Formal Arguments, Preferences, and Natural Language Interfaces to Humans: an Empirical Evaluation

Federico Cerutti and **Nava Tintarev** and **Nir Oren**[1]

Abstract.
It has been claimed that computational models of argumentation provide support for complex decision making activities in part due to the close alignment between their semantics and human intuition. In this paper we assess this claim by means of an experiment: people's evaluation of formal arguments — presented in plain English — is compared to the conclusions obtained from argumentation semantics. Our results show a correspondence between the acceptability of arguments by human subjects and the justification status prescribed by the formal theory in the majority of the cases. However, post-hoc analyses show that there are some significant deviations, which appear to arise from implicit knowledge regarding the domains in which evaluation took place. We argue that in order to create argumentation systems, designers must take implicit domain specific knowledge into account.

1 INTRODUCTION

Like other systems for automatic reasoning, argumentation approaches can suffer from "opacity": humans can find difficult to understand why some course of actions are chosen, and what alternatives exist [17]. Other domains with similar issues (e.g. expert and recommender systems [15, 22, 24, 28]) have attempted to overcome this issue through the use of natural language interfaces, and it has been suggested that such an approach can also aid argumentation systems. However, with the exception of [20, 3], no user experiments have yet been carried out to determine whether humans agree with the reasoning of instantiated argument systems.

This paper describes an experiment, first outlined in [4], with human participants which studies the correspondences between formal arguments — using Prakken and Sartor's *Formal System II* (§ 5 of [18]), hereafter abbreviated *P&S* — and a natural language representation of them (explanatory interface). The experiment evaluates whether people ascribe the same status to the natural language statements as the one suggested by formal argumentation semantics.

P&S satisfies two important desiderata for a formal argument system, namely that (1) it is based on a model of human reasoning (so as to ease the transition between the formal system and natural language); and (2) it provides us with the ability to create arguments about preferences. Dealing with preferences is an important aspect of many argumentation systems [14, 10, 2] and seems to be a core source for defeasibility. Unlike other systems, which have to be extended with meta-arguments in order to derive a preference argument between two arguments or rules, *P&S* encompasses defeasible reasoning about preferences.

In the experiment, participants read a paragraph in natural language — handcrafted in order to be natural and fluent — depicting an indirect dialogue between fictitious actors: each actor plays a role by defending a specific position. Since *P&S* was developed to support legal reasoning, we hypothesize that there is a correspondence between statement acceptability (as judged by humans) and justification status (according to the formal model of *P&S*). In particular, we expect that the majority of the participants agree with the skeptically accepted arguments, but not with the credulously accepted ones.

The paper is structured as follows. Section 2 summarises the *P&S* approach. Section 3 describes the experimental methodology and the research hypotheses. Section 4 analyses our experimental results. Section 5 compares this work with the relevant literature, while Section 6 concludes the paper. [5] presents the text scenarios and formal arguments used in the experiment, as well as screenshots.

2 THE PRAKKEN AND SARTOR APPROACH

P&S [18] considers an object language similar to that used in logic programming. Within the language, an atom $p(t)$ and its negation $\neg p(t)$ are literals. The connective $\sim$ represents *negation as failure*. In addition, the language contains a distinguished binary predicate symbol $\prec$ with which information about priorities can be expressed in the object language itself.

Definition 1. *A* rule *is an expression of the form:*

$$r : L_0 \wedge \ldots \wedge L_j \wedge \sim L_k \wedge \ldots \wedge \sim L_m \Rightarrow L_n$$

*where r, a first-order term, is the rule name, L_i ($0 \leq i \leq n$) are strong literals, $\mathtt{B}(r) = \{L_0, \ldots L_j, \sim L_k, \ldots, \sim L_m\}$ is the set (*body*) of* antecedents*, and $\mathtt{H}(r) = L_n$ is the* consequent *or* head *of the rule.*

A strong *literal is an atomic first-order formula, or a formula of the form $r \prec r'$, or such formulae preceded by strong negation $\neg$. A* weak *literal is a literal of the form $\sim L$, where L is a strong literal.*

A strict *rule contains no weak literals. Priorities are allowed only between defeasible rules: i.e. $r \prec r'$ is in the language iff both r and r' are defeasible.*

The complement *of a literal L is denoted with $\overline{L}$. For any atom A, $\overline{A} = \neg A$ and $\overline{\neg A} = A$.*

An *ordered theory* is a tuple of sets of strict and defeasible rules with some axioms (formally, strict rules) ensuring a strict partial order of preferences (transitivity, contraposition of transitivity, and asymmetry).

[1] University of Aberdeen, School of Natural and Computing Science, Kings College, AB24 3UE, Aberdeen, UK, email: {f.cerutti,n.tintarev,n.oren}@abdn.ac.uk

Definition 2. *An* ordered theory *is a* $\langle S, D\rangle$*, where* S *and* D *are sets of, respectively, strict and defeasible rules. The set* S *always contains*[2]*:*
$(x \prec y) \wedge (y \prec z) \Rightarrow (x \prec z)$, $(x \prec y) \wedge \neg(x \prec z) \Rightarrow \neg(y \prec z)$
$(y \prec z) \wedge \neg(x \prec z) \Rightarrow \neg(x \prec y)$, $(x \prec y) \Rightarrow \neg(y \prec x)$

The set of rules of an ordered theory can be combined to form *arguments*.

