Graphical Representation Enhances Human Compliance with Principles for Graded Argumentation Semantics

Srdjan Vesic, Bruno Yun, Predrag Teovanovic

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

ABSTRACT

We examined principles of graded argumentation semantics (independence, anonymity, void precedence, and maximality) to explore if (a) they realistically model human reasoning, (b) graphical representation of arguments facilitates compliance with the principles, (c) there is a positive correlation between compliance with different principles, and (d) this compliance is related to cognitive reflection, need for cognition and faith in intuition. The participants (N = 96) were randomly assigned to one of two experimental conditions - the graph group was presented with textual and graphical representations, while the second group was presented only with textual arguments. Our results indicate that there are major differences in the compliance with the several argumentation principles studied in this paper. However, compliance with argumentation principles was consistently better and more consistent in the graph group. Moreover, cognitive reflection correlated with compliance to some principles, but only in the graph group.

KEYWORDS

Argumentation; Human Reasoning; Principles; Graded Semantics

ACM Reference Format:
Srdjan Vesic, Bruno Yun, and Predrag Teovanovic. 2022. Graphical Representation Enhances Human Compliance with Principles for Graded Argumentation Semantics. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 9 pages.

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.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

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 $\mathcal{AS} = (\mathcal{A}, C)$, where $\mathcal{A}$ is a finite set of arguments and $C \subseteq \mathcal{A} \times \mathcal{A}$ is a set of binary attacks between arguments. The set of attackers of $a \in \mathcal{A}$ is $Att(a) = \{b \in \mathcal{A} \mid (b, a) \in C\}$.

Given an argumentation graph $\mathcal{AS} = (\mathcal{A}, C)$, 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 $1 \leq j \leq n-1$, $(a_j, a_{j+1}) \in C$ or $(a_{j+1}, a_j) \in C$. The connected components of $\mathcal{AS}$, denoted $cc(\mathcal{AS})$, 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 $a \prec c \prec b$, whereas a graded semantics provides the degrees (sometimes called scores), for example $Deg(a) = 0$, $Deg(c) = 0.8$, $Deg(b) = 1$, 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 σ that takes as input any $\mathcal{AS} = (\mathcal{A}, C)$ and returns a function $Deg^{\sigma}_{\mathcal{AS}} : \mathcal{A} \to [0, 1]$. The notation $Deg^{\sigma}_{\mathcal{AS}}(a) \leq Deg^{\sigma}_{\mathcal{AS}}(b)$ means that b is at least as acceptable as a w.r.t. σ.

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 graph $\mathcal{AS} = (\mathcal{A}, C)$ and $\mathcal{AS}' = (\mathcal{A}', C')$ is a function $\gamma : \mathcal{A} \to \mathcal{A}'$ such that for every $(a, b) \in C$ iff $(\gamma(a), \gamma(b)) \in C'$. 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 σ satisfies anonymity iff for every two argumentation graphs $\mathcal{AS} = (\mathcal{A}, C)$ and $\mathcal{AS}' = (\mathcal{A}', C')$ such that $\mathcal{AS}' = \gamma(\mathcal{AS})$, it holds that for all $a, b \in \mathcal{A}$, $Deg^{\sigma}_{\mathcal{AS}}(a) \leq Deg^{\sigma}_{\mathcal{AS}}(b)$ iff $Deg^{\sigma}_{\mathcal{AS}'}(\gamma(a)) \leq Deg^{\sigma}_{\mathcal{AS}'}(\gamma(b))$.

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 $a \in \mathcal{A}$, $Deg^{\sigma}_{\mathcal{AS}}(a) = Deg^{\sigma}_{\mathcal{AS}'}(\gamma(a))$.

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 σ satisfies independence iff for every argumentation graph $\mathcal{AS}$, for every $\mathcal{AS}' = (\mathcal{A}', C') \in cc(\mathcal{AS})$ and for every $a, b \in \mathcal{A}'$, if $Deg^{\sigma}_{\mathcal{AS}'}(a) \leq Deg^{\sigma}_{\mathcal{AS}'}(b)$ then $Deg^{\sigma}_{\mathcal{AS}}(a) \leq Deg^{\sigma}_{\mathcal{AS}}(b)$.

