors in order to evaluate this hypothesis.

Our second hypothesis when designing the study was that people who were non-native English speakers would find ArgTrust helpful in constructing reasoning that involved subtleties of the language used in the narrative (English). So we asked participants to include information on the pre-survey about the language that they speak at home ("What language(s) do you and your parents speak at home?"). Although 17 (77%) of participants speak English at home, many of the participants also speak another language; in fact, 15 (68%) of participants speak languages other than English at home. Overall, 7 (32%) people indicated that they live in bi-lingual households, and one person indicated that they speak four languages at home. The analysis (below) considers the participants grouped according to **English** speaking and **Non-English** speaking homes in order to evaluate this hypothesis.

We also asked participants about computer usage in the Pre-survey, in order to whether their level of familiarity with technology. The results are shown in Table 2. These indicate that all members of the cohort involved in the study were quite familiar with computers and technology in general, regardless of whether they had studied a technical subject as an undergraduate and regardless of their native language. Use of social media was more varied across the cohort. As a result, the analysis that follows does not attempt to derive any correlations between computer usage habits and results with respect to understanding of the test scenario or ease-of-use interacting with ArgTrust.

4.2 Mid-survey

The Mid-survey was designed to reflect users' understanding of the scenario after only reading the text narrative (i.e., before interacting with ArgTrust). Some sample questions are discussed here. Note that the participants from Group I did not complete the Mid-survey, due to logistical issues when the study was administered at that site, so only 16 participants (Groups II and III) completed the Mid-survey. Percentages reported in this section are thus computed out of 16 instead of 22 users.

Users responded to six multiple choice questions (with possible answers of TRUE, FALSE and Inconclusive), followed by an indication of their confidence in their answer, ranging from 1 (least confident) to 10 (most confident). Users also responded to questions about their *trust* in the informants depicted in the scenario and the *likelihood* that an intermediate conclusion drawn from some evidence provided in the scenario is valid. Finally, users responded to the ultimate question posed in the scenario narrative:

group	*several times per day*		*at least once per day*		*several times per week*		*infrequently*		*never*	
How often do you use a computer? (i.e., laptop or desktop)										
everyone	**20**	**(91%)**	1	(5%)	1	(5%)	0	(0%)	0	(0%)
Group I	4	(67%)	1	(17%)	1	(17%)	0	(0%)	0	(0%)
Group II	7	(100%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
Group III	9	(100%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
Tech	12	(100%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
Non-Tech	8	(80%)	1	(10%)	1	(10%)	0	(0%)	0	(0%)
English	15	(88%)	1	(6%)	1	(6%)	0	(0%)	0	(0%)
Non-English	5	(100%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
How often do you use a computer-based device? (other than a laptop or desktop)										
everyone	**20**	**(91%)**	1	(5%)	1	(5%)	0	(0%)	0	(0%)
Group I	6	(100%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
Group II	6	(86%)	1	(14%)	0	(0%)	0	(0%)	0	(0%)
Group III	8	(89%)	0	(0%)	1	(11%)	0	(0%)	0	(0%)
Tech	11	(92%)	0	(0%)	1	(8%)	0	(0%)	0	(0%)
Non-Tech	9	(90%)	1	(10%)	0	(0%)	0	(0%)	0	(0%)
English	15	(88%)	1	(6%)	1	(6%)	0	(0%)	0	(0%)
Non-English	5	(100%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
How often do you use social media? (e.g., Facebook, Twitter)										
everyone	6	(27%)	2	(9%)	6	(27%)	7	(32%)	1	(5%)
Group I	2	(33%)	1	(17%)	2	(33%)	0	(0%)	1	(17%)
Group II	3	(43%)	1	(14%)	1	(14%)	2	(29%)	0	(0%)
Group III	1	(11%)	0	(0%)	3	(33%)	5	(56%)	0	(0%)
Tech	2	(17%)	1	(8%)	3	(25%)	6	(50%)	0	(0%)
Non-Tech	4	(40%)	1	(10%)	3	(30%)	1	(10%)	1	(10%)
English	5	(29%)	1	(6%)	5	(29%)	5	(29%)	1	(6%)
Non-English	1	(20%)	1	(20%)	1	(20%)	2	(40%)	0	(0%)

Table 2 Participants' computer usage habits.

You have to decide which rebel militia the Braxian Military efforts should focus on defending against.

along with their confidence in their answer.

A. Mid-Survey Questions about Facts

The multiple choice questions were designed to determine how carefully the users read the narrative text and how well they understood the scenario. Specifically, we are looking for whether users extracted facts (or *evidence*, to use the *argumentation* terminology) and implications associated with the facts (or *rules*, again using the *argumentation* terminology). Table 3 tallies the responses to Mid-survey questions about facts presented in the scenario.

In answer to the question:

The Red rebel militia is operating in the area devastated by an earthquake.

none of the users answered FALSE. Only just over half felt confident with a TRUE answer (63%) and the rest were unable to reach a conclusion (38%).

Close examination of the narrative produces two sentences that clearly relate to this question:

> A week ago, a powerful earthquake struck Brax...
> There are two rebel insurgent cells operating in the region: Reds and Lions.

But there is indeed ambiguity with respect to the phrase "in the region"—does "region" refer to the same area where the earthquake has occurred?

In contrast, in response to question:

> Many vehicles have been seen near Lion headquarters in Waga.

all the users answered TRUE, and their confidence was high (9.06). Referring back to the narrative, one of the bulleted items states:

> – The Braxian Military reconnaissance reports that they have seen lots of vehicles outside the Lion Headquarters both in Tapel and in Waga.

The unanimous consensus could stem from two reasons: first, readers tend to pay more attention to details that are enumerated in lists; and second, users trusted the Braxian Military as the source of this information more than other sources (such as informants).

The first reason (readers pay attention to details presented in lists) is backed up by the results shown in response to the question:

> Twitter feeds are unreliable sources of information.

The following excerpt from the narrative provides evidence for the answer:

> – Twitter feeds are inundated with reports of violence which are often contradictory. Twitter feeds are not considered very reliable.

Here, most users provided a TRUE answer (75%), though a few could not draw a conclusion. One participant answered FALSE, which we could interpret as an indication that that person did not read the narrative carefully or fully comprehend the details; however, this person also indicated in the Pre-survey that they used social media several times a day, which may bias his/her answer. Note that there were other participants who also indicated that they use social media several times a day but still answered TRUE to this question about the unreliability of Twitter.

Some questions were designed to be misleading with respect to the facts presented in the scenario. Consider the statement from the Mid-survey:

> Vehicles belonging to the Tiger rebel militia have been seen near the Braxian Military headquarters.

This statement indicates a group called "Tiger", which does not exist in the scenario and was included as a contrast to the previous question which asks about "Lion"—attempting to highlight whether users confused the names "Lion" and "Tiger". More than half the users entered FALSE (63%), indicating that most were paying attention, one user entered TRUE, indicating that they were not paying attention; as well, a non-trivial number of users (5) thought the information was Inconclusive. This could either be because they didn't read the

group	TRUE		FALSE		Inconclusive		*confidence*
The Red rebel militia is operating in the area devastated by an earthquake.							
everyone	10	(63%)	0	(0%)	6	(38%)	8.12 (1.63)
Group II	5	(71%)	0	(0%)	2	(29%)	7.57
Group III	5	(56%)	0	(0%)	4	(44%)	8.56
Tech	7	(58%)	0	(0%)	5	(42%)	8.08
Non-Tech	3	(75%)	0	(0%)	1	(25%)	8.25
English	8	(62%)	0	(0%)	5	(38%)	8.31
Non-English	2	(67%)	0	(0%)	1	(33%)	7.33
Many vehicles have been seen near Lion headquarters in Waga.							
everyone	16	(100%)	0	(0%)	0	(0%)	9.06 (0.93)
Group II	7	(100%)	0	(0%)	0	(0%)	9.57
Group III	9	(100%)	0	(0%)	0	(0%)	8.67
Tech	12	(100%)	0	(0%)	0	(0%)	8.83
Non-Tech	4	(100%)	0	(0%)	0	(0%)	9.75
English	13	(100%)	0	(0%)	0	(0%)	9.00
Non-English	3	(100%)	0	(0%)	0	(0%)	9.33
Twitter feeds are unreliable sources of information.							
everyone	12	(75%)	1	(6%)	3	(19%)	8.12 (1.78)
Group II	5	(71%)	1	(14%)	1	(14%)	8.86
Group III	7	(78%)	0	(0%)	2	(22%)	7.56
Tech	9	(75%)	0	(0%)	3	(25%)	8.08
Non-Tech	3	(75%)	1	(25%)	0	(0%)	8.25
English	9	(69%)	1	(8%)	3	(23%)	8.00
Non-English	3	(100%)	0	(0%)	0	(0%)	8.67
Vehicles belonging to Tiger rebels have been seen near the Brax Military headquarters.							
everyone	1	(6%)	10	(63%)	5	(31%)	9.12 (1.20)
Group II	1	(14%)	4	(57%)	2	(29%)	8.86
Group III	0	(0%)	6	(67%)	3	(33%)	9.33
Tech	0	(0%)	8	(67%)	4	(33%)	9.25
Non-Tech	1	(25%)	2	(50%)	1	(25%)	8.75
English	1	(8%)	7	(54%)	5	(38%)	9.00
Non-English	0	(0%)	3	(100%)	0	(0%)	9.67
The Lions have access to chemical weapons.							
everyone	0	(0%)	7	(44%)	9	(56%)	9.00 (1.10)
Group II	0	(0%)	5	(71%)	2	(29%)	8.86
Group III	0	(0%)	2	(22%)	7	(78%)	9.11
Tech	0	(0%)	5	(42%)	7	(58%)	9.08
Non-Tech	0	(0%)	2	(50%)	2	(50%)	8.75
English	0	(0%)	5	(38%)	8	(62%)	9.00
Non-English	0	(0%)	2	(67%)	1	(33%)	9.00