Definition 3. *An* argument *is a finite sequence* $\mathbf{a} = \langle r_0, \ldots, r_n\rangle$ *of ground instances of rules such that:*
1. $\forall\, r_i \in \mathbf{a}, \forall L_j \in \mathtt{B}(r_i)$ *s.t.* L_j *is a strong literal, there is a* $r_k, k < i$ *such that* $L_j = \mathtt{H}(r_k)$*;*
2. $\forall r_i, r_j \in \mathbf{a}, r_i \neq r_j$, $\mathtt{H}(r_i) \neq \mathtt{H}(r_j)$.

Given an argument $\mathbf{a} = \langle r_1, \ldots, r_n\rangle$*, and a finite sequence of rules* $T = \langle r_{n+1}, \ldots, r_m\rangle$*:*
- $\mathtt{C}(\mathbf{a}) = \{L \mid L = \mathtt{H}(r), \forall r \in \mathbf{a}\}$ *is the* set of conclusions *of* $\mathbf{a}$*;*
- $\mathtt{A}(\mathbf{a}) = \{L \mid \sim \overline{L} \in \mathtt{B}(r), \forall r \in \mathbf{a}\}$ *is the* set of assumptions *of* $\mathbf{a}$*;*
- $\mathbf{a} + T = \langle r_1, \ldots, r_m\rangle = \mathbf{a}'$*, and* $\mathbf{a}'$ *is an argument.*

An argument $\mathbf{a}$ *is* based on *the ordered theory* $\Gamma = \langle S, D\rangle$ *iff* $\forall r \in \mathbf{a}, r \in S \cup D$. $\mathcal{A}rgs_\Gamma$ *is the* set of arguments *based on* Γ*. For any set* $\mathcal{A}rgs$ *of arguments,* $<_{\mathcal{A}rgs} = \{(r_i < r_j) \mid \exists \mathbf{a} \in \mathcal{A}rgs$ *s.t.* $(r_i \prec r_j) \in \mathtt{C}(\mathbf{a})\}$.

Let us now recall the notions of conflict (*attack*) and defeat as prescribed by *P&S*.

Definition 4. *Given a set of arguments* $\mathcal{A}rgs$*, and* $\mathbf{a}_1, \mathbf{a}_2 \in \mathcal{A}rgs$*.* $\mathbf{a}_1$ attacks $\mathbf{a}_2$ *iff there are sequences* S_1*,* S_2 *of strict rules such that* $L = \mathtt{C}(\mathbf{a}_1')$ *where* $\mathbf{a}_1' = \mathbf{a}_1 + S_1$ *and either* $\overline{L} = \mathtt{C}(\mathbf{a}_2')$ *where* $\mathbf{a}_2' = \mathbf{a}_2 + S_2$*, or* $\overline{L} \in \mathtt{A}(\mathbf{a}_2)$.

An argument is coherent *iff it does not attack itself.*

$\mathcal{A}rgs$ *is* conflict-free *iff* $\nexists \mathbf{a}_1, \mathbf{a}_2 \in \mathcal{A}rgs$ *s.t.* $\mathbf{a}_1$ *attacks* $\mathbf{a}_2$.

P&S further divides the notion of conflict or attack into rebutting and undercutting attacks.

Definition 5. *Given* $\mathcal{A}rgs$ *a set of arguments,* $\mathbf{a}_1, \mathbf{a}_2 \in \mathcal{A}rgs$*,* S_1 *and* S_2 *two sequences of strict rules, and* $L = \mathtt{C}(\mathbf{a}_1')$ *where* $\mathbf{a}_1' = \mathbf{a}_1 + S_1$*. Then*
1. $\mathbf{a}_1$ undercuts $\mathbf{a}_2$ *iff* $\overline{L} \in \mathtt{A}(\mathbf{a}_2)$*;*
2. $\mathbf{a}_1$ rebuts $\mathbf{a}_2$ *iff* $\overline{L} = \mathtt{C}(\mathbf{a}_2')$ *where* $\mathbf{a}_2' = \mathbf{a}_2 + S_2$*, provided that* $R_L(\mathbf{a}_1') \not<_{\mathcal{A}rgs} R_{\overline{L}}(\mathbf{a}_2')$.

Given $\mathbf{a} \in \mathcal{A}rgs$ *and* S *a sequence of strict rules, the* set of defeasible rules relevant to L *is*

$$R_L(\mathbf{a}+S) = \begin{cases} \{r_d\} \;\; \textit{iff} \;\; r_d \in \mathbf{a},\ r_d \textit{ is defeasible, and} \\ \qquad\qquad L = \mathtt{H}(r_d) \\ R_{L_1}(\mathbf{a}+S) \cup \ldots \cup R_{L_n}(\mathbf{a}+S) \\ \qquad \textit{iff} \;\; \mathbf{a} \textit{ is defeasible, and} \\ \qquad\qquad r_s : L_1 \wedge \ldots \wedge L_n \Rightarrow L, r_s \in S \end{cases}$$

Given R_1 *and* R_2 *two sets of defeasible rules,* $R_1 <_{\mathcal{A}rgs} R_2$ *iff* $\exists r_1 \in R_1$ *such that* $\forall r_2 \in R_2, (r_1 <_{\mathcal{A}rgs} r_2)$.

Given the above definition of rebut and undercut, which explicitly consider preferences among rules, the concept of *defeat* between arguments follows easily.