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 σ satisfies void precedence iff for every argumentation graph $\mathcal{AS} = (\mathcal{A}, C)$ and $a, b \in \mathcal{A}$ such that $Att(a) = \emptyset$ and $Att(b) \neq \emptyset$ then $Deg^{\sigma}_{\mathcal{AS}}(a) > Deg^{\sigma}_{\mathcal{AS}}(b)$.

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

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

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.

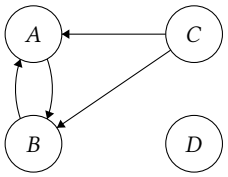

Figure 1: Graphical representation of natural language arguments and their attacks for the Zuber use-case.

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.

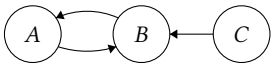

Figure 2: Graphical representation of natural language arguments and their attacks for the Staphylococcus use-case.

(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.

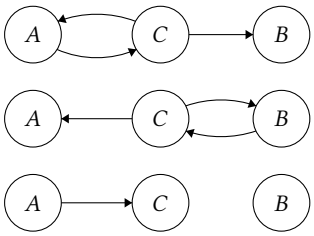

Figure 3: Graphical representation of the three graphical representations (top, middle, bottom) proposed in a control question.

(3) Third, the participants were asked to perform 16 different tasks (T_1 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 to NFC_5) and five statements assessing faith in intuition (FI_1 to FI_5) that were part of the rational-experiential inventory [20] (see Table 1).

NFC_1	I do not like to have to do a lot of thinking. (R)
NFC_2	I try to avoid situations that require thinking in depth about something. (R)
NFC_3	I prefer to do something that challenges my thinking abilities rather than something that requires little thought.
NFC_4	I prefer complex to simple problems.
NFC_5	Thinking hard and for a long time about something gives me little satisfaction. (R)
FI_1	I trust my initial feelings about people.
FI_2	I believe in trusting my hunches.
FI_3	My initial impressions of people are almost always right.
FI_4	When it comes to trusting people, I can usually rely on my "gut feelings".
FI_5	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

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[1]. 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).

[1]This distinction is new and does not exist in the literature.

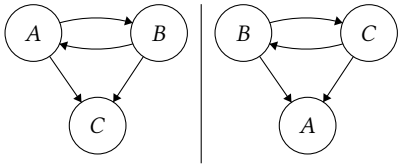

Figure 4: Anonymity states that the strength of C in the graph from Example 3.4 (left) should be the same as the strength of A in the graph from Example 3.5 (right).

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 contents. The anonymity principle says that the evaluation of the arguments should depend only on the graph and not on the arguments' names and their 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