Table 3 Responses to Mid-survey questions about facts. Standard deviation is shown for averages across all participants (in parentheses).

scenario carefully, or because they thought that there might really be another rebel group called Tiger, but that the scenario did not provide any conclusive evidence about the Tiger group.

The statement:

The Lions have access to chemical weapons.

is in clear contraction to the following statement in the narrative:

> The rebel militia forces have access to only small arms weapons and limited explosives.

Nonetheless, less than half the users (44%) answered FALSE, while the remainder could not draw any conclusion (56%); and confidence overall was high (9.00).

B. Mid-survey Questions about Applying Rules

One question attempts to measure how well the users applied rules presented in the scenario:

> How likely is it that vehicles outside the Lion militia HQ indicate that the group is planning an attack?

which puts together this fact:

> The Braxian Military reconnaissance reports that they have seen lots of vehicles outside the Lion Headquarters...

and this rule:

> The presence of large number of vehicles outside a rebel militia headquarters can indicate that the rebel militia is planning many attacks...

from the narrative. The results (see Table 4) show that users did not assimilate this information well from reading the scenario, because the mean answer (and standard deviation) were 6.75 (1.61), on a scale of 1 (not likely at all) to 10 (very likely).

group	*average likelihood (1-10)*
How likely is it that vehicles outside the Lion militia headquarters indicate that the group is planning an attack?	
everyone	6.75 (1.61)
Group II	7.00
Group III	6.56
Tech	7.00
Non-Tech	6.00
English	6.31
Non-English	8.67

Table 4 Responses to Mid-survey questions about applying rules, ranging from 1 (not likely at all) to 10 (very likely).

C. Mid-survey Questions about Trust

One question attempts to measure how much users *trust* sources of information:

> How much do you trust the paid Lion militia informants?

The mean answer (and standard deviation) were 5.69 (1.66), on a scale of 1 (not much at all) to 10 (very much). Results are tallied in Table 5.

group	*average level of trust (1-10)*
How much do you trust the paid Lion militia informants?	
everyone	5.69 (1.66)
Group II	6.00
Group III	5.44
Tech	5.58
Non-Tech	6.00
English	5.85
Non-English	5.00

Table 5 Responses to Mid-survey questions about trust, ranging from 1 (not much trust) to 10 (very much).

D. Mid-survey Questions about Conclusions

Finally, one question asks about users' conclusions with respect to the overall predicament posed by the narrative:

> Which rebel militia is the most violent, implying that Braxian Military efforts should focus on defending against attacks from that insurgent group (as opposed to other insurgents)?

The responses are tallied in Table 6. Results are fairly evenly split (44 : 56) between the two groups (*Lions*) and (*Reds*), and confidence is moderate (7.19)—lower with respect to questions about facts (which averaged 8.7 confidence measures).

group	Lions		Reds		*confidence*
Which rebel militia is the most violent?					
everyone	7	(44%)	9	(56%)	7.19 (1.17)
Group II	4	(57%)	3	(43%)	7.71
Group III	3	(33%)	6	(67%)	6.78
Tech	5	(42%)	7	(58%)	6.83
Non-Tech	2	(50%)	2	(50%)	8.25
English	5	(38%)	8	(62%)	7.31
Non-English	2	(67%)	1	(33%)	6.67

Table 6 Responses to Mid-survey questions about conclusion drawn.

4.3 Post-survey

The Post-survey is divided into two parts. The first part contains a series of questions that, like the Mid-survey, are designed to reflect how well the users understand the scenario and are supported in making decisions with the information presented. The second part contains a number of questions that are designed to assess the users' impressions of ArgTrust, independent from

the particular scenario. Here, we discuss both parts. Comparative analysis of answers given in the Post-survey and Mid-survey is deferred to Section 4.3.2. We begin, next, by presenting the Post-survey responses.

4.3.1 Post-survey, Part 1

The first part of the Post-survey was designed to reflect users' understanding of the scenario after interacting with ArgTrust. Some sample questions are discussed here. Percentages reported in this section are computed out of 22 users, since all users completed the Post-survey.

Similarly to the Mid-survey, users responded to four multiple choice questions (with possible answers of TRUE, FALSE and Inconclusive), as well as indicating their confidence in their answer, ranging from 1 (least confident) to 10 (most confident). Users also responded to questions about their *trust* in the informants depicted in the scenario and the *likelihood* that an intermediate conclusion drawn from some evidence provided in the scenario is valid. Finally, users responded to the ultimate question posed in the scenario narrative:

> You have to decide which rebel militia the Braxian Military efforts should focus on defending against.

along with their confidence in their answer.

A. Post-survey Questions about Facts

As with the Mid-survey, the multiple choice questions in the Post-survey were designed to determine how carefully the users read the narrative text and how well they understood the scenario, in particular after interacting with ArgTrust to work with the facts and rules contained in the narrative. The responses are tallied in Table 7.

The following statement was contained in the Post-survey:

> Information from the Braxian Military reconnaissance team indicates that there are many rebel militia forces in the areas devastated by the earthquake.

Evidence supporting this observation is contained in the scenario narrative:

> A week ago, a powerful earthquake struck Brax...
> There are two rebel insurgent cells operating in the region: Reds and Lions.

which can reasonably derive a TRUE response to the question. As Table 7 indicates, more than half of participants (64%) responded with TRUE. However, most of the rest (7 people, 32%) were unable to draw a conclusion.

Another Post-survey question:

> Vehicles being seen near a rebel headquarters implies that rebels are planning an attack.

paraphrases an item from the narrative:

– The presence of large number of vehicles outside a rebel militia headquarters can indicate that the rebel militia is planning many attacks/incidents of violence on relief personnel.

However, a literal interpretation of the narrative phrase "can indicate" is less conclusive than the phrasing in the question: "implies". Perhaps this is why more users responded to the question with an Inconclusive answer, as shown in Table 7. Here we note that all hesitation with respect to this fact was reported by those who speak English at home.

Confidence with respect to questions about facts was lower than the Mid-survey, 7.78 on average, versus 8.7 average confidence on the Mid-survey.

group	TRUE		FALSE		Inconclusive		*confidence*
Information from the Braxian Military reconnaissance team indicates that there are many rebel militia forces in the areas devastated by the earthquake.							
everyone	14	(64%)	1	(5%)	7	(32%)	7.64 (1.99)
Group I	2	(33%)	1	(17%)	3	(50%)	6.50
Group II	6	(86%)	0	(0%)	1	(14%)	8.14
Group III	6	(67%)	0	(0%)	3	(33%)	8.00
Tech	9	(75%)	0	(0%)	3	(25%)	8.25
Non-Tech	5	(50%)	1	(10%)	4	(40%)	6.90
English	11	(65%)	0	(0%)	6	(35%)	7.88
Non-English	3	(60%)	1	(20%)	1	(20%)	6.80
Vehicles being seen near a rebel headquarters implies that rebels are planning an attack.							
everyone	10	(45%)	1	(5%)	11	(50%)	7.91 (1.57)
Group I	3	(50%)	1	(17%)	2	(33%)	7.67
Group II	3	(43%)	0	(0%)	4	(57%)	8.57
Group III	4	(44%)	0	(0%)	5	(56%)	7.56
Tech	6	(50%)	0	(0%)	6	(50%)	7.92
Non-Tech	4	(40%)	1	(10%)	5	(50%)	7.90
English	5	(29%)	1	(6%)	11	(65%)	8.00
Non-English	5	(100%)	0	(0%)	0	(0%)	7.60

Table 7 Responses to Post-survey questions about facts. Standard deviation is shown for averages across all participants (in parentheses).