Definition 6. *Given* $\mathcal{A}rgs$ *a set of arguments, and* $\mathbf{a}_1, \mathbf{a}_2 \in \mathcal{A}rgs$*.* $\mathbf{a}_1$ $\mathcal{A}rgs$-defeats $\mathbf{a}_2$ *iff:*
1. $\mathbf{a}_1 = \langle\rangle$ *and* $\mathbf{a}_2$ *is incoherent; or*
2. $\mathbf{a}_1$ *undercuts* $\mathbf{a}_2$*; or*
3. $\mathbf{a}_1$ *rebuts* $\mathbf{a}_2$ *and* $\mathbf{a}_2$ *does not undercut* $\mathbf{a}_1$.

Moreover, $\mathbf{a}_1$ strictly $\mathcal{A}rgs$-defeats $\mathbf{a}_2$ *iff* $\mathbf{a}_1$ $\mathcal{A}rgs$*-defeats* $\mathbf{a}_2$ *and* $\mathbf{a}_2$ *does not* $\mathcal{A}rgs$*-defeat* $\mathbf{a}_1$.

P&S defines an argument as acceptable with respect to a set of arguments if they defend it against the defeats it receives.

Definition 7. *An argument* $\mathbf{a}_1$ *is* acceptable *with respect to a set* $\mathcal{A}rgs$ *of arguments iff* $\forall \mathbf{a}_2$ *s.t.* $\mathbf{a}_2$ $\mathcal{A}rgs$*-defeats* $\mathbf{a}_1$*,* $\exists \mathbf{a}_3 \in \mathcal{A}rgs$ *s.t.* $\mathbf{a}_3$ *strictly* $\mathcal{A}rgs$*-defeats* $\mathbf{a}_2$.

Similarly to Dung's abstract argumentation framework (AF) [7][3], *P&S* considers two kinds of semantics: *skeptical* and *credulous*. A skeptical semantics generally selects a smaller but "stronger" set of arguments, and in *P&S* is defined on the basis of the characteristic function of an ordered theory.

Definition 8. *Let* $\Gamma = \langle S, D\rangle$ *be an ordered theory,* $S \subseteq \mathcal{A}rgs_\Gamma$ *and* $CF_\Gamma = \{C \subseteq \mathcal{A}rgs_\Gamma \mid C$ *is conflict-free*$\}$*. Then the* characteristic function *of* Γ *is:*

$$\begin{array}{rcl} G_\Gamma & : & CF_\Gamma \mapsto 2^{\mathcal{A}rgs_\Gamma} \\ G_\Gamma(S) & = & \{\mathbf{a} \in \mathcal{A}rgs_\Gamma \mid \mathbf{a} \textit{ is acceptable with respect to } S\} \end{array}$$

$Just(\mathcal{A}rgs_\Gamma)$ *denotes the set of* justified *arguments, namely the least fixpoint of* G_Γ.

Sometimes it is of interest to determine which sets of arguments are based on the same coherent point of view. Such a notion is captured by the stable semantics, which has a credulous flavour.

Definition 9. *A conflict-free set of arguments* $\mathcal{A}rgs$ *is a* stable extension *iff* $\forall \mathbf{a}_1 \notin \mathcal{A}rgs$*,* $\exists \mathbf{a}_2 \in \mathcal{A}rgs$ *s.t.* $\mathbf{a}_2$ $\mathcal{A}rgs$*-defeats* $\mathbf{a}_1$.

Given the above definitions, the notion of justification status of arguments is as follows:

Definition 10. *For any ordered theory* $\Gamma = \langle S, D\rangle$ *and* $\mathbf{a} \in \mathcal{A}rgs_\Gamma$*:*
1. $\mathbf{a}$ *is* justified *iff* $\mathbf{a} \in Just(\mathcal{A}rgs_\Gamma)$.
2. $\mathbf{a}$ *is* overruled *iff* $\mathbf{a}$ *is attacked by a justified argument.*
3. $\mathbf{a}$ *is* defensible and stable *iff* $\mathbf{a}$ *is in a stable extension.*

3 THE EXPERIMENT[4]

The experiment consists of presenting each participant with a text, written in natural language, followed by a questionnaire. The questionnaire serves to assess the justification status of the natural language statements from the participants' point of view.

The experiment follows a *between subjects design* across eight texts, i.e. each participant is shown a single (randomly selected) text. Each text is derived from a knowledge base (KB) formalised using *P&S*. We considered the following four domains:
1. weather forecast (derived from an example discussed in [13]);
2. political debate;
3. used car sale;
4. romantic relationship.

For every domain i, we generated two related KBs: a *base case* ($i.B$), which we then modified to create its *extended case* ($i.E$).

[2] Round brackets are not element of the language, they are used informally to improve readability.

[3] Although it is beyond the scope of this paper, intuitively an AF corresponding to a *P&S* theory can be derived by considering the same set of arguments, and the relation of $\mathcal{A}rgs$-defeats as the attack relation.

[4] Additional technical material can be found in [5].

Each base case considers two arguments, $\mathbf{a}_1$ and $\mathbf{a}_2$. These arguments support two contradicting conclusions. Moreover, a preference, $\mathbf{a}_3$, is stated in favour of $\mathbf{a}_2$: this determines the successful defeat of $\mathbf{a}_1$.