Principle	Item description	Percentage of correct answers		Test of difference		
		No-graph group ($n = 57$)	Graph group ($n = 41$)	χ^2	p	r_φ
Anonymity between tasks	ABT_1 ($T_3 - T_{10}$)	48.8%	68.4%	3.84	0.05	0.20
	ABT_2 ($T_5 - T_7$)	31.7%	49.1%	2.97	0.08	0.08
	ABT_3 ($T_8 - T_{12}$)	26.8%	73.7%	21.08	$< .001$	0.46
	ABT_4 ($T_9 - T_{14}$)	7.3%	64.9%	32.75	$< .001$	0.58
	ABT_5 ($T_{15} - T_{16}$)	9.8%	61.4%	26.55	$< .001$	0.52
Anonymity within tasks	AWT_1 T_1 ($A = B$)	95.1%	98.2%	0.78	0.38	0.09
	AWT_2 T_3 ($A = B$)	87.8%	93.0%	0.77	0.38	0.09
	AWT_3 T_4 ($A = B$)	61.0%	87.7%	9.50	0.002	0.31
	AWT_4 T_5 ($A = B$)	90.2%	100%	5.80	0.02	0.24
	AWT_5 T_7 ($A = B$)	58.5%	86.0%	9.44	0.02	0.31
	AWT_6 T_8 ($A = B = C = D$)	26.8%	68.4%	16.51	< 0.001	0.41
	AWT_6a T_8 ($A = B$)	70.7%	91.2%	7.00	0.008	0.27
	AWT_6b T_8 ($C = D$)	48.8%	91.2%	22.04	< 0.001	0.47
	AWT_7 T_9 ($A = B = C$)	24.4%	82.5%	32.04	< 0.001	0.58
	AWT_8 T_{10} ($B = C$)	85.4%	93.0%	1.51	0.22	0.12
	AWT_9 T_{11} ($A = B = C$)	90.2%	94.7%	0.73	0.39	0.09
	AWT_10 T_{12} ($A = B = C = D$)	58.5%	87.7%	10.98	0.001	0.34
	AWT_10a T_{12} ($A = B$)	95.1%	98.2%	0.78	0.37	0.09
	AWT_10b T_{12} ($C = D$)	95.1%	98.2%	0.78	0.37	0.09
	AWT_11 T_{13} ($A = B$)	58.5%	94.7%	19.24	< 0.001	0.44
	AWT_12 T_{14} ($B = C = D$)	90.2%	96.5%	1.62	0.20	0.13
	AWT_13 T_{15} ($A = B$)	58.5%	91.2%	14.64	< 0.001	0.39
	AWT_14 T_{16} ($A = B$)	92.7%	94.7%	0.18	0.67	0.04
Independence	IND_1 ($T_1 - T_5$)	90.2%	98.2%	3.15	0.08	0.18
	IND_2 ($T_1 - T_{12}$)	92.7%	96.5%	0.71	0.39	0.08
	IND_3 ($T_2 - T_6$)	53.7%	73.7%	4.22	0.04	0.21
	IND_4 ($T_4 - T_{13}$)	39.0%	82.5%	19.59	$< .001$	0.45
	IND_5 ($T_{11} - T_{14}$)	90.2%	93.0%	0.24	0.63	0.05
Void precedence	VP_1 ($T_2; C > A, B$)	58.5%	87.7%	10.98	0.001	0.34
	VP_2 ($T_4; C, D > A, B$)	51.2%	59.6%	0.69	0.41	0.08
	VP_3 ($T_5; D > A, B, C$)	61.0%	77.2%	3.01	0.08	0.17
	VP_4 ($T_6; C, D > A, B$)	41.5%	68.4%	8.08	0.008	0.27
	VP_5 ($T_7; C > A, B, D$)	31.7%	73.7%	17.06	< 0.001	0.42
	VP_6 ($T_9; B > A, C, D$)	17.1%	61.4%	19.14	< 0.001	0.44
	VP_7 ($T_{13}; C > A, B$)	63.4%	75.4%	1.65	0.20	0.13
	VP_8 ($T_{14}; A > B, C, D$)	51.2%	74.4%	6.17	0.01	0.25
	VP_9 ($T_{15}; D > A, B, C$)	43.9%	71.9%	7.82	0.005	0.28
	VP_10 ($T_{16}; D > A, B, C$)	53.7%	78.9%	7.05	0.008	0.27
Maximality	MAX_1 ($T_2; C = 4$)	39.0%	61.4%	4.78	0.03	0.22
	MAX_2 ($T_4; C = 4, D = 4$)	14.6%	29.8%	3.06	0.08	0.18
	MAX_3 ($T_5; D = 4$)	53.7%	68.4%	2.21	0.14	0.15
	MAX_4 ($T_6; C = 4, D = 4$)	12.2%	21.1%	1.31	0.25	0.12
	MAX_5 ($T_7; C = 4$)	41.5%	64.9%	5.30	0.02	0.23
	MAX_6 ($T_9; B = 4$)	22.0%	57.9%	12.58	< 0.001	0.36
	MAX_7 ($T_{13}; C = 4$)	41.5%	52.6%	1.19	0.27	0.11
	MAX_8 ($T_{14}; A = 4$)	56.1%	71.9%	2.64	0.10	0.16
	MAX_9 ($T_{15}; D = 4$)	41.5%	71.9%	9.16	0.002	0.31
	MAX_10 ($T_{16}; D = 4$)	65.9%	71.9%	0.41	0.52	0.07

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. $T_3 - T_{10}$ 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.

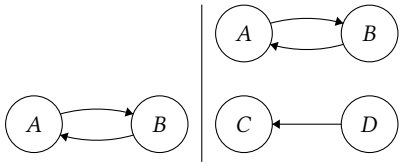

Figure 5: Two graphs shown in two different tasks. Independence states that the ranking between A and B should be the same in the two graphs.

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.

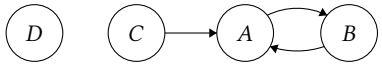

Figure 6: Void precedence states that C and D should be ranked higher then A and B.

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 $\alpha = 0.73$ 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 $\alpha = 0.87$ for faith in intuition subscale was quite high, but it was much lower for need for cognition subscale (Cronbach's $\alpha = 0.46$) 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 $r_\varphi = 0.27$.