B. Post-survey Questions about Applying Rules

As with the Mid-survey, one question attempts to measure how well the users applied rules presented in the scenario:

How likely is it that vehicles outside the Lion militia HQ indicate that the group is planning an attack?

Results are shown in Table 8. The mean answer (and standard deviation) were 6.91 (1.72), on a scale of 1 (not likely at all) to 10 (very likely). This is an increase from the Mid-survey (which was 6.75), though not statistically significant.

group	*average likelihood (1-10)*
How likely is it that vehicles outside the Lion militia headquarters indicate that the group is planning an attack?	
everyone	6.91 (1.72)
Group I	7.50
Group II	6.86
Group III	6.56
Tech	6.83
Non-Tech	7.00
English	6.47
Non-English	8.40

Table 8 Responses to Post-survey questions about applying rules, ranging from 1 (not likely at all) to 10 (very likely).

C. Post-survey Questions about Trust

Again, as with the Mid-survey, one question attempts to measure how much users *trust* sources of information:

How much do you trust the paid Lion militia informants?

Results are tallied in Table 9. The mean answer (and standard deviation) were 5.36 (1.71), on a scale of 1 (not much at all) to 10 (very much). This is a small decrease from the Mid-survey (5.69), though not statistically significant. It is notable, however, that all participants *except* those from non-English speaking households lowered their level of trust between the Mid- and Post-surveys.

group	*average level of trust (1-10)*
How much do you trust the paid Lion militia informants?	
everyone	5.36 (1.71)
Group I	5.33
Group II	5.71
Group III	5.11
Tech	5.00
Non-Tech	5.80
English	5.29
Non-English	5.60

Table 9 Responses to Post-survey questions about trust, ranging from 1 (not much trust) to 10 (very much).

D. Post-survey Questions about Conclusions

Two of the multiple-choice questions in Part 1 of the Post-survey ask users to draw their own intermediate conclusions based on the narrative:

Red leaders have tweeted that they are planning an attack.

and

> Braxian Military should prioritize rescue operations over attacks on rebel militia.

The answers, tallied in Table 10 indicate that users were unable to draw their own intermediate conclusions. This is understandable, because relatively little information pertinent to either question was provided in the narrative text, and providing a TRUE or FALSE answer to either question would require the user to make unsubstantiated inferences. This is a positive effect of interacting with ArgTrust, because users are given the opportunity to modulate the facts and rules contained in the narrative according to their own beliefs and trust values. The act of working directly with the content forces users to be careful not to draw conclusions for which there is no supporting evidence in the scenario.

group	TRUE		FALSE		Inconclusive		*confidence*
Red leaders have tweeted that they are planning an attack.							
everyone	0	(0%)	8	(36%)	14	(64%)	8.50 (2.44)
Group I	0	(0%)	2	(33%)	4	(67%)	8.00
Group II	0	(0%)	4	(57%)	3	(43%)	8.57
Group III	0	(0%)	2	(22%)	7	(78%)	8.78
Tech	0	(0%)	4	(33%)	8	(67%)	8.92
Non-Tech	0	(0%)	4	(40%)	6	(60%)	8.00
English	0	(0%)	4	(24%)	13	(76%)	8.71
Non-English	0	(0%)	4	(80%)	1	(20%)	7.80
Braxian Military should prioritize rescue operations over attacks on rebel militia.							
everyone	5	(23%)	2	(9%)	15	(68%)	7.45 (1.92)
Group I	1	(17%)	0	(0%)	5	(83%)	7.50
Group II	4	(57%)	1	(14%)	2	(29%)	8.29
Group III	0	(0%)	1	(11%)	8	(89%)	6.78
Tech	2	(17%)	1	(8%)	9	(75%)	7.33
Non-Tech	3	(30%)	1	(10%)	6	(60%)	7.60
English	3	(18%)	2	(12%)	12	(71%)	7.41
Non-English	2	(40%)	0	(0%)	3	(60%)	7.60

Table 10 Responses to Post-survey questions about intermediate conclusions.

Finally, users are asked to draw a conclusion with respect to the overall predicament posed by the narrative:

> You have to decide which rebel militia the Braxian Military efforts should focus on defending against.

The results are displayed in Table 11. The responses lean heavily toward defending against the Reds militia (82%), though confidence is moderate (6.55). It is notable that confidence was lower for *all* groupings of participants and more sharply lowered by non-technical participants and participants from non-English speaking homes.

group	Lions		Reds		*confidence*
Which rebel group should the Braxian Military provide defense against?					
everyone	4	(18%)	18	(82%)	6.55 (1.63)
Group I	2	(33%)	4	(67%)	5.67
Group II	1	(14%)	6	(86%)	7.14
Group III	1	(11%)	8	(89%)	6.67
Tech	1	(8%)	11	(92%)	6.67
Non-Tech	3	(30%)	7	(70%)	6.40
English	2	(12%)	15	(88%)	6.94
Non-English	2	(40%)	3	(60%)	5.20

Table 11 Responses to Post-survey questions about conclusion drawn.

4.3.2 Changes from Mid-survey to Post-survey

This section highlights responses that changed between the Mid-survey and the Post-survey. Note that because (as mentioned earlier) only participants in Groups II and III completed the Mid-survey, percentages presented in this section are computed out of 16 instead of 22.

We report on three types of changes. The first type is the extent to which users changed their opinions about facts and/or rules presented in the scenario. The second type is the extent to which users changed their confidence in their answers. The third type is the extent to which users changed the number of Inconclusive answers to conclusive answers (i.e., TRUE or FALSE).

With respect to reflecting the facts and/or rules presented in the scenario, the following changes in opinion were detected:

- In response to questions about whether there are rebel groups in the same region as the earthquake, two (13%) users changed their opinions from FALSE to TRUE.
- Half the users (50%) also changed their opinion about how likely is it that vehicles outside the Lion militia HQ indicate that the group is planning an attack. The mean (and standard deviation) rose from 6.75 (1.61) to 6.69 (1.62).
- Half of the users (50%) changed how much they trust paid Lion militia informants, from a mean (and standard deviation) of 5.69 (1.66) to 5.38 (1.45).
- Five (31%) people changed their opinion about which group is most violent between the Mid-survey and Post-survey, and all changed from Reds to Lions. However, users' confidence in their answers to this question declined, from a mean (and standard deviation) of 7.19 (1.17) to 6.88 (1.59).

The quantitative differences recorded are not statistically significant, nonetheless, it is an important result that many users changed their opinions as a result of interacting with ArgTrust.

With respect to changed levels of confidence, the differences in confidence levels are shown in Table 12. Averaged across all users who completed both Mid- and Post-surveys, the confidence level *decreased* after working with ArgTrust.

Even considering averages within a number of groupings (Tech majors vs Non-Tech majors; native English speakers vs non-native English speakers; Groups II and III), all the averages showed a decline in confidence. When looking at individuals, the confidence level only increased for 2 out of 16 participants. Both of these individuals were native English speakers; only one was a Tech major. This result is unexpected. Again, although the quantitative values are not statistically significant, nonetheless, it is an important result that users' confidence levels declined.

group	*Mid-survey*		*Post-survey*		*change*
everyone	8.48	(0.70)	7.81	(0.90)	-0.67
Group II	8.57	(0.70)	8.14	(0.78)	-0.43
Group III	8.41	(0.72)	7.56	(0.94)	-0.86
Tech	8.45	(0.73)	7.82	(0.94)	-0.64
Non-Tech	8.57	(0.68)	7.80	(0.91)	-0.77
English	8.45	(0.70)	7.72	(0.91)	-0.73
Non-English	8.62	(0.81)	8.20	(0.87)	-0.42

Table 12 Differences in confidence levels, from Mid-survey to Post-survey. Negative *change* means that level of confidence decreased. Only data from participants who completed both Mid-survey and Post-survey is included.

With respect to changed number of Inconclusive answers to conclusive answers (i.e., TRUE or FALSE), results are shown in Table 13. For all users who completed both Mid- and Post-surveys, the percentage of multiple choice questions that were given Inconclusive answers *increased* between the Mid-survey and the Post-survey. The same holds true for each sub-group of users (Tech majors vs Non-Tech majors; from English vs non-English speaking homes). However, we note that the "Non-English" group was significantly less likely to provide an INCONCLUSIVE answer. While we cannot draw any statistically significant conclusions from this observation because our sample size is too small, we believe this trend warrants further investigation in future studies.