To exemplify, let us consider a political debate (domain 2). A politician and an economist discuss the potential financial outcome of the independence of a region X. The politician puts forward an argument in favour of the conclusion "If Region X becomes independent, X's citizens will be poorer than they are now". Another argument holding a contradicting conclusion (i.e. Region X will not be poorer) is advanced by the economist. The economist's opinion is likely to be preferred to that of the politician, and is supported by a technical document. Formally, this situation (cf. case $2.B$, § 1.3 of [5]) is represented by a theory $\Gamma = \langle S, D \rangle$ with the following rules:

S	D
$s_1: \Rightarrow s_{AAA}$	$r_1: s_{AAA} \wedge \sim ex_{AAA} \Rightarrow poorer$
$s_2: \Rightarrow s_{BBB}$	$r_2: s_{BBB} \wedge s_{doc} \wedge \sim ex_{BBB} \wedge \sim ex_{doc} \Rightarrow \neg\, poorer$
$s_3: \Rightarrow s_{doc}$	$r_3: \sim ex_{expert} \Rightarrow r_1 \prec r_2$

Γ gives rise to the following set of arguments:

$$\mathcal{A}rgs = \{\mathbf{a}_1 = \langle s_1, r_1 \rangle, \mathbf{a}_2 = \langle s_2, s_3, r_2 \rangle, \mathbf{a}_3 = \langle r_3 \rangle\}$$

where $\mathbf{a}_2$ $\mathcal{A}rgs$-defeats $\mathbf{a}_1$. For each base case, the set of justified arguments is $\{\mathbf{a}_2, \mathbf{a}_3\}$, and $\mathbf{a}_1$ is always overruled.

Each *extended case* adds another argument, $\mathbf{a}_4$, whose effect is to reinstate $\mathbf{a}_1$. We considered three ways to perform this reinstatement (See Table 1): undercut of the preference argument ($1.E$, § 1.2 of [5], and $3.E$, § 1.6 of [5]); rebuttal of the argument $\mathbf{a}_2$ ($2.E$, § 1.4 of [5]); rebuttal of the preference argument ($4.E$, § 1.8 of [5]).

For instance, in our running example ($2.E$, § 1.4 of [5]), argument $\mathbf{a}_4$ states that more recent research disputes the claim of the economist. Thus Γ is enlarged with: $s_4 :\Rightarrow s_{research}$ and $s_5 : s_{research} \Rightarrow \neg s_{doc}$. $\mathbf{a}_4 = \langle s_4, s_5 \rangle$ is then derived. This gives rise to two stable extensions, $\{\mathbf{a}_1, \mathbf{a}_3, \mathbf{a}_4\}$ and $\{\mathbf{a}_2, \mathbf{a}_3\}$.

Table 1 summarises the various scenarios considered. Let us note that the knowledge bases for scenarios $1.B$, $3.B$, and $4.B$ are identical except for the names of atoms. The same holds for $1.E$ and $3.E$.

We created natural language text from each knowledge base which was shown to the participants. This text was hand-crafted, though we intend to investigate text generation techniques such as Controlled English [19] in future work.

Table 1: Domains and scenarios for the experiment. $1.B$, $3.B$, and $4.B$ are logically equivalent: AAA and BBB arguments are perfectly symmetrical in contrast to $2.B$, where BBB's position is supported by a technical document.

Domain	Base Case	Extended Case	Type of reinstatement
1, weather	$1.B$ § 1.1 of [5]	$1.E$ § 1.2 of [5]	preference undercut
2, politics	$2.B$ § 1.3 of [5]	$2.E$ § 1.4 of [5]	$\mathbf{a}_2$ rebuttal
3, buying car	$3.B$ § 1.5 of [5]	$3.E$ § 1.6 of [5]	preference undercut
4, romance	$4.B$ § 1.7 of [5]	$4.E$ § 1.8 of [5]	preference rebuttal

3.1 Method

The experiment was administered as an online questionnaire using Amazon's Mechanical Turk (MT) service [16]. Screenshots of the experiment can be found in § 2 of [5]. Before the experiment, the English fluency of each participant was tested using a very short Cloze Test [23]. Participants who failed the test were excluded from the experiment.

In a *between subjects* design, participants are given one of the eight scenarios, where the first two contradicting arguments ($\mathbf{a}_1$ and $\mathbf{a}_2$) are presented by two fictitious actors, AAA and BBB. This example is from scenario $2.E$ (domain 2, extended case § 1.4 of [5]):

> In a TV debate, the politician AAA argues that if Region X becomes independent then X's citizens will be poorer than now. Subsequently, financial expert Dr. BBB presents a document; which scientifically shows that Region X will not be worse off financially if it becomes independent. After that, the moderator of the debate reminds BBB of more recent research by several important economists that disputes the claims in that document.

Then, the participants are asked to determine which of the following positions they think is accurate:

- $\mathcal{P}_\mathbf{A}$: I think that AAA's position is correct (e.g. *"X's citizens will be poorer than now"*);
- $\mathcal{P}_\mathbf{B}$: I think that BBB's position is correct (e.g. *"X's citizens will not be worse off financially"*);
- $\mathcal{P}_\mathbf{U}$: I cannot determine if either AAA's or BBB's position is correct (e.g. *"I cannot conclude anything about Region X's finances"*).

Next, participants are asked to rate a number of statements in terms of relevance (for the conclusion) and agreement: this provided the necessary data for investigating the support for the preference statement. For agreement, participants are asked *"How much do you agree with the following statements?"* and respond on a 7 point scale from Disagree to Agree for each statement. The question about relevance is phrased in terms of the final conclusion, e.g. *"How relevant are the following statements for your conclusion?"* (from 'can be ignored' to 'has a large impact').

