Graphical Representation Enhances Human Compliance with Principles for Graded Argumentation Semantics

Srdjan Vesic
CNRS, Univ. Artois, CRIL
Lens, France
vesic@cril.fr

Bruno Yun
University of Aberdeen
Aberdeen, United Kingdom
bruno.yun@abdn.ac.uk

Predrag Teovanovic
University of Belgrade
Belgrade, Serbia
teovanovic@fasper.bg.ac.rs

1 INTRODUCTION

Group decision-making and negotiation are fundamental issues for modern multi-agent systems [18, 21]. In such a context, information exchange between agents is crucial for agents to coordinate and cooperate. Argumentation theory offers an intuitive interface for stating and explaining agents’ positions, allowing them to share or withhold their goals or intentions during the negotiation process.

Researchers in AI often point out that one of the strong points of the argumentation approach is that it uses the format which is intuitive and easy to grasp by humans [5, 6, 13, 28]. This is because arguments are often constructed by humans to defend or challenge a viewpoint in a debate or a dialogue. We often make decisions by relying on arguments in favor or against a particular action (e.g., buying an object). The scholars in artificial intelligence (AI) developed formal models of reasoning, dialogue, and decision-making based on argumentation.

Most of the recent approaches in argumentation are based on the formalism introduced by Dung in 1995 [12]. He represents the arguments as nodes of a graph where each directed edge models an attack between two arguments. Once an argumentation graph is generated, one can make use of one of the so-called “argumentation semantics”. A semantics allows calculating the set of extensions, which are the sets of arguments that can be accepted together. Those extensions represent the possible viewpoints in an argumentation debate. As a result, arguments can have three levels of acceptability. An argument is skeptically accepted if it belongs to all the extensions; credulously accepted if it belongs to at least one extension; rejected if it does not belong to any extension.

Researchers also introduced the so-called “ranking-based semantics” as a way to compare arguments from the least to the most contested. To navigate the plethora of the existing ranking-based semantics, desirable principles have been defined to classify them [1, 7, 27]. For example, the void precedence principle claims that every non-attacked argument is stronger than any attacked argument; the independence principle claims that two arguments that are not connected by attacks should not have any influence on each other’s ranking. It should be noted that there are more than 20 principles in ranking-based semantics [3, 7, 27].

We tackle three research questions in this paper. First, to the extent of our knowledge, the principles we mentioned have never been experimentally evaluated to check whether they are intuitive or acceptable for humans (and especially non-experts). Hence, we aim to examine if the argumentation principles proposed by the AI researchers realistically model human reasoning. In other words, if a given principle prescribes that a certain argument, say X, should be stronger than another argument Y in a certain context (i.e., argumentation graph), do the human participants also rank X as stronger than Y? The first goal of our experiment is to verify this.

Second, it is often claimed by the researchers in AI that argument-based models are easier to grasp by humans than logical formulas. In particular, they hypothesized that the graph-based representation is helpful to understand the problem and facilitate reasoning. Thus, the second aim of our study was to examine whether the participants who see graph-based arguments draw better conclusions than those who do not have access to the argumentation graph.

Third, we investigated if there is a positive manifold among compliances with different normative principles of ranking-based semantics, i.e., whether people who more frequently behave in accordance to expectations that derive from one principle (e.g., anonymity) are also prone to behave more frequently in accordance to expectations that derive from other principles such that independence, maximality, and void precedence among others [1, 7, 26]. Finally, we aimed to examine if these individual differences can be
predicted by more traditional psychological measures such as the Need for cognition, Faith in intuition, and Cognitive reflection.

This paper is structure as follows. In Section 2, we recall the argumentation setting and formally introduce the principles for graded semantics as well as the necessary notations. In Section 3, we provide the reader with details on the design of our experiments, including explanations on how the principle compliances and the cognitive styles were measured. In Section 4, we thoroughly present our results via item-level, scale-level, and correlational analyses. Lastly, in Section 5, we discuss our findings and reflect on their wider impact on the argumentation community and AI in general.

2 BACKGROUND

We start this section by recalling the definition of an argumentation graph as defined by Dung in his seminal paper [12].

Definition 2.1 (Argumentation graph). An argumentation graph is a pair \( \mathcal{AS} = (\mathcal{A}, \mathcal{C}) \), where \( \mathcal{A} \) is a finite set of arguments and \( \mathcal{C} \subseteq \mathcal{A} \times \mathcal{A} \) is a set of binary attacks between arguments. The set of attackers of \( a \in \mathcal{A} \) is \( \text{Att}(a) = \{b \in \mathcal{A} \mid (b, a) \in \mathcal{C}\} \).