Table 13 Percentage of multiple choice questions that were answered Inconclusive. Positive *change* means level of inconclusiveness increased. Only data from participants who completed both Mid-survey and Post-survey is included.

group	*Mid-survey*		*Post-survey*		*change*
everyone	0.34	(0.20)	0.52	(0.28)	0.17
Group II	0.10	(0.18)	0.16	(0.27)	0.05
Group III	0.26	(0.24)	0.38	(0.35)	0.13
Tech	0.27	(0.24)	0.41	(0.35)	0.14
Non-Tech	0.19	(0.19)	0.29	(0.29)	0.10
English	0.34	(0.20)	0.52	(0.29)	0.17
Non-English	0.02	(0.08)	0.03	(0.12)	0.01

4.3.3 Post-survey, Part 2

Table 14 lists the fifteen questions contained in Part 2 of the Post-survey that aim to assess users' impressions of ArgTrust; Table 15 contains the responses. These questions are adapted from the *NASA Task Load Index (TLX)* [26, 27]. Figure 7 plots the mean and standard deviation computed across all 22 participants in the study. The bars labelled (a)–(r) correspond to the questions listed in Tables 14 and 15. Discussion follows.

(a)	How difficult was the scenario to understand?
(b)	How well would you say you understood the scenario BEFORE using ArgTrust?
(c)	How well would you say you understood the scenario AFTER using ArgTrust?
(d)	How much did the ArgTrust software help to visualize the scenario?
(e)	How mentally demanding was the scenario BEFORE using ArgTrust?
(f)	How mentally demanding was the scenario AFTER using ArgTrust?
(g)	How hard was it to make a decision?
(h)	How much did the ArgTrust software help with your decision making?
(j)	Did you think the ArgTrust system easy to use?
(k)	Overall, how helpful did you find the ArgTrust system?
(m)	How physically demanding was the session?
(n)	How hurried or rushed was the pace of the session?
(p)	How successful were you in accomplishing what you were asked to do?
(q)	How hard did you have to work to accomplish your level of performance?
(r)	How insecure, discouraged, irritated, stressed, and annoyed were you?

Table 14 Post-survey questions assessing users' impressions of ArgTrust

Table 16 lists the average changes in *understanding* and *mental demand*, respectively, as perceived by users before and after interacting with ArgTrust. Overall, the mean is positive in both cases, indicating that both level of understanding and mental demand *increased* as a result of using ArgTrust. However, the high standard deviations of both values render the results inconclusive. Figure 8 plots the changes for each individual who participated in the study. The values are sorted in ascending order, to preserve the anonymity of participants. These plots illustrate that the understanding increased for most users. Only 2 participants reported large decreases in understanding as a result of using ArgTrust, and both of these were in the Non-Technical group. Approximately a third of participants reported increase in mental demand as a result of using ArgTrust, though this is understandable because everyone was new to using the software and it is expected that they would have to think harder when using new software (versus not using it).

The second part of the Post-survey also posed the question:

> Did you know anything about logical argumentation before participating in this study?

which is essentially the same as the following question posed in the Pre-survey:

group	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
everyone	4.68	7.45	7.59	5.36	5.05	5.36	6.36	5.18
	(2.25)	(1.84)	(1.89)	(2.15)	(2.50)	(2.11)	(2.42)	(2.56)
Group I	3.83	7.83	7.00	4.00	5.17	7.00	7.33	4.67
Group II	5.00	7.86	8.57	6.29	5.43	5.00	5.14	6.14
Group III	5.00	6.89	7.22	5.56	4.67	4.56	6.67	4.78
Tech	5.42	7.25	7.75	5.50	4.75	4.83	6.42	4.92
Non-Tech	3.80	7.70	7.40	5.20	5.40	6.00	6.30	5.50
English	4.65	7.47	7.24	5.29	5.18	5.35	6.53	5.06
Non-English	4.80	7.40	8.80	5.60	4.60	5.40	5.80	5.60

group	(j)	(k)	(m)	(n)	(p)	(q)	(r)
everyone	6.50	5.23	2.50	3.05	4.05	4.86	3.59
	(2.69)	(2.41)	(2.18)	(1.94)	(2.44)	(2.17)	(2.65)
Group I	5.50	4.17	4.67	2.67	4.50	5.33	6.33
Group II	6.00	6.29	1.71	1.86	3.00	5.29	1.71
Group III	7.56	5.11	1.67	4.22	4.56	4.22	3.22
Tech	6.58	5.33	1.83	3.75	4.50	4.92	2.92
Non-Tech	6.40	5.10	3.30	2.20	3.50	4.80	4.40
English	6.82	4.94	2.18	3.29	4.12	4.71	3.47
Non-English	5.40	6.20	3.60	2.20	3.80	5.40	4.00

Table 15 Tallies of Post-survey questions assessing users' impressions of ArgTrust. Responses were given on a scale of 1 (least) to 10 (most).

group	*understanding*		*mental demand*	
everyone	0.14	(2.01)	0.32	(2.10)
Group I	-0.83	(3.31)	1.83	(2.71)
Group II	0.71	(0.95)	-0.43	(1.99)
Group III	0.33	(1.41)	-0.11	(1.27)
Tech	0.50	(1.31)	0.08	(1.56)
Non-Tech	-0.30	(2.63)	0.60	(2.67)
English	-0.24	(2.11)	0.18	(2.13)
Non-English	1.40	(0.89)	0.80	(2.17)

Table 16 Post-survey responses: average self-reported changes in understanding and mental demand

Did you know anything about "logical argumentation" or "argumentation graphs" before today?

Table 17 shows the results from both surveys. It is interesting to note that the number of "yes" responses to questions about knowledge of logical argumentation prior to the study increased from the Pre-survey to the Post-survey. In fact, 5 people changed their answer from "no" to "yes" and 2 people changed their answer from "yes" to "no". Unfortunately, none of the participants who changed their answer provided any explanation in their comments for why their answer changed. We speculate that the people who changed their answer from "no" to "yes" found that they knew about the concept but were unfamiliar with the term "logical argumentation", and that the people who changed their answer from "yes" to "no" found that the concept they thought of as

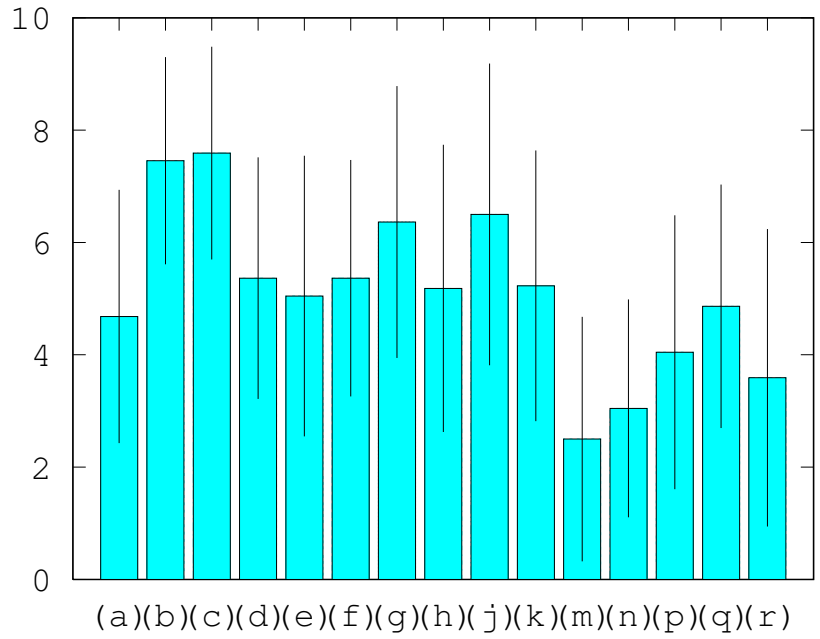


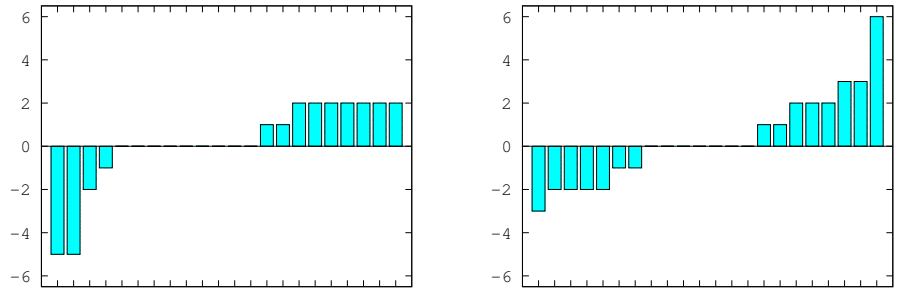
Fig. 7 Answers to Post-survey assessment questions. The key to the bar labels (a-r) is listed in Table 14.