3.2 Hypotheses

This experiment aims to verify the link between formal argumentation semantics and human reasoning. Our hypotheses therefore revolve around the assumption that participants are able to use the natural language text to reach a conclusion in agreement with the one obtained by *P&S*:

H1: In the base cases (Scenarios $1.B$, $2.B$, $3.B$ and $4.B$), the majority of participants will agree with BBB's statement (*position* $\mathcal{P}_\mathbf{B}$).

H2: In the extended cases (Scenarios $1.E$, $2.E$, $3.E$ and $4.E$), the majority of participants will agree that they cannot conclude anything from the text (*position* $\mathcal{P}_\mathbf{U}$).

H3: The majority of participants who view a base case scenario will agree with the preference argument, and find it relevant.

4 RESULTS

Of the 366 people who began the experiment, 199 failed the English test and did not proceed further. An additional 6 participants did not complete the whole experiment. The remaining 161 were roughly equally split across the eight scenarios.

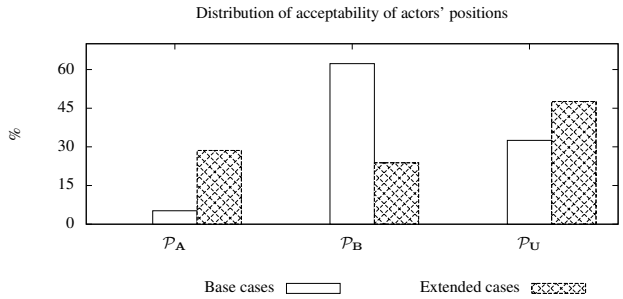

Figure 1: Distribution of the final conclusion $\mathcal{P}_\mathbf{A}/\mathcal{P}_\mathbf{B}/\mathcal{P}_\mathbf{U}$, comparing base cases with extended cases, in percent.

Non-parametric statistical tests are used: Chi-square within a sample, Mann-Whitney for pair-wise comparisons, Kruskal-Wallis in comparisons across more than two groups, and Fisher for associations between two sets of groups.

4.1 Hypothesis Verification

Figure 1 depicts the participants' choices among the positions $\mathcal{P}_\mathbf{A}$, $\mathcal{P}_\mathbf{B}$, and $\mathcal{P}_\mathbf{U}$, distinguishing between base and extended cases. The majority of participants in the base cases selected $\mathcal{P}_\mathbf{B}$, and the majority of participants in the extended cases selected $\mathcal{P}_\mathbf{U}$. The response categories differ significantly in both distributions[5]. The majority categories of the distributions are also in line with **H1** and **H2**, which can thus be considered confirmed.

Let us now turn our attention to hypothesis **H3**. In the experiment, we asked the participants to rate how much (on a scale of 1 to 7) they agree with the following statement (*agreement*), and whether it is relevant in drawing their conclusion (*relevance*): *"BBB is more trustworthy than AAA."* Mann-Whitney tests indicate that there is a significant difference between the base and the extended cases for agreement[6] and relevance[7]. In addition, the median values both for agreement and relevance are greater for the base cases than for the extended cases. Hypothesis **H3** is thus also confirmed. Table 2 shows some variation across scenarios, such as the reversal of the conclusion in domain 4, extended case. We therefore include an analysis by scenario in the next section.

4.2 Post-Hoc Analysis

Table 2 summarises the distribution of the final conclusion across scenarios. As we expected, the Fisher test (conflating over base and extended cases) shows that there is a significant ($p < 0.01$) dependency[8] on the scenario of the positions taken by participant. Table 2, however, highlights two anomalies: (1) there is a substantially greater number of $\mathcal{P}_\mathbf{U}$ for scenario $1.B$ than $3.B$ and the others scenarios; (2) the majority of participants chose $\mathcal{P}_\mathbf{A}$ instead of $\mathcal{P}_\mathbf{U}$ in scenario $4.E$.

[5] Base cases, χ^2 analysis (2, N=77)=37.74, $p < 0.001$; extended cases χ^2 (2, N=84)=8.0, $p < 0.02$.
[6] Mann-Whitney $U(1778)$, $Z = -5.0$, $p < 0.001$.
[7] Mann-Whitney $U(1852)$, $Z = -4.7$, $p < 0.001$.
[8] Fisher $(N = 161) = 48.756$, $p < 0.001$, 10000 sampled tables, Monte Carlo approach with 99% confidence interval (MC99).

Table 2: Distribution of the final conclusion $\mathcal{P}_\mathbf{A}/\mathcal{P}_\mathbf{B}/\mathcal{P}_\mathbf{U}$ in percent, for each scenarios. Shading denotes the most likely conclusions.

	Base Cases			Extended Cases		
	$\mathcal{P}_\mathbf{A}$	$\mathcal{P}_\mathbf{B}$	$\mathcal{P}_\mathbf{U}$	$\mathcal{P}_\mathbf{A}$	$\mathcal{P}_\mathbf{B}$	$\mathcal{P}_\mathbf{U}$
1, weather	5.0	50.0	45.0	15.8	21.1	63.2
2, politics	5.3	63.2	31.6	21.1	10.5	68.4
3, buying car	0.0	68.2	31.8	23.8	23.8	52.4
4, romance	12.5	68.8	18.8	48.0	36.0	16.0

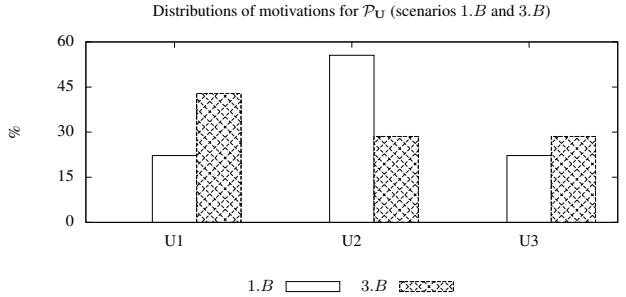

Figure 2: Distribution across three categories of justification (U1: lack of information, U2: domain specific reasons; U3: other) for agreement with the $\mathcal{P}_\mathbf{U}$ position in scenarios $1.B$ and $3.B$.