Given an argumentation graph \( \mathcal{AS} = (\mathcal{A}, \mathcal{C}) \), an undirected path in \( \mathcal{AS} \) is a sequence \( (a_1, \ldots, a_n) \) such that for all \( i \in \{1, \ldots, n\} \), \( a_i \in \mathcal{A} \) and for all \( 1 \leq i \leq n - 1 \), \( (a_i, a_{i+1}) \in \mathcal{C} \) or \( (a_{i+1}, a_i) \in \mathcal{C} \). The connected components of \( \mathcal{AS} \), denoted \( \text{cc}(\mathcal{AS}) \), is the set of largest subgraphs of \( \mathcal{AS} \), such that two arguments are in the same component of \( \mathcal{AS} \) iff there is an undirected path between them.

In Dung’s approach of argumentation [12], the acceptability of arguments is based on their belonging to some sets, called extensions [4, 8, 12, 15]. Different semantics are used to calculate the extensions; we refer the reader to the work of Baroni et al. for an introduction on argumentation semantics [4]. A different approach consists in ranking the arguments from the most to the least acceptable ones depending on how much they are contested. The latter semantics are called ranking-based semantics and have been widely explored in the literature [1, 7, 26]. Indeed, there are two similar approaches: the ranking-based approach and the graded approach. A ranking-based semantics provides a rank (or an order) on arguments, for example \( a < c < b \), whereas a graded semantics provides the degrees (sometimes called scores), for example \( \text{Deg}(a) = 0 \), \( \text{Deg}(c) = 0.8 \), \( \text{Deg}(b) = 1 \), meaning that e.g. \( a \) has the acceptability degree 0. Note that each graded semantics naturally induces the corresponding ranking-based semantics.

Definition 2.2 (Graded semantics). A graded semantics is a function \( \sigma \) that takes as input any \( \mathcal{AS} = (\mathcal{A}, \mathcal{C}) \) and returns a function \( \text{Deg}_{\mathcal{AS}} : \mathcal{A} \rightarrow [0, 1] \). The notation \( \text{Deg}_{\mathcal{AS}}(a) \leq \text{Deg}_{\mathcal{AS}}(b) \) means that \( b \) is at least as acceptable as \( a \) w.r.t. \( \sigma \).

To perform a more systematic study of graded semantics, researchers in argumentation have defined several desirable principles that should govern the behaviour of these semantics. Since this is the first study that aims at assessing to which extent the principles are compatible with human reasoning, we focus on a limited subset of principles. To test the compliance we need several examples (at least five) per principle and to keep the experiment length within reasonable limits, we decided not to exceed five principles in this study. We choose the principles that were considered the simplest to model using up to four arguments to avoid large graphs, which are difficult to grasp and evaluate by non-experts. Note that although there are more restrictive versions of the anonymity and independence principles, where the isomorphic images of an argument need to have the same acceptability degree, we decided to encode the “ranking-based” version of the principles since humans tend to compare arguments between them, rather than associate absolute scores to each argument.

To define anonymity, we first introduce the notion of isomorphism between argumentation graphs.

Definition 2.3 (Isomorphism). An isomorphism between two argumentation graphs \( \mathcal{AS} = (\mathcal{A}, \mathcal{C}) \) and \( \mathcal{AS'} = (\mathcal{A'}, \mathcal{C'}) \) is a function \( \gamma : \mathcal{A} \rightarrow \mathcal{A'} \) such that for every \( (a, b) \in \mathcal{C} \) iff \( (\gamma(a), \gamma(b)) \in \mathcal{C'} \). With a slight abuse of notation, we use \( \mathcal{AS'} = \gamma(\mathcal{AS}) \).

Anonymity states that the names of the argument should not influence their acceptability.

Definition 2.4 (Anonymity). We say that a graded semantics \( \sigma \) satisfies anonymity iff for every two argumentation graphs \( \mathcal{AS} = (\mathcal{A}, \mathcal{C}) \) and \( \mathcal{AS'} = (\mathcal{A'}, \mathcal{C'}) \) such that \( \mathcal{AS'} = \gamma(\mathcal{AS}) \), it holds that for all \( a, b \in \mathcal{A} \), \( \text{Deg}_{\mathcal{AS}}^\sigma(a) \leq \text{Deg}_{\mathcal{AS}}^\sigma(b) \) iff \( \text{Deg}_{\mathcal{AS'}}^\sigma(\gamma(a)) \leq \text{Deg}_{\mathcal{AS'}}^\sigma(\gamma(b)) \).

As stated before, there is a more restrictive version of anonymity exists where the degree of each argument must be exactly the same as its isomorphic image, i.e., for all argument \( a \in \mathcal{A} \), \( \text{Deg}_{\mathcal{AS}}^\sigma(a) = \text{Deg}_{\mathcal{AS'}}^\sigma(\gamma(a)) \).

The independence principle states that the acceptability degree of an argument should only affected by other arguments in its connected component.