"logical argumentation" was not aligned with the concept as presented in the study.

	Pre-survey				*Post-survey*			
group	yes		no		yes		no	
Did you know anything about logical argumentation before participating in this study?								
everyone	2	(9%)	20	(91%)	5	(23%)	17	(77%)
Group I	0	(0%)	6	(100%)	1	(17%)	5	(83%)
Group II	1	(14%)	6	(86%)	1	(14%)	6	(86%)
Group III	1	(11%)	8	(89%)	3	(33%)	6	(67%)
Tech	2	(17%)	10	(83%)	3	(25%)	9	(75%)
Non-Tech	0	(0%)	10	(100%)	2	(20%)	8	(80%)
English	1	(6%)	16	(94%)	5	(29%)	12	(71%)
Non-English	1	(20%)	4	(80%)	0	(0%)	5	(100%)

Table 17 Changes from Pre-survey to Post-survey with respect to prior knowledge of argumentation.

Finally, the Post-survey ends with a free-form comments question, asking users to provide feedback. Table 18 contains some of the free-form comments entered. Though the comments also contain some negative sentiments, we be-



(a) understanding (b) mental demand

Fig. 8 Post-survey responses: individual self-reported changes in understanding and mental demand. Values are sorted numerically, to preserve anonymity of individuals. Positive values indicate *increases* from "before" to "after".

lieve that these indicate that some users understand the benefits of structuring decisions in the way that we do in ArgTrust.

Table 18 Comments from participants

- The tool is successful in representing the factors that go into making a decision and displaying the relationships between those facts and how the influence a decision.
- It somewhat helped filter out the information that is not necessarily true.
- It helped me break the component information down a little bit but I had already created an outline of my own that helped me just as much if not more. Seeing the beliefs and the trust broken down was helpful, though.
- It helped to tell me what I'm supposed to think. . . how much I'm supposed to trust people, and how I was supposed to interpret the statements in the given scenario. However, it bothered me that a tool was telling me how to simplify a complex problem, since I don't believe the tool can possibly take all the details and subtleties into account. but if I accept that a complex situation can and must be simplified, then yes, the tool is helpful as a place to plug in parameters and let it do the math.
- By breaking down the situation into smaller bits and displaying how much you believe each situation to be true, it was much easier to make decisions because I was considering the situation asked only, not the entire situation.
- When I put how I felt into numbers, it organized and simplified my concerns and weighed all of the factors into the equation for me. It made it easier to see the results.

4.4 Analysis

In analysing the data, there are a number of interesting observations to make. There was no correlation found between participants' level of education or most other Pre-survey demographics to any of the Mid-survey or Post-survey responses. The factors that did correlate are academic major, split between

Tech and Non-Tech majors, and native language, split between native English speakers and non-native English speakers. The results are analysed below.

First, we look at the extent to which working with ArgTrust caused users to change their minds.

- *After using ArgTrust, users expressed less trust in paid Lion militia informants.*
- *After using ArgTrust, users believed that an attack by the Lion rebel group was less likely.*
- *After using ArgTrust, most users believed that the Reds rebel group was the most violent.*

Each of these findings confirms our working hypothesis that ArgTrust will have impact on users' decision-making processes.

Next, we look at differences in users' responses to the fifteen questions in Post-survey Part 2, which aim to measure users' perceptions about the effectiveness of interacting with ArgTrust. While Figure 7 shows the aggregate data across all users, Figures 9, 10 and 11 illustrate the results by looking at the users divided into different groupings. While there are no statistically significant quantitative differences, there are some interesting qualitative observations that seem important.

Specifically, we discuss users' perceived differences between *understanding* and *mental demand* before and after working with ArgTrust. The most distinct differences can be seen in Figure 9, bars `s` and `t`, respectively.

- With respect to *understanding*, the score for Group II participants rose from 7.86 to 8.57 and the score for Group III participants rose from 6.89 to 7.22. However, Group I reported a decrease in understanding from 7.83 to 7.00. Scores were reported on a scale of 1 (Very poor) to 10 (Very well). If we look at the numbers of participants reporting an increase/decrease in understanding rather than the average scores, we find that 3 people in Group I reported better understanding and 2 reported worse understanding, 3 people in Group II reported better understanding and none reported worse understanding, and 3 people in Group III reported better understanding and 2 reported worse understanding. Overall, that makes 9 reporting better understanding and 7 reporting worse understanding. Examining the groupings according to Tech majors vs Non-Tech majors, the score went up for Tech majors, but went down slightly for Non-Tech majors.

  Examining the groupings according to native English speakers vs non-native English speakers, there is a much greater increase in score for non-native English speakers. difference in users' perceived understanding after working with ArgTrust (`c`).
- With respect to *mental demand*, the results were more mixed. Groups II and III both showed a decrease in difficulty from pre-software to post-software (5.43 to 5.00 and 4.67 to 4.56, respectively). Group I, however, showed a large rise in difficulty (5.17 to 7.00). Scores were reported on a scale of 1 (Not demanding at all) to 10 (Very demanding).

If we again look at numbers of participants reporting, we find that 1 person from Group I reports the task is less demanding with the software and 4 report it is more demanding, and the figures for Group II and Group III are 4 (less demanding) and 2 (more) and 2 (less) and 2 (more) respectively. Thus, overall, we have 7 reporting it is less demanding and 8 reporting it is more demanding; amongst participants with technical backgrounds, more (6) reported that the task was less demanding when using the ArgTrust software than those who reported the task was more demanding (4).

Next, we examine users' perceived level of frustration when interacting with ArgTrust. Users were asked How insecure, discouraged, irritated, stressed, and annoyed were you? in the Post-survey, with responses ranging from 1 (Not annoyed or stressed at all) to 10 (Very annoyed and stressed). The average score for Group II was 1.71, and for Group III the score was 3.22. However, for Group I, the score was significantly higher: 6.33.

Other observations include:

- Group I users did not find the software helpful with respect to visualizing the scenario, as compared to Group II and III users (d).
- Group II found it much easier to make decisions than Group I and III users (g).
- Group III found the ArgTrust system easier to use than Group I and II users (j), while native English speakers found it easier to use than non-native English speakers.
- Non-native English speakers found the system more helpful overall (k).
- For some unknown reason, Group I users found it physically demanding (m) to use the system, which may tie in with their high level of frustration. Otherwise, there was no difference in how the study was implemented; there was no physical element in the study.
- Tech majors felt that they were more rushed or hurried than Non-Tech majors (n).
- Tech majors felt that they were more successful than Non-Tech majors (p).

In summary, the results of the user study provide evidence that participants who are presented with argumentation-based support for making decisions found that interacting with ArgTrust helped their understanding (both directly reported and inferred from the changes made in their decisions), albeit at the cost of increased difficulty in reaching a decision and decreased confidence in their decisions. Nonetheless, users were more able to make decisions after interacting with ArgTrust. Factors that seem to have an impact on users' experience include whether they are native English speakers or not, and whether they majored in Technical subjects as undergraduates or not. Other demographic factors (gender, age, ethnicity, level of education) did not play a part.

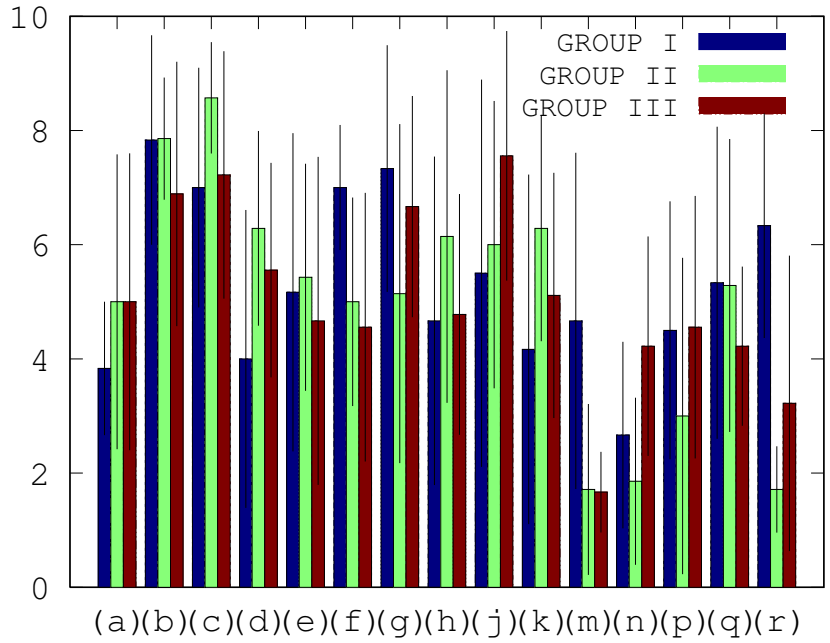


Fig. 9 Groups I, II and III. The key to the bar labels (a-r) is listed in Table 14.