4.2.1 *Distribution of Positions in Base Cases*

From Table 2 it appears that the scenarios do not affect the distribution of choices in base cases[9] — the majority of participants chose $\mathcal{P}_\mathbf{B}$ in any scenario. A Fisher test found no significant effect of association for scenario[10]. However, Table 2 shows a greater number of $\mathcal{P}_\mathbf{U}$ positions for scenario $1.B$ (weather, base case) compared to the other scenarios.

We classified the motivations added by participants who chose $\mathcal{P}_\mathbf{U}$ into three categories: "lack of information" (U1), "domain specific reasons" (U2), "other" (U3). Figure 2 summarises the distributions of motivations according to these three categories for participants who chose the $\mathcal{P}_\mathbf{U}$ position in scenarios $1.B$ and $3.B$, which are logically equivalent (cf. Table 1). Figure 2 shows that the majority of motivations supporting the choice of the $\mathcal{P}_\mathbf{U}$ position in the case of the weather forecast ($1.B$) are domain-dependent (e.g. one participant wrote *"All weather forecasts are notoriously inaccurate."*). On the other hand, in scenario $3.B$, the $\mathcal{P}_\mathbf{U}$ position is mainly justified by a lack of information, e.g. *"I have two conflicting reports."*.

4.2.2 *Distribution of Positions in Base/Extended Cases*

To further investigate the anomaly described in Section 4.2.1, we analysed the relationships between base and extended cases. In particular, we expect that the distributions of choices (among $\mathcal{P}_\mathbf{A}$, $\mathcal{P}_\mathbf{B}$, and $\mathcal{P}_\mathbf{U}$) in the base case is significantly different from the distribution of choices in the corresponding extended case. This is the case for the third domain ($3.B$ and $3.E$, buying a car)[11], but not for the first one ($1.B$ and $2.B$, weather forecasts)[12].

[9] This despite the fact that scenario $2.B$ is formulated in a slightly different way with respect to scenarios $1.B$, $3.B$, and $4.B$, cf. Table 1.
[10] Fisher $(N = 77) = 5.268$, $p = 0.488$, 10000 sampled tables, MC99.
[11] Fisher $(N = 43) = 10.693$, $p < 0.001$, 10000 sampled tables, MC99.
[12] Fisher $(N = 39) = 3.832$, $p = 0.187$, 10000 sampled tables, MC99.

Let us notice here that the formal representation of $1.B$ is equivalent to $3.B$ and that $1.E$ is equivalent to $3.E$. This seems to suggest that participants used "collateral knowledge" [12] (i.e. domain dependent knowledge) when they performed their task.

4.2.3 *Distribution of Positions in Extended Cases*

In the case of extended cases, as Table 2 suggests, the domain has a significant effect[13] on the distribution of positions. The main effect appears to be in domain 4 (romance), where a "reversal of preference" occurs. This result might suggest that the nature of the domain — subjective, high investment — is the cause of this. A second explanation is the different logical form of the scenario (cf. Table 1).

4.2.4 *Relevance and Agreement*

Considering the analysis underlying the acceptance of hypotheses **H3** about *agreement* and *relevance*, according to the Kruskal-Wallis test, the distributions of the answers for each scenario are statistically independent. This suggests the post-hoc analysis summarised in Table 3a (following the approach described by [21], cited in [9]). This table highlights how, in the case of political debate (domain 3), there is a statistically significant ($p < 0.05$) difference between the distributions of the answers in the base and in the extended cases. In particular, the wide difference of median values for base and extended cases suggests that the agreement and relevance for preference statements is much higher for the base case than in the extended case, giving even stronger support for hypothesis **H3**.

Moreover, looking at Table 3b, the median values both for relevance and agreement in scenarios $3.B$ and $4.B$ are substantially different — resp. 6.50 and 2.00 — while their formal representations are equivalent (cf. Table 1). Table 3b shows that this difference is significant ($p < 0.05$), and supports our claim that domains and their "collateral knowledge" (buying a car vs. romance in this case) have a significant impact on people's choices.

4.3 Discussion

Our post-hoc analysis suggests that people evaluate preference relations in a domain-dependant way. In particular, two interesting situations arose during the experiment. The first one regards a direct comparison between scenario $1.B$ (about weather forecasts) and $3.B$ (buying a car), which are the base cases with the highest percentage of people choosing $\mathcal{P}_{\mathrm{U}}$ (Table 2). In the case of the weather forecast, the majority of people who chose $\mathcal{P}_{\mathrm{U}}$ justified that choice with domain-specific rationalisations (e.g. *"All weather forecasts are notoriously inaccurate"*). However, in the buying-a-car domain, the majority of justification are related to lack of information (Figure 2), which seems to be rational if the preference relation is not strong enough to convince. One possible interpretation is that the participants used the preference as a support for conclusions and that the acceptance of a preference is strictly related to the domain.