Definition 2.5 (Independence). We say that a graded semantics \( \sigma \) satisfies independence iff for every argumentation graph \( \mathcal{AS} \), for every \( \mathcal{AS'} = (\mathcal{A'}, \mathcal{C'}) \in \text{cc}(\mathcal{AS}) \) and for every \( a, b \in \mathcal{A'} \), if \( \text{Deg}_{\mathcal{AS}}^\sigma(a) \leq \text{Deg}_{\mathcal{AS}}^\sigma(b) \) then \( \text{Deg}_{\mathcal{AS'}}^\sigma(a) \leq \text{Deg}_{\mathcal{AS'}}^\sigma(b) \).

Void precedence states that an unattacked argument is more acceptable than an attacked one.

Definition 2.6 (Void precedence). We say that a graded semantics \( \sigma \) satisfies void precedence iff for every argumentation graph \( \mathcal{AS} = (\mathcal{A}, \mathcal{C}) \) and \( a, b \in \mathcal{A} \) such that \( \text{Att}(a) = \emptyset \) and \( \text{Att}(b) \neq \emptyset \) then \( \text{Deg}_{\mathcal{AS}}^\sigma(a) > \text{Deg}_{\mathcal{AS}}^\sigma(b) \).

Maximality states that an unattacked argument should have the maximum acceptability degree.

Definition 2.7 (Maximality). We say that a graded semantics \( \sigma \) satisfies maximality iff for every argumentation graph \( \mathcal{AS} = (\mathcal{A}, \mathcal{C}) \) and \( a \in \mathcal{A} \) such that \( \text{Att}(a) = \emptyset \) then \( \text{Deg}_{\mathcal{AS}}^\sigma(a) = 1 \).

We now present the details of the experimental setting.

3 EXPERIMENTAL DESIGN

We recruited a total of 98 participants (91 females) from a pool of master’s students that attended statistics course at the Faculty for Special Education and Rehabilitation at the University of Belgrade. The mean age was 25.92 years (SD = 4.25), the participants did not have any previous knowledge about computer science nor argumentation theory, and all participants received course credit for taking part in the study.
Each participant was randomly assigned to one of two experimental groups. The first group was presented with both textual and graphical representations of arguments \((n = 57)\), while the second group was presented with only textual arguments \((n = 41)\). For example, the first group was shown the text from Example 3.1 and the graphical representation from Figure 1, while the second group was only shown the textual arguments from Example 3.1.

**Example 3.1.** Let us consider the following four textual arguments (A, B, C, and D) about the Zuber transportation company (which graphical representation is provided in Figure 1).

- A: The train 4147 at 7.45 am for Paris will be on time. I checked this morning on the Zuber transportation website.
- B: The train 4147 at 7.45 am for Paris will be 20 minutes late. I checked this morning on the Zuber transportation mobile app.
- C: The Zuber company has been reported to show incorrect information recently due to hackers’ activity that targeted their database.
- D: The Zuber company stocks have dropped in value since the recent hacker attack.

For each use-case, we designed the corresponding argumentation graph by following the well established rules from argumentation theory. In particular, we chose the types of attack relations (e.g. attacking an explicit generic) that are the easiest to understand by humans [11] to insure that their mental representation matches the graph displayed. The experiment was split in four major parts:

1. First, we explained to participants what an argument is and what attacks are. This was done through a short tutorial that consisted of three examples. Each example was composed of a text describing three arguments and the corresponding graphical representation with the arguments and the attacks between them. Let us present, in Example 3.2 below, one of the examples from the tutorial.

**Example 3.2.** Let us consider the following three textual arguments (A, B, and C) about Staphylococcus (which graphical representation is provided in Figure 2).

- A: Smith et al. have published a paper in 2013 that concludes that cyclic antibiotics (and only them) can treat Staphylococcus.
- B: Doe et al. have published a paper in 2013 that concludes that non-cyclic antibiotics (and only them) can treat Staphylococcus.
- C: Wang et al. published a paper in 2016 that corrects the mistake of Doe et al. and concludes that cyclic antibiotics can treat Staphylococcus.

2. Second, we gave the participants three control questions to test their understanding of the notions of argument and attack. In each question, the participants were first shown a text (in natural language) describing three arguments as well as three different argumentation graphs. Then, they were asked to choose which graph corresponded to the given text. In Example 3.3, we present one of the control questions.

**Example 3.3.** Let us consider the following three textual arguments (A, B, and C) about the whereabouts of Sophie Schmoe. The three graphical representations given to the participants are represented in Figure 3.

- A: Joe Bloggs made a statement that he saw Sophie Schmoe at the Palais Garnier yesterday at 8 pm.
- B: Romana Leech says that there was a protest in front of the Eiffel tower yesterday.
- C: Dick Harry declared that he was with Sophie Schmoe at the Champs Elysée yesterday at 8 pm and that Romana Leech is a compulsive liar so her statements cannot be trusted.