5 Related work

There are four main areas of work that are related to the results reported here: modelling trust; reasoning about trust using argumentation; argumentation-based decision making; and the argumentation in interaction between humans and agents. We briefly cover each of these below stating the differences with our work.

5.1 Modelling trust

As computer systems have become increasingly distributed, and control of those systems has become more decentralised, computational approaches to trust have become steadily more important [23]. Some of this work has directly been driven by changes in technology, for example considering the trustworthiness of nodes in peer-to-peer networks [1, 15, 32], or dealing with wireless networks [22, 34, 65]. Other research has been driven by changes in the way that technology is used, especially involving the Internet. One early challenge is related to the establishment of trust in e-commerce [47, 62, 79], and the use of reputation systems to enable this trust [36, 37]. Another issue is the problem

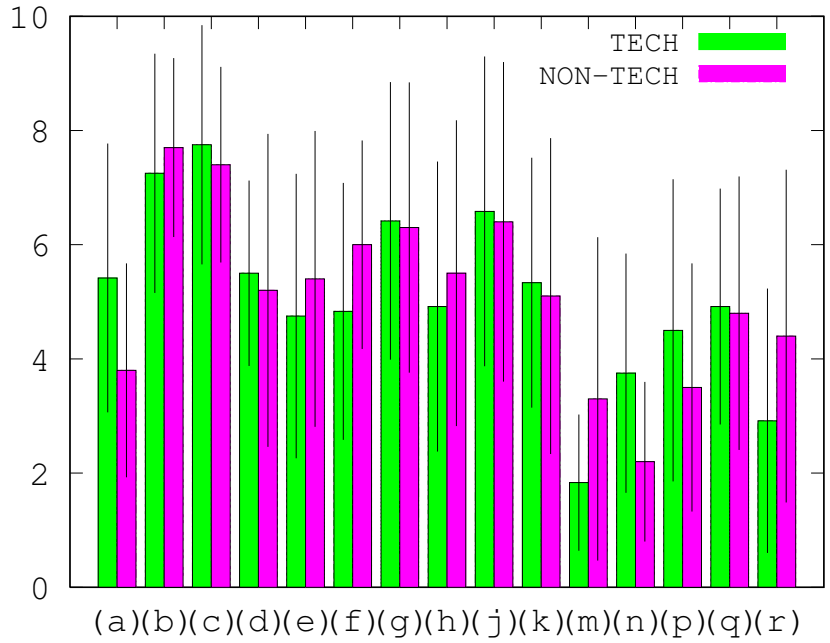


Fig. 10 Technical vs Non-Technical undergraduate majors. The key to the bar labels (a-r) is listed in Table 14.

of deciding which of several competing sources of conflicting information one should trust [2, 11].

Additional issues have arisen with the development of the social web, for example, the questions of how social media can be manipulated [41, 42] and how one should revise one's notions of trust based on the past actions of individuals [25]. In this area is some of the work that is most relevant for that we describe here, work that investigates how trust should be propagated through networks of individuals [24, 29, 35, 78], and we have drawn on this in our implementation of ArgTrust.

In all of this work, the focus is on the computation of the right numerical trust value to assign to individuals, and is often concerned with showing that the approach advanced in the paper is better than some other approach. In contrast, our work is concerned with how the values computed by these methods can be used in combination with argumentation, and *ArgTrust* provides a choice of ways of computing trust measures drawn from existing models.

5.2 Reasoning about trust using argumentation

The second area of work to consider is that which looks at the use of argumentation to handle trust. While the literature on trust is considerable, prior work

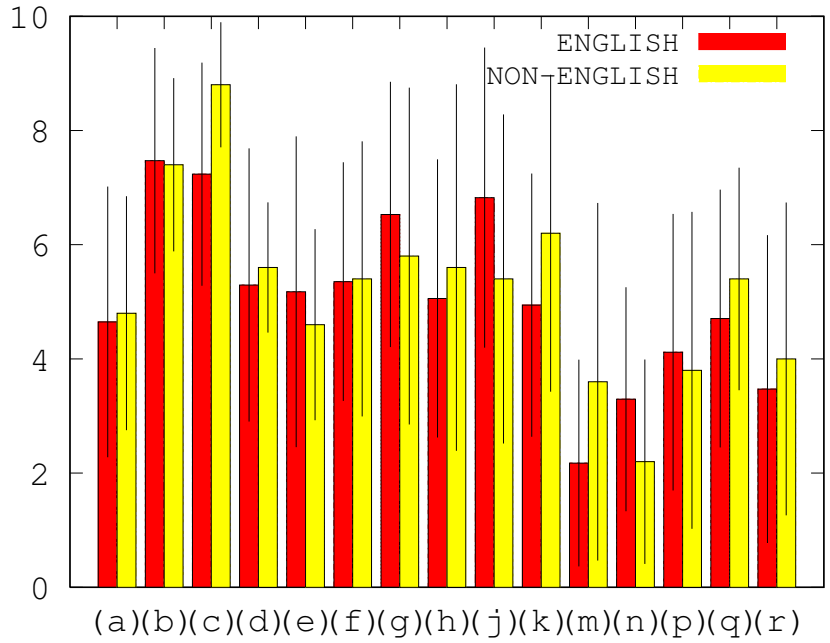


Fig. 11 Participants from English vs Non-English speaking homes. The key to the bar labels (a-r) is listed in Table 14.

on argumentation and trust is much more sparse. Existing work includes the application of argumentation techniques to networks of trust and distrust [28]; the combination of argumentation with fuzzy trust measures [64], with statistical data [44] and with subjective logic [49] (used in [29] to handle trust); the use of metalevel argumentation to describe trust [73]; and the description of a set of argument schemes for deriving trust [50]. However, none of this covers the same ground as the work reported here which is not concerned with the detail of how argumentation about trust is carried out, but how the results of the argumentation process is used by humans.

5.3 Argumentation-based decision making

The third area of work to consider is that on argumentation-based decision making. Here the work by Fox and colleagues [14, 17] showed that constructing arguments for and against a decision option, and then simply combining these arguments[8] could provide a decision mechanism that rivalled the accuracy of probabilistic models. This basic method was extended in [18, 51] to create a symbolic mechanism that, like classical decision theory, distinguished between

[8] Even the very straightforward mechanism of counting arguments for and against.

belief in propositions and the values of decision outcomes, while [30] showed the usefulness of arguments in communicating evidence for decision options to human users. More recent work on argumentation and decision making is described in [31, 71].

The relationship that we consider between argumentation and decision-making is different from all of this work. All the above work tries to build argumentation systems that identify the best decision to take. Even [30], which of the work in this area is closest to what we are doing, tries to identify the best decision and explain it to a human user in terms the human can understand. In work that evaluates the effectiveness of the systems, the aim is to show that the system gets the decisions right [14, 77]. Our focus, in contrast, is just to present information and test whether the users find the information to be useful—whether it helps them to feel more comfortable with their decisions, and whether they alter their opinion as a result of being able to use *ArgTrust* to visualise and manipulate the information on which they base their decisions.

5.4 Argumentation for human-agent interaction

A number of authors have examined whether software that presents arguments to users can help their argumentation skills. Work on assessing the impact of software that diagrams arguments has been carried out for several of the better known tools for visualizing arguments. For example, [10] tests the use of the Questmap system in teaching legal argumentation; [66] describes formative evaluation of the Belvedere system to support students in scientific argumentation; [33] examines the use of the TC3 tool to help in collaborative argumentation in the pursuit of a writing task; and [63] summarises an experiment to assess whether the "reasoner's workbench" *Convince Me* helps students, again in the area of scientific reasoning. All these papers find some support for the usefulness of the software, but this line of work also includes [72], which provides a sharp critique of much of the work listed above. Also noteworthy is [13], which investigates argument mapping—on paper, rather than using software—as a means for improving understanding of text, and finds that the presentation of arguments improves comprehension and recall.