The second situation is highlighted by Table 3b: participants' answers regarding their agreement with the preference argument and its relevance for the final decision are statistically different when we consider base cases for the domains of political debate and romance (Scenarios $3.B$ and $4.B$ resp.). However, the formal representations of these scenarios are logically equivalent and also the tests on the distribution of choices among the three given alternatives showed no dependencies. This lends further tentative support to the hypothesis that people may use preference differently in different domains.

[13] Fisher $(N = 84) = 16.308$, $p < 0.05$, 10000 sampled tables, MC99.

Table 3: Post-hoc analysis regarding *relevance* and *agreement*: pairwise comparison base-extended cases (a); and between $1.B$ and $4.B$ (b). Statistically significant cases (i.e. when $|\overline{R_x} - \overline{R_y}| >$ C.D) are highlighted in grey.
† Mean rank as computed with the Kruskal-Wallis test
* Median
‡ *Critical Difference*, as computed in [21] cited by [9] with $\alpha = 0.05$.

		Base cases		Extended cases		
		$\overline{R_B}^{\dagger}$	Md^{*}_{B}	$\overline{R_E}^{\dagger}$	Md^{*}_{E}	C.D.‡
Relevance	1, weather	110.38	6.00	82.92	4.00	46.60
	2, politics	107.45	6.00	69.45	4.00	47.19
	3, buying car	118.05	6.50	67.45	4.00	44.38
	4, romance	48.34	2.00	44.40	2.00	46.57
Agreement	1, weather	116.38	6.00	87.18	4.00	46.60
	2, politics	103.34	6.00	65.05	4.00	47.19
	3, buying car	121.93	6.50	64.33	4.00	44.38
	4, romance	44.94	2.00	44.20	2.00	46.57

(a)

	Scenario $3.B$		Scenario $4.B$		
	$\overline{R_{3.B}}^{\dagger}$	$Md^{*}_{3.B}$	$\overline{R_{4.B}}^{\dagger}$	$Md^{*}_{4.B}$	C.D.‡
Relevance	118.05	6.50	48.34	2.00	47.79
Agreement	121.93	6.50	44.94	2.00	47.79

(b)

5 RELATED WORK

Similarly to this work, [20] investigates how reinstatement may affect acceptance of claims. Participants looked at an original claim which is attacked by an additional argument. [20] provides evidence that participants believe that the original claim, once reinstated, is acceptable. Unlike the current work, [20] does not rely on an instantiated theory, does not consider preferences, and focuses on two types of reinstatement only. Another difference is that it considers a restricted set of participants who have been interviewed in person: participants were requested to assess the "degree of acceptability" of arguments.

One strand of research looks at the use of argumentation in online debates and engaging citizens in policy: participants can vote on arguments and attacks [1, 8]. The work aims to develop a formal semantics that would allow for aggregation of votes, both for and against claims and arguments. While a promising approach, it is still very novel and has yet to become established.

Other work examines the creation of argumentation frameworks from natural language text [3]. Here, the textual entailment approach is used for mapping linguistic objects by means of semantic inferences at a textual level. From a formal counterpart, Dung's argumentation framework has been used in order to automatically evaluate the acceptability of arguments. Like [20], this work does not consider either structured arguments, or the impacts of preferences.

Finally, others suggest the use of aspects of narrative coherence to help specify conditions of well-formedness to arguments and identify arguments from unstructured text [27]. There is also a wealth

of research studying how to represent arguments in natural language [6], particularly in the legal domain [18]. Most of these approaches do not have an empirical grounding with human participants.

6 CONCLUSION AND FUTURE WORK

This paper presents an investigation into the relationship between formal systems of defeasible argumentation and arguments in natural language. We conducted an experiment aimed at evaluating argumentation models in relation to human cognition. In this experiment, participants read a text written in natural English depicting an indirect dialogue among some fictitious actors. Several domains were considered (weather forecast, political debate, used car sale, romantic relationship), with formal similarities among them (cf. Table 1).

The results suggest a correspondence between the formal theory and its representation in natural language. Moreover, when people apply preference rules, they generally follow what the *P&S* theory would prescribe. If a reinstatement to the less preferred argument is added, then the majority of the participants agreed that the situation is undecided, showing a "skeptical attitude" (cf. Fig. 1, Table 3a).

Since there is a suggestion that humans evaluate preference differently depending on domain, a deeper understanding of the relationship between the "collateral knowledge" [12] associated to the domain and the outcome is one of the intended directions of future work. Moreover, we intend to investigate whether the reversal of conclusion in the extended case for the 'romance' domain is due to the logical form containing a preference rebuttal or due to the subjective and high risk nature of the domain. We also plan to consider state-of-the-art argumentation formalisms such as [14], and more complex argument sets than those studied in this paper, possibly derived from argument corpora that can be formalised using either argument schemes [26] or formal systems like Carneades [11]. To achieve these goals, we intend to continue our investigation in the context of intelligence analysis [19, 25]. Finally, we also intend to compare the properties of natural language explanatory interfaces (e.g. story-telling approach vs direct arguments, linguistic indicators) and other types of interfaces to arguments (e.g. visual arguments [12]).

Acknowledgments

We thank the anonymous reviewers for their helpful comments.

This research has been carried out within the project "Scrutable Autonomous Systems" (SAsSY), funded by the UK Engineering and Physical Sciences Research Council, grant ref. EP/J012084/1.