3. Third, the participants were asked to perform 16 different tasks \((T_1 \text{ to } T_{16})\). Our goal was to check the participants’ level of compliance with the different argumentation principles. More details about how each individual principle was assessed will be given in Section 3.1. In each task, the participants were shown between two and four textual arguments and the corresponding argumentation graph if the participant was assigned in the graph group, otherwise, the participant was only shown the textual arguments (no-graph group). Then, the participants were asked to estimate the strength of each argument by using a 4-point Likert scale: 1 (very weak), 2 (weak), 3 (strong), and 4 (very strong).

4. Lastly, participants were presented with three cognitive reflection test tasks [14] (see Section 3.2.1) following five statements assessing need for cognition \((NFC_1 \text{ to } NFC_5)\) and five statements assessing faith in intuition \((FI_1 \text{ to } FI_5)\) that were part of the rational-experiential inventory [20] (see Table 1).
I do not like to have to do a lot of thinking. (R)

I try to avoid situations that require thinking in depth about something. (R)

I prefer to do something that challenges my thinking abilities rather than something that requires little thought.

I prefer complex to simple problems.

Thinking hard and for a long time about something gives me little satisfaction. (R)

I trust my initial feelings about people.

My initial impressions of people are almost always right.

When it comes to trusting people, I can usually rely on my "gut feelings".

I can usually feel when a person is right or wrong even if I can’t explain how I know.

Table 1: Statements shown to participants. (R) is used to denote reversely scored items

| NFC1 | I do not like to have to do a lot of thinking. (R) |
| NFC2 | I try to avoid situations that require thinking in depth about something. (R) |
| NFC3 | I prefer to do something that challenges my thinking abilities rather than something that requires little thought. |
| NFC4 | I prefer complex to simple problems. |
| NFC5 | Thinking hard and for a long time about something gives me little satisfaction. (R) |
| FI1 | I trust my initial feelings about people. |
| FI2 | My initial impressions of people are almost always right. |
| FI3 | When it comes to trusting people, I can usually rely on my "gut feelings". |
| FI4 | I can usually feel when a person is right or wrong even if I can’t explain how I know. |

3.1 Evaluating Principle Compliance

In this subsection, we give more details on how the participants’ compliance with the graded semantics principles studied in this paper have been assessed.

3.1.1 Anonymity. This principle claims that the ranking of arguments will remain the same, independently of the arguments’ content, as long as the structure of the argumentation graph is preserved. We study two variants of anonymity. The first variant, which we call anonymity between tasks is more general and requires an isomorphism between any two graphs with the same structure. Five pairs of tasks were used to measure this principle (ABT_1 to ABT_5; see first row of Table 2). For each pair, both tasks consist of the same number of arguments (ranging from two to four) such that the topology of the graph is preserved. Formally, the argumentation graphs can be represented using the same graph if the arguments are renamed. To illustrate, let us present one pair of tasks (see Example 3.4 and 3.5 below).

Example 3.4. Let us consider the following three textual arguments (A, B, and C) about the painter Philippe Zhao (which graphical representation is provided on the left side of Figure 4):

A: The historian Petrović deduced from the recently found letters of the painter Philippe Zhao that he only had one wife and she was Serbian.

B: The historian Albin deduced from the recently found letters of the painter Philippe Zhao that he only had one wife and she was Polish.

C: Roger says that Philippe Zhao had multiple wives in his life.

Note that both A and B attack C since each of them contains a justification, and also since a historian is more trustworthy than an average person (Roger).

Figure 4: Anonymity states that the strength of C in the graph from Example 3.4 (left) should be the same as the strength of A in the graph from Example 3.5 (right).

Example 3.5. Let us consider the following three textual arguments (A, B, and C) about Charlotte’s murder (which graphical representation is provided on the right side of Figure 4):

A: Carla’s son says that she is innocent.

B: Detective Decker recently contradicted Carla’s son and found that Carla killed Charlotte with a pistol.

C: Detective Dan recently contradicted Carla’s son and found that Carla killed Charlotte with a knife.

B and C mutually attack each other and each of them attacks A.

Note that the graphs in the two tasks are “the same” (two mutually conflicting arguments attacking the third argument) except for the arguments’ names and content. The anonymity principle says that the evaluation of the arguments should depend only on the graph and not on the arguments’ names and content. For instance, if a participant claims that A and B are equally strong and are both stronger than C in the first task, according to anonymity, they should claim that in the second task, B and C are equally strong, and are both stronger than A.