Another line of related work is that which assesses whether human reasoning matches that captured in formal argumentation. While work such as [74], which looks at reasoning with defaults, is somewhat related, the only work we know of that explicitly examines the differences between argumentation theory and human reasoning is [60]. This paper looks at the question of reinstatement. Argumentation theory says that if an argument A is attacked by an argument B and that is all we know then A is OUT and B is IN. However, if B is then attacked by C, A and C are IN and B is OUT—A is reinstated by the attack that C makes on B. After reinstatement, the theory says that A is as strongly held as it was before any attack by B. The study in [60] reveals that while the human subjects understood that A should be reinstated by the attack from C on B—so they agreed with the general pattern of reinstatement—they did not

agree that A was held with the same force that it was held before the attack from B.

Of the work in this final area, [13] is the closest to ours, though differs in its normative focus—it is concerned with whether the students better understand and recall information when it is presented to them using argument diagramming—whereas our work is concerned with how users feel about decisions reached when using argumentation as opposed to decisions reached given only text.

6 Summary

We have described *ArgTrust*, an interactive application designed to help users balance information from multiple sources and draw conclusions from that information, using logical argumentation and a computational model of trust in information sources. The underlying inference engine is an implementation of earlier work with ArgTrust, here developed using a mySQL and Python back-end and a web-based front-end, largely written in PHP and HTML/CSS. This allows more flexibility with scenario input and the ability to transition to dynamic scenarios, one of the next steps with this research. As well, this architecture offers more options for user interface development.

The main contribution of this paper is the presentation and in-depth analysis of a user study, conducted to evaluate the effectiveness of ArgTrust in helping human users to make decisions. Users were presented with an ambiguous and complex scenario and instructed to formulate an evidence-backed recommendation for an action to take, based on information provided in the scenario. Participants completed Pre-, Mid- and Post-surveys that provided data about their backgrounds (demographics), and their understanding of the scenario before and after interacting with ArgTrust.

The results demonstrate, most importantly, that ArgTrust helped users consider their decisions more carefully. Analysis of the results provided other qualitative observations: differences in users' experiences with the application, with respect to difficulty understanding the scenario and effort expended in making decisions, were more closely aligned to the language users speak at home (English or not) and users' undergraduate majors (Technical or not), versus other demographic factors. An unexpected result is that users' questioned their answers more, after working with ArgTrust, indicated by their level of confidence in answers declining after using the application.

Future work involves extending the software to handle dynamic scenarios, applying the methodology to additional domains and testing with a wider range of users.

Acknowledgements This research was funded under Army Research Laboratory Cooperative Agreement Number W911NF-09-2-0053, by the National Science Foundation under grant #1117761 , and by the National Security Agency under the Science of Security Lablet grant (SoSL). Additional funding was provided by a University of Liverpool Research Fellowship and by a Fulbright-King's College London Scholar Award. The views and conclusions

contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funders. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. Z. Abrams, R. McGrew, and S. Plotkin. Keeping peers honest in EigenTrust. In *Proceedings of the 2nd Workshop on the Economics of Peer-to-Peer Systems*, 2004.
2. B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International World Wide Web Conference*, Banff, Alberta, May 2007.
3. L. Amgoud. *Contribution a l'integration des préferences dans le raisonnement argumentatif*. PhD thesis, Université Paul Sabatier, Toulouse, July 1999.
4. L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artifical Intelligence*, 34(3):197–215, 2002.
5. L. Amgoud, N. Maudet, and S. Parsons. Modelling dialogues using argumentation. In *Proceedings of the Fourth International Conference on Multi-Agent Systems*, pages 31–38, Boston, MA, USA, 2000. IEEE Press.
6. M. Q. Azhar, E. Schneider, J. Salvit, H. Wall, and E. I. Sklar. Evaluation of an argumentation-based dialogue system for human-robot collaboration. In *Proceedings of the Workshop on Autonomous Robots and Multirobot Systems (ARMS) at Autonomous Agents and MultiAgent Systems (AAMAS)*, St Paul, MN, USA, May 2013.
7. P. Baroni, M. Caminada, and M. Giacomin. An introduction to argumentation semantics. *The Knowledge Engineering Review*, 2011.
8. P. Besnard and A. Hunter. A logic-based theory of deductive arguments. *Artificial Intelligence*, 128:203–235, 2001.
9. L. Birnbaum, M. Flowers, and R. McGuire. Towards an AI model of argumentation. In *Proceedings of the 1st National Conference on Artificial Intelligence*, pages 313–315, Los Altos, CA, 1980. William Kaufmann.
10. C. S. Carr. Using computer supported argument visualization to teach legal argumentation. In P. A. Kirschner, S. J. Buckingham-Shum, and C. S. Carr, editors, *Visualizing argumentation: Software tools for collaborative and educational sense-making*, pages 75–96. Springer, 2003.
11. X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. In *Proceedings of the 35th International Conference on Very Large Databases*, Lyon, France, August 2009.
12. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

13. C. P. Dwyer, M. J. Hogan, and I. Stewart. An examination of the effects of argument mapping on students??? memory and comprehension performance. *Thinking Skills and Creativity*, 8:11–24, 2013.
14. J. Emery, R. Walton, A. Coulson, D. Glasspool, S. Ziebland, and J. Fox. Computer support for recording and interpreting family histories of breast and ovarian cancer in primary care (RAGs): Qualitative evaluation with simulated patients. *British Medical Journal*, 319(7201):32–36, 1999.
15. M. Feldman, C. Papadimitriou, J. Chuang, and I. Stoica. Free-riding and whitewashing in Peer-toPeer systems. In *Proceedings of the 3rd Annual Workshop on Economics and Information Security*, 2004.
16. S. P. Ferrando and E. Onaindia. Defeasible argumentation for multi-agent planning in ambient intelligence applications. In V. Conitzer, W. van der Hoek, L. Padgham, and M. Winikoff, editors, *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, Valencia, Spain, 2012. IFAAMAS.
17. J. Fox, A. Glowinski, C. Gordon, S. Hajnal, and M. O'Neil. Logic engineering for knowledge engineering: design and implementation of the oxford system of medicine. *Artificial Intelligence in Medicine*, 2(6):323–339, 1990.
18. J. Fox and S. Parsons. Arguing about beliefs and actions. In A. Hunter and S. Parsons, editors, *Applications of uncertainty formalisms*. Springer-Verlag, 1998.
19. A. J. García and G. Simari. Defeasible logic programming: an argumentative approach. *Theory and Practice of Logic Programming*, 4(1):95–138, 2004.
20. J. Golbeck. *Computing and Applying Trust in Web-based Social Networks*. PhD thesis, University of Maryland, College Park, 2005.
21. J. Golbeck. Combining provenance with trust in social networks for semantic web content filtering. In *Proceedings of the International Provenance and Annotation Workshop*, Chicago, Illinois, May 2006.
22. K. Govindan, P. Mohapatra, and T. F. Abdelzaher. Trustworthy wireless networks: Issues and applications. In *Proceedings of the International Symposium on Electronic System Design*, Bhubaneswar, India, December 2010.
23. T. Grandison and M. Sloman. A survey of trust in internet applications. *IEEE Communications Surveys and Tutorials*, 4(4):2–16, 2000.
24. R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International Conference on the World Wide Web*, 2004.
25. C.-W. Hang, Y. Wang, and M. P. Singh. An adaptive probabilistic trust model and its evaluation. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, Estoril, Portugal, 2008.
26. S. G. Hart. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 904–908, 2006.