Research was sponsored by US Army Research Laboratory (ARL) and the UK Ministry of Defence (MoD) under Agreement Number W911NF-06-3-0001. 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 US ARL, the U.S. Government, the UK MoD, or the UK Government. The US and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza, 'An Argumentation-Based Approach for Automatic Evaluation of Design Debates', in *Workshop on Computational Logic in Multi-Agent Systems*, (2013).

[2] Philippe Besnard and Anthony Hunter, *Elements of Argumentation*, MIT Press, 2008.

[3] Elena Cabrio and Serena Villata, 'Natural Language Arguments: A Combined Approach', in *European Conference on AI (ECAI)*, pp. 205–210, (2012).

[4] Federico Cerutti, Nava Tintarev, and Nir Oren, 'Formal Argumentation: A Human-centric Perspective', in *Eleventh International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2014)*, (2014).

[5] Federico Cerutti, Nava Tintarev, and Nir Oren. Human-Argumentation Experiment Pilot 2013. `http://goo.gl/FvgC08`, 2014.

[6] Carlos Iván Chesñevar, Ana Gabriela Maguitman, and Ronald Prescott Loui, 'Logical models of argument', *ACM Computing Surveys (CSUR)*, **32**(4), 337–383, (2000).

[7] Phan Minh Dung, 'On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games', *Artificial Intelligence*, **77**(2), 321–357, (1995).

[8] Sinan Egilmez, João Martins, and João Leite, 'Extending social abstract argumentation with votes on attacks', in *Workshop on Theory and Applications of Formal Argumentation*, (2013).

[9] Andy Field, *Discovering Statistics Using SPSS (Introducing Statistical Methods series)*, SAGE Publications Ltd, 2009.

[10] Alejandro J. García and Guillermo R. Simari, 'Defeasible logic programming: an argumentative approach', *Theory and Practice of Logic Programming*, **4**(1+2), 95–138, (2004).

[11] Thomas F. Gordon, Henry Prakken, and Douglas N. Walton, 'The Carneades model of argument and burden of proof', *Artificial Intelligence*, **171**(10-15), 875–896, (July 2007).

[12] Michael H.G. Hoffmann, 'Logical argument mapping: A method for overcoming cognitive problems of conflict management', *International Journal of Conflict Management*, **16**(4), 304–334, (2005).

[13] Sanjay Modgil, 'Reasoning about preferences in argumentation frameworks', *Artificial Intelligence*, **173**(9-10), 901–934, (2009).

[14] Sanjay Modgil and Henry Prakken, 'A general account of argumentation with preferences', *Artificial Intelligence*, **195**, 361–397, (2013).

[15] Bernard Moulin, Hengameh Irandoust, Micheline Belanger, and Gaëlle Desbordes, 'Explanation and Argumentation Capabilities: Towards the Creation of More Persuasive Agents', *Artificial Intelligence Review*, **17**, 169–222, (2002).

[16] MT. Amazon Mechanical Turk. http://www.mturk.com.

[17] John O'Hara, James Higgins, Stephen Fleger, and Valarie Barnes, 'Human-system interfaces to automatic systems: Review guidance and technical basis.', Technical Report BNL-91017-2010, Brookhaven National Laboratory, (2010).

[18] Henry Prakken and Giovanni Sartor, 'Argument-based extended logic programming with defeasible priorities', *Journal of Applied Non-Classical Logics*, **7**(1-2), 25–75, (1997).

[19] Alun Preece, Diego Pizzocaro, Dave Braines, David Mott, Geeth de Mel, and Tien Pham, 'Integrating hard and soft information sources for D2D using controlled natural language', in *Information Fusion (FUSION), 2012 15th International Conference on*, pp. 1330–1337, (2012).

[20] Iyad Rahwan, Mohammed Iqbal Madakkatel, Jean-Francois Bonnefon, Ruqiyabi Naz Awan, and Sherief Abdallah, 'Behavioural Experiments for Assessing the Abstract Argumentation Semantics for Reinstatement', in *Cognitive Science*, (2010).

[21] Sidney Siegel and N. John Castellan Jr., *Nonparametric Statistics for The Behavioral Sciences*, McGraw-Hill Humanities/Social Sciences/Languages, 1988.

[22] William Swartout, Cecile Paris, and Johanna Moore, 'Explanations in knowledge systems: Design for explainable expert systems', *IEEE Expert*, **6**(3), 58–64, (1991).

[23] Wilson L. Taylor, 'Cloze procedure: A new tool for measuring readability', *Journalism Quarterly,*, **30**, 415–433, (1953).

[24] Nava Tintarev, 'Explaining Recommendations', in *User Modeling*, pp. 470–474, (2007).

[25] Alice Toniolo, Federico Cerutti, Nir Oren, and Timothy J Norman, 'Reasoning about provenance for collaborative intelligence analysis', Technical report, (2013).

[26] Douglas N. Walton, Chris Reed, and Fabrizio Macagno, *Argumentation Schemes*, Cambridge University Press, NY, 2008.

[27] Adam Wyner, 'Arguments as Narratives', in *Proceedings of the Third Workshop on Computational Models of Narrative (CMN 2012)*, ed., Mark Finlayson, pp. 178–180, (2012).

[28] L Richard Ye, 'The value of explanation in expert systems for auditing: An experimental investigation', *Expert Systems with Applications*, **9**(4), 543–556, (1995).