The second variant of the anonymity principle, called anonymity within tasks, requires a graph when there exists an isomorphism (different from identity) from the graph to itself. Fourteen tasks were employed to measure this principle (AWT_1 to AWT_14; second row in Table 2). For each task, we compared the acceptability values attached to the arguments by the participants and compared them with the different arguments that are isomorphic within the graph. For instance, in Example 3.4, A and B mutually attack each other and they both attack C, anonymity within tasks says that A and B should have the same acceptability degree.

3.1.2 Independence. This principle claims that two arguments that are not connected by attacks should not have any influence on each other’s ranking. Five pairs of tasks were used in our study (IND_1 to IND_5; third row in Table 2). For each pair of tasks, one task consisted of two or three arguments, whereas the other task consisted of the same arguments with one or two additional arguments that were not conflicting with (and thus, not connected to) the existing arguments. Formally, there was no attack between any of the existing arguments from the first task and the new arguments introduced in the second task - therefore, this principle states that the ranking of the arguments from the first task should remain unchanged after the introduction of new unrelated arguments.

In Example 3.6, we illustrate one pair of tasks. The first task was composed of two mutually conflicting arguments A and B. In the second task, the participant is shown the previous two arguments.

This distinction is new and does not exist in the literature.
<table>
<thead>
<tr>
<th>Principle</th>
<th>Item description</th>
<th>Percentage of correct answers</th>
<th>Test of difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No-graph group (n = 57)</td>
<td>Graph group (n = 41)</td>
</tr>
<tr>
<td>Anonymity between tasks</td>
<td>ABT_1 (T₃ – T₁₀)</td>
<td>48.8%</td>
<td>68.4%</td>
</tr>
<tr>
<td></td>
<td>ABT_2 (T₅ – T₇)</td>
<td>31.7%</td>
<td>49.1%</td>
</tr>
<tr>
<td></td>
<td>ABT_3 (T₆ – T₁₂)</td>
<td>26.8%</td>
<td>73.7%</td>
</tr>
<tr>
<td></td>
<td>ABT_4 (T₀ – T₁₄)</td>
<td>7.3%</td>
<td>64.9%</td>
</tr>
<tr>
<td></td>
<td>ABT_5 (T₁₅ – T₁₆)</td>
<td>9.8%</td>
<td>61.4%</td>
</tr>
<tr>
<td>Anonymity within tasks</td>
<td>AWT_1 T₁ (A = B)</td>
<td>95.1%</td>
<td>98.2%</td>
</tr>
<tr>
<td></td>
<td>AWT_2 T₃ (A = B)</td>
<td>87.8%</td>
<td>93.0%</td>
</tr>
<tr>
<td></td>
<td>AWT_3 T₄ (A = B)</td>
<td>61.0%</td>
<td>87.7%</td>
</tr>
<tr>
<td></td>
<td>AWT_4 T₅ (A = B)</td>
<td>90.2%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>AWT_5 T₇ (A = B)</td>
<td>58.5%</td>
<td>94.4%</td>
</tr>
<tr>
<td></td>
<td>AWT_6 T₈ (A = B = C = D)</td>
<td>26.8%</td>
<td>68.4%</td>
</tr>
<tr>
<td></td>
<td>AWT_6a T₆ (A = B)</td>
<td>70.7%</td>
<td>91.2%</td>
</tr>
<tr>
<td></td>
<td>AWT_6b T₆ (C = D)</td>
<td>48.8%</td>
<td>91.2%</td>
</tr>
<tr>
<td></td>
<td>AWT_7 T₉ (A = B = C)</td>
<td>24.4%</td>
<td>82.5%</td>
</tr>
<tr>
<td></td>
<td>AWT_8 T₁₀ (B = C)</td>
<td>85.4%</td>
<td>93.0%</td>
</tr>
<tr>
<td></td>
<td>AWT_9 T₁₁ (A = B = C)</td>
<td>90.2%</td>
<td>94.7%</td>
</tr>
<tr>
<td></td>
<td>AWT_10 T₁₂ (A = B = C = D)</td>
<td>58.5%</td>
<td>87.7%</td>
</tr>
<tr>
<td></td>
<td>AWT_10a T₁₂ (A = B)</td>