27. S. G. Hart and L. E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139–183, 1988.
28. W. T. Harwood, J. A.Clark, and J. L. Jacob. Networks of trust and distrust: Towards logical reputation systems. In D. M. Gabbay and L. van der Torre, editors, *Logics in Security*, Copenhagen, Denmark, 2010.
29. A. Jøsang, R. Hayward, and S. Pope. Trust network analysis with subjective logic. In *Proceedings of the 29th Australasian Computer Society Conference*, Hobart, January 2006.
30. P. N. Judson, J. Fox, and P. J. Krause. Using new reasoning technology in chemical information systems. *Journal of Chemical Information and Computer Sciences*, 36:621–624, 1996.
31. A. Kakas and P. Moraitis. Argumentation based decision making for autonomous agents. In *2nd International Conference on Autonomous Agents and Multi-Agent Systems*, New York, NY, 2003. ACM Press.
32. S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The EigenTrust algorithm for reputation management in P2P networks. In *Proceedings of the 12th World Wide Web Conference*, May 2004.
33. G. Kanselaar, G. Erkens, J. Andriessen, M. P. andA. Veerman, and J. Jaspers. Designing argumentation tools for collaborative learning. In P. A. Kirschner, S. J. Buckingham-Shum, and C. S. Carr, editors, *Visualizing argumentation: Software tools for collaborative and educational sense-making*, pages 51–73. Springer, 2003.
34. C. Karlof and D. Wagner. Secure routing in wireless sensor networks: attacks and countermeasures. *Ad Hoc Networks*, 1:293–315, 2003.
35. Y. Katz and J. Golbeck. Social network-based trust in prioritzed default logic. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.
36. T. Khopkar, X. Li, and P. Resnick. Self-selection, slipping, salvaging, slacking and stoning: The impacts of. In *Proceedings of the 6th ACM Conference on Electronic Commerce*, Vancouver, Canada, June 2005. ACM.
37. B. Khosravifar, J. Bentahar, A. Moazin, and P. Thiran. On the reputation of agent-based web services. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 1352–1357, Atlanta, July 2010.
38. P. A. Kirschner, S. J. Buckingham-Shum, and C. S. Carr, editors. *Using computer supported argument visualization to teach legal argumentation.* Springer, 2003.
39. E. Kok, J.-J. Meyer, H. Prakken, and G. Vreeswijk. Testing the benefits of structured argumentation in multi-agent deliberation dialogues. In *Proceedings of the 9th International Workshop on Argumentation in Multiagent Systems*, Valencia, Spain, 2012.
40. S. Kraus, K. Sycara, and A. Evenchik. Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence*, 104(1–2):1–69, 1998.
41. J. Lang, M. Spear, and S. F. Wu. Social manipulation of online recommender systems. In *Proceedings of the 2nd International Conference on*

Social Informatics, Laxenburg, Austria, 2010.
42. K. Lerman and A. Galstyan. Analysis of social voting patterns on Digg. In *Proceedings of the 1st Workshop on Online Social Networks*, Seattle, August 2008.
43. C.-J. Liau. Belief, information acquisition, and trust in multi-agent systems — a modal logic formulation. *Artificial Intelligence*, 149:31–60, 2003.
44. P.-A. Matt, M. Morge, and F. Toni. Combining statistics and arguments to compute trust. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagents Systems*, Toronto, Canada, May 2010.
45. H. Mercier and D. Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74, 2011.
46. S. Modgil and H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.
47. L. Mui, M. Moteashemi, and A. Halberstadt. A computational model of trust and reputation. In *Proceedings of the 35th Hawai'i International Conference on System Sciences*, 2002.
48. S. Naylor. *Not a Good Day Day to Die: The Untold Story of Operation Anaconda*. Berkley Caliber Books, New York, 2005.
49. N. Oren, T. Norman, and A. Preece. Subjective logic and arguing with evidence. *Artificial Intelligence*, 171(10–15):838–854, 2007.
50. S. Parsons, K. Atkinson, Z. Li, P. McBurney, E. Sklar, M. Singh, K. Haigh, K.Levitt, and J. Rowe. Argument schemes for reasoning about trust. *Argument and Computation*, 5(2–3):160–190, 2014.
51. S. Parsons and S. Green. Argumentation and qualitative decision making. In *Proceedings of the 5th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 1999.
52. S. Parsons, P. McBurney, and E. Sklar. Reasoning about trust using argumentation: A position paper. In *Proceedings of the Workshop on Argumentation in Multiagent Systems*, Toronto, Canada, May 2010.
53. S. Parsons, C. Sierra, and N. R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261—292, 1998.
54. S. Parsons, E. I. Sklar, J. Salvit, H. Wall, and Z. Li. ArgTrust: Decision making with information from sources of varying trustworthiness (Demonstration). In *Proceedings of Autonomous Agents and Multiagent Systems (AAMAS)*, St Paul, MN, USA, May 2013.
55. S. Parsons, Y. Tang, E. Sklar, P. McBurney, and K. Cai. Argumentation-based reasoning in agents with varying degrees of trust. In *Proceedings of the 10th International Conference on Autonomous Agents and Multi-Agent Systems*, Taipei, Taiwan, 2011.
56. S. Parsons, M. Wooldridge, and L. Amgoud. Properties and complexity of formal inter-agent dialogues. *Journal of Logic and Computation*, 13(3):347–376, 2003.
57. P. Pasquier, R. Hollands, I. Rahwan, F. Dignum, and L. Sonenberg. An empirical study of interest-based negotiation. *Journal of Autonomous Agents and Multi-Agent Systems*, 22(2):249–288, 2011.

58. H. Prakken. On dialogue systems with speech acts, arguments, and counterarguments. In *Proceedings of the Seventh European Workshop on Logic in Artificial Intelligence*, Berlin, Germany, 2000. Springer Verlag.
59. H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 2005.
60. I. Rahwan, M. I. Madakkatel, J. F. Bonnefon, R. N. Awan, and S. Abdallah. Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science*, 34(8):1483–1502, 2010.
61. I. Rahwan and G. R. Simari, editors. *Argumentation in Artificial Intelligence*. Springer Verlag, Berlin, Germany, 2009.
62. P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. In M. R. Baye, editor, *The Economics of the Internet and E-Commerce*, pages 127–157. Elsevier Science, Amsterdam, 2002.
63. P. Schank and M. Ranney. Improved reasoning with *Convince Me*. In *CHI'95 Conference Companion*, pages 276–277, 1995.
64. R. Stranders, M. de Weerdt, and C. Witteveen. Fuzzy argumentation for trust. In F. Sadri and K. Satoh, editors, *Proceedings of the Eighth Workshop on Computational Logic in Multi-Agent Systems*, volume 5056 of *Lecture Notes in Computer Science*, pages 214–230. Springer Verlag, 2008.
65. Y. Sun, W. Yu, Z. Han, and K. J. R. Liu. Trust modeling and evaluation in ad hoc networks. In *Proceedings of the YYth Annual IEEE Global Communications Conference*, pages 1862–1867, 2005.
66. D. Suthers, A. Weiner, J. Connelly, and M. Paolucci. Belvedere: Engaging students in critical discussion of science and public policy issues. In *Proceedings of the 7th World Conference on Artificial Intelligence in Education*, pages 266–273, Washington, DC, August 1995.
67. K. Sycara. Persuasive argumentation in negotiation. *Theory and Decision*, 28:203–242, 1990.
68. Y. Tang, K. Cai, P. McBurney, E. Sklar, and S. Parsons. Using argumentation to reason about trust and belief. *Journal of Logic and Computation*, 22(5):979–1018, 2012.
69. Y. Tang, K. Cai, E. Sklar, and S. Parsons. A prototype system for argumentation-based reasoning about trust. In *Proceedings of the 9th European Workshop on Multiagent Systems*, Maastricht, Netherlands, November 2011.
70. Y. Tang, E. I. Sklar, and S. Parsons. An Argumentation Engine: ArgTrust. In *Proceedings of the Workshop on Argumentation in Multiagent Systems (ArgMAS) at Autonomous Agents and MultiAgent Systems (AAMAS)*, Valencia, Spain, June 2012.
71. P. Tolchinsky, S. Modgil, U. Cortes, and M.Sanchez-Marre. Cbr and argument schemes for collaborative decision making. In *Proceedings of the First International Conference on Computational Models of Argument*, pages 71–82, Liverpool, UK, 2006.

72. S. W. van den Braak, H. van Oostendorp, H. Prakken, and G. A. Vreeswijk. A critical review of argument visualization tools: do users become better reasoners? In *Proceedings of the Workshop on Computational Models of Natural Argument*, pages 67–75, 2006.
73. S. Villata, G. Boella, D. M. Gabbay, and L. van der Torre. Arguing about the trustworthiness of the information sources. In *Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Belfast, UK, 2011.
74. C. M. Vogel. *Inheritance reasoning: Psychological plausibility, proof theory and semantics*. PhD thesis, University of Edinburgh, Centre for Cognitive Science, 1995.
75. G. Vreeswijk and H. Prakken. Credulous and sceptical argument games for preferred semantics. In *Proceedings of the 7th European Workshop on Logics in Artificial Intelligence*, 2000.
76. D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, Albany, NY, USA, 1995.
77. R. Walton, C. Gierl, H. Mistry, M. P. Vessey, and J. Fox. Evaluation of computer support for prescribing (CAPSULE) using simulated cases. *British Medical Journal*, 315:791–795, 1997.
78. Y. Wang and M. P. Singh. Trust representation and aggregation in a distributed agent system. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, MA, 2006.
79. B. Yu and M. Singh. Distributed reputation management for electronic commerce. *Computational Intelligence*, 18(4):535–349, 2002.