<td>95.1%</td>
<td>98.2%</td>
</tr>
<tr>
<td></td>
<td>AWT_10b T₁₂ (C = D)</td>
<td>95.1%</td>
<td>98.2%</td>
</tr>
<tr>
<td></td>
<td>AWT_11 T₁₃ (A = B)</td>
<td>58.5%</td>
<td>94.7%</td>
</tr>
<tr>
<td></td>
<td>AWT_12 T₁₄ (B = C = D)</td>
<td>90.2%</td>
<td>96.5%</td>
</tr>
<tr>
<td></td>
<td>AWT_13 T₁₅ (A = B)</td>
<td>58.5%</td>
<td>91.2%</td>
</tr>
<tr>
<td></td>
<td>AWT_14 T₁₆ (A = B)</td>
<td>92.7%</td>
<td>94.7%</td>
</tr>
<tr>
<td>Independence</td>
<td>IND_1 T₁ (T₂)</td>
<td>90.2%</td>
<td>98.2%</td>
</tr>
<tr>
<td></td>
<td>IND_2 T₁ (T₂)</td>
<td>92.7%</td>
<td>96.3%</td>
</tr>
<tr>
<td></td>
<td>IND_3 T₁ (T₂)</td>
<td>53.7%</td>
<td>73.7%</td>
</tr>
<tr>
<td></td>
<td>IND_4 T₁ (T₂)</td>
<td>39.0%</td>
<td>82.5%</td>
</tr>
<tr>
<td></td>
<td>IND_5 T₁ (T₂)</td>
<td>90.2%</td>
<td>93.0%</td>
</tr>
<tr>
<td>Void precedence</td>
<td>VP_1 T₂ (C &gt; A, B)</td>
<td>58.5%</td>
<td>87.7%</td>
</tr>
<tr>
<td></td>
<td>VP_2 T₃ (C, D &gt; A, B)</td>
<td>51.2%</td>
<td>59.6%</td>
</tr>
<tr>
<td></td>
<td>VP_3 T₅ (D &gt; A, B, C)</td>
<td>61.0%</td>
<td>77.2%</td>
</tr>
<tr>
<td></td>
<td>VP_4 T₆ (D &gt; A, B, C)</td>
<td>41.5%</td>
<td>68.4%</td>
</tr>
<tr>
<td></td>
<td>VP_5 T₇ (C &gt; A, B, D)</td>
<td>31.7%</td>
<td>73.7%</td>
</tr>
<tr>
<td></td>
<td>VP_6 T₉ (B &gt; A, C, D)</td>
<td>17.1%</td>
<td>61.4%</td>
</tr>
<tr>
<td></td>
<td>VP_7 T₁₃ (C &gt; A, B)</td>
<td>63.4%</td>
<td>75.4%</td>
</tr>
<tr>
<td></td>
<td>VP_8 T₁₄ (A &gt; B, C, D)</td>
<td>51.2%</td>
<td>74.4%</td>
</tr>
<tr>
<td></td>
<td>VP_9 T₁₅ (D &gt; A, B, C)</td>
<td>43.9%</td>
<td>71.9%</td>
</tr>
<tr>
<td></td>
<td>VP_10 T₁₆ (D &gt; A, B, C)</td>
<td>53.7%</td>
<td>78.9%</td>
</tr>
<tr>
<td>Maximaliy</td>
<td>MAX_1 T₂ (C = 4)</td>
<td>39.0%</td>
<td>61.4%</td>
</tr>
<tr>
<td></td>
<td>MAX_2 T₃ (C = 4, D = 4)</td>
<td>14.6%</td>
<td>29.8%</td>
</tr>
<tr>
<td></td>
<td>MAX_3 T₅ (D = 4)</td>
<td>53.7%</td>
<td>68.4%</td>
</tr>
<tr>
<td></td>
<td>MAX_4 T₆ (C = 4, D = 4)</td>
<td>12.2%</td>
<td>21.1%</td>
</tr>
<tr>
<td></td>
<td>MAX_5 T₇ (C = 4)</td>
<td>41.5%</td>
<td>64.9%</td>
</tr>
<tr>
<td></td>
<td>MAX_6 T₈ (B = 4)</td>
<td>22.0%</td>
<td>57.9%</td>
</tr>
<tr>
<td></td>
<td>MAX_7 T₁₃ (C = 4)</td>
<td>41.5%</td>
<td>52.6%</td>
</tr>
<tr>
<td></td>
<td>MAX_8 T₁₄ (A = 4)</td>
<td>56.1%</td>
<td>71.9%</td>
</tr>
<tr>
<td></td>
<td>MAX_9 T₁₅ (D = 4)</td>
<td>41.5%</td>
<td>71.9%</td>
</tr>
<tr>
<td></td>
<td>MAX_10 T₁₆ (D = 4)</td>
<td>65.9%</td>
<td>71.9%</td>
</tr>
</tbody>
</table>

Table 2: Percentage of correct answers in two groups and tests of difference. ABT_1 stands for the first pair of tasks that is used to check the compliance with anonymity between tasks. T₃ – T₁₀ means that tasks 3 and 10 are used in ABT_1. A = B means that we check whether the strength of argument A is equal to the strength of argument B. The notation C > A, B means that we check if C is stronger than A and B. Regarding maximality, C = 4 means that we check if C has the maximal strength.
and the following additional arguments C and D. Since newly introduced arguments C and D are not related to A and B, the rating of A and B should remain unchanged, i.e. if a participant claimed that A is as strong as B in the first task, they should claim that A is still as strong as B in the second task (after C and D are added).

**Example 3.6.** Let us consider the following four textual arguments (A, B, C and D) about Lady Gaga (which graphical representation is provided in Figure 5).

A : Benedicte says that the dress worn by Lady Gaga this morning is green.
B : Michael says that the dress worn by Lady Gaga this morning is red.
C : Lady Gaga was nominated for the Liechtenstein Neuroscience Association award this year.
D : Liechtenstein Neuroscience Association committee announced that only scientists can be nominated for their award.

![Figure 5](image)

**Figure 5:** Two graphs shown in two different tasks. Independence states that the ranking between A and B should be the same in the two graphs.

3.1.3 **Void precedence.** This principle claims that every non-attacked argument should be stronger than any attacked argument. Ten tasks were employed to measure this principle. For each task, we checked that arguments that were not attacked were ranked higher than any arguments that were attacked by at least one other argument.

**Example 3.7.** Let us consider the following four textual arguments (A, B, C and D) about tennis (which graphical representation is provided in Figure 6).

A : John thinks that each tennis game should end after one player wins three sets.
B : Pierre thinks that each tennis game should end after one player wins two sets.
C : Gerhard claims that the players will be too tired at the end of the season if all the tournaments are played on three sets.
D : Ichiro says that the Association of Tennis Professionals (ATP) should provide more money for young players since the sponsors have too much impact.

![Figure 6](image)

**Figure 6:** Void precedence states that C and D should be ranked higher than A and B.

Here, in addition to the mutual conflicts between A and B, C attacks A but is not attacked by the other arguments. The void precedence principle states that non-attacked arguments, here C and D, should be ranked strictly higher than A and B.

3.1.4 **Maximality.** This principle claims that non-attacked arguments should have the maximal acceptability value. Ten tasks were employed to measure this principle. In each task, we checked whether all non-attacked arguments had degree 4 (very strong). For instance, in Example 3.7, since arguments C and D are not attacked, we checked if the participant ranked C and D as “very strong”.

3.2 **Measures of Cognitive Styles**

3.2.1 **Cognitive reflection.** Cognitive reflection was assessed via the cognitive reflection test (CRT [14]), consisting of three items that cue a fast but incorrect response. One example of an item is the following question: “A racket and a ball cost 1100 RSD (Serbian currency) in total. The racket costs 1000 RSD more than the ball. How much does the ball cost?”. Although the correct answer is “50 RSD” approximately 60% of participants answered “100 RSD” (an incorrect but intuitive answer). A total score was calculated as a sum of intuitive responses (Cronbach’s α = 0.73 indicating fair level of internal consistency).

3.2.2 **Need for cognition and faith in intuition.** Those styles were assessed by employing a short, 10-item form of rational-experiential inventory [20] (see Table 1). Each subscale consists of five items, and participants were instructed to assess statements by using a 5-point Likert-type scale ranging from 1 (completely disagree) to 5 (completely agree). Scores for each subscale were calculated as average item values after reversely coding data for negatively formulated items. Cronbach’s α = 0.87 for faith in intuition subscale was quite high, but it was much lower for need for cognition subscale (Cronbach’s α = 0.46) indicating suboptimal level of internal consistency for these measures.

4 **EXPERIMENTAL RESULTS**

Results presented in Table 2 relate to the item-level of analysis. As described in Section 3, for each of the five principles, several (pairs of) tasks, ranging from five to fourteen, were used to examine if people behave in accordance with these principles. For three principles (independence, anonymity between tasks and anonymity within tasks), consistency of responses was used as a criterion of normative behavior, while response accuracy was used as a criterion for two other principles (void precedence and maximality). Principles and related (pairs of) items are presented in the first two columns of Table 2. In the third and fourth columns of Table 2, we show the percentage of normatively correct responses in the no-graph and graph groups respectively. Results indicate that anonymity between tasks in the no-graph group was the hardest principle to comply with; with a percentage of correct answers ranging from 7.3% to 48.8%, while participants behaved the most frequently in accordance to anonymity between tasks in the graph group; with a percentage of correct answers ranging from 68.4% to 100%. Furthermore, the percentage of normatively correct responses was higher in the graph group in comparison to the no-graph group on all 44 items, and this difference was statistically significant (ps < 0.05) on 27 items. The mean effect of graph representation on the participants’ performance across items was r = 0.27.

Results of scale-level analyses, presented in Table 3, further demonstrate that graphical representation of argument structure significantly enhances participants’ performance. The effect of
We showed that participants exhibit a higher level of compliance with the principles when they are provided the graphical representation of arguments not only enhances participants’ performance in a way that it amplifies compliance with principles, but it also increases the reliability of people’s behavior (see Tables 3 and 5). Namely, their responses were more consistent in two ways. First, as shown by Cronbach’s alpha coefficients, internal consistency of measures of anonymity between tasks and void precedence was much higher in the graph group. In other words, a participant who solves correctly one of void precedence tasks (resp. tasks intended to measure anonymity between tasks) has higher chance to solve correctly other void precedence tasks (resp. tasks intended to measure anonymity between tasks). Second, as indicated by the coefficients of correlations and results of exploratory factor analysis, overall strength of relations between measures of adherence to different principles was much higher in the graph group. In other words, the person who complies with independence (or any other principle) has a higher chance to comply with all other principles. Taken together, correlational results indicate that people were not always choose arguments that are justified according to the given semantic choice. Cetutti et al. [9] showed that although there is 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, there are some significant deviations, which appear to arise from implicit knowledge”. We refer the reader to the survey of empirical cognitive studies about formal argumentation by Cetutti et al. [9].

In future work, we plan to explore whether weighted argumentation graphs [2] could better model human reasoning. The idea is to consider not only arguments and attacks, but also arguments’ initial weights. An initial weight represents the intrinsic strength of an argument which does not take into account the graph and the attacks but only internal factors, for example, the trust in the source providing the argument or the participant’s belief in an argument.

We already explained that participants exhibit a higher level of compliance with the principles when they are provided the graphical representation of the arguments. Furthermore, graphical representation of arguments not only enhances participants’ performance in a way that it amplifies compliance with principles, but it also increases the reliability of people’s behavior (see Tables 3 and 5). Namely, their responses were more consistent in two ways. First, as shown by Cronbach’s alpha coefficients, internal consistency of measures of anonymity between tasks and void precedence was much higher in the graph group. In other words, a participant who solves correctly one of void precedence tasks (resp. tasks intended to measure anonymity between tasks) has higher chance to solve correctly other void precedence tasks (resp. tasks intended to measure anonymity between tasks). Second, as indicated by the coefficients of correlations and results of exploratory factor analysis, overall strength of relations between measures of adherence to different principles was much higher in the graph group. In other words, the person who complies with independence (or any other principle) has a higher chance to comply with all other principles. Taken together, correlational results indicate that people were not always choose arguments that are justified according to the given semantic choice. Cetutti et al. [9] showed that although there is 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, there are some significant deviations, which appear to arise from implicit knowledge”. We refer the reader to the survey of empirical cognitive studies about formal argumentation by Cetutti et al. [9].

In future work, we plan to explore whether weighted argumentation graphs [2] could better model human reasoning. The idea is to consider not only arguments and attacks, but also arguments’ initial weights. An initial weight represents the intrinsic strength of an argument which does not take into account the graph and the attacks but only internal factors, for example, the trust in the source providing the argument or the participant’s belief in an argument.

We already explained that participants exhibit a higher level of compliance with the principles when they are provided the graphical representation of the arguments. Furthermore, graphical representation of arguments not only enhances participants’ performance in a way that it amplifies compliance with principles, but it also increases the reliability of people’s behavior (see Tables 3 and 5). Namely, their responses were more consistent in two ways. First, as shown by Cronbach’s alpha coefficients, internal consistency of measures of anonymity between tasks and void precedence was much higher in the graph group. In other words, a participant who solves correctly one of void precedence tasks (resp. tasks intended to measure anonymity between tasks) has higher chance to solve correctly other void precedence tasks (resp. tasks intended to measure anonymity between tasks). Second, as indicated by the coefficients of correlations and results of exploratory factor analysis, overall strength of relations between measures of adherence to different principles was much higher in the graph group. In other words, the person who complies with independence (or any other principle) has a higher chance to comply with all other principles. Taken together, correlational results indicate that people were not always choose arguments that are justified according to the given semantic choice. Cetutti et al. [9] showed that although there is 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, there are some significant deviations, which appear to arise from implicit knowledge”. We refer the reader to the survey of empirical cognitive studies about formal argumentation by Cetutti et al. [9].
Our results are aligned with other studies recommending the usage of collaborative computer-supported argument visualisation tools [17], such as Rationale (https://www.rationaleonline.com/) or Kialo (https://www.kialo.com/).

One remaining open question is, however, how does one provide such a graph in a real-life setting? Are the participants able to provide it themselves and what mechanisms can we use to ensure the correctness of the created graph, i.e., that the graph models the problem in question. We plan to address those questions as a part of future work. Namely, we plan to investigate how well people translate arguments and attacks into a graphical representation and whether drawing the graph themselves improves their compliance with the reasoning principles. Note that although a similar approach was used by Cramer and Guillaume [11], they only focused on the human perception of the directionality of attacks and not on how drawing the graph could improve cognitive reasoning. From the control tasks, we showed that roughly 75% of the participants choose the appropriate graph among the three offered.
REFERENCES