Results of scale-level analyses, presented in Table 3, further demonstrate that graphical representation of argument structure significantly enhances participants' performance. The effect of

graphical representation was consistently significant across scales (values of t statistics ranged from 4.04 to 6.41, all p-values were equal or less than 0.001) and its size ranged from 0.70 (medium effect) to 1.37 (very large) in terms of Cohen's d statistic, i.e., from 0.33 to 0.55 in terms of point-biserial correlation coefficient.

Results of the correlational analysis are shown within Table 4 (no-graph group) and Table 5 (graph group) with Cronbach's alpha coefficients on diagonal lines. Presented results indicate that internal consistency of responses was somewhat higher in the group that was presented with graphs, and this difference was statistically significant for anonymity between tasks ($\chi^2(1) = 8.11$, $p = 0.004$) and void precedence ($\chi^2(1) = 11.38$, $p < 0.001$) scales.

Measures of individual differences in participants' propensity to comply with different normative principles showed positive manifold, but this tendency was far more pronounced in the graph group, in which correlation coefficients ranged from 0.38 to 0.67. Consequently, a single latent factor ($\lambda = 3.17$), that loaded highly on each scale (all rs above 0.70) and accounted for 63.4% of the total variance, was extracted. In other words, participants who were more prone to comply with one of principles were also more prone to comply with all the other principles, which indicates that these behaviors are rooted in more general ability to comply with argumentation principles. This ability was significantly related to cognitive reflection ($r = 0.34$, $p < 0.001$), but not to faith in intuition and need for cognition ($rs < 0.20$). This indicates that propensity to resist reporting the response that first comes to mind and to engage in further reflection is highly relevant for understanding the general ability to comply with argumentation principles. On the other side, two latent factors were retained in the no-graph group and rotated using the Promax procedure. The first latent factor accounted for 46.3% variance and it loaded highly on maximality (0.88), void precedence (0.78) and independence (0.77), while the second explained an additional 30.3% of the variance and loaded on two anonymity scales - between (0.94) and within tasks (0.87). These two factors were practically independent ($r = 0.19$, $p = 0.24$) and neither one correlated significantly with cognitive reflection, faith in intuition and need for cognition ($ps > 0.05$).

5 DISCUSSION

We showed that participants exhibit a higher level of compliance with the principles when they are provided the graphical representation of the arguments (see Table 3). We now provide some more detailed comments. We start by considering the principles that were the least complied with and analyse possible reasons.

Anonymity between tasks is the least satisfied principle for the no-graph group. Furthermore, there is a huge difference in satisfaction between the graph and the no-graph group. We hypothesise that this is because participants had difficulties in observing that there is the same structure of conflicts without formalizing the structure via an argumentation graph or another formalism. This would explain both the low degree of compliance in the no-graph group and the difference between the graph and the no-graph group.

The second least satisfied principle in the no-graph group - and the least satisfied principle in the graph group - is maximality. We hypothesise that maximality is just too strong a principle. Let us illustrate this on the task T_6, in which there was the least degree of compliance with maximality for both groups. This task is shown in Example 3.7 and Figure 6. Consider argument D; even if it is not attacked in this task, the participants do not necessarily completely agree with it due to their beliefs and their background knowledge. They can consider the strength of D to be less than 4 (for example, 3). However, maximality states that since D is not attacked, it should have the maximal strength, which in our experiment was 4.

The data from other empirical studies on argumentation [10, 22, 23] provides possible reasons of non-compliance with some of the principles. For instance, Polberg and Hunter [22] observe that "the data shows that people use their own personal knowledge in order to make judgments". It is obvious that this can yield non-compliance with anonymity and maximality. Rosenfeld et al. [23] showed that in the context of argumentative conversations, people do not always choose arguments that are justified according to the given semantic choice. Cerutti et al. [10] 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 Cerutti 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 4 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 far more consistent in their evaluation of arguments (i.e., less prone to respond randomly) when arguments were accompanied by graphical representations.

Conclusions presented here should be taken with some caution considering the exploratory nature of our study and scarcity of similar previous research. One might ask if results would be replicated under other conditions. For example, participants could be instructed to rank arguments in each task (instead of rating each of

Scale	No-graph (n = 57)		Graph (n = 41)		Test of difference			
	M	SD	M	SD	$t(96)$	p	r	d
Independence	73.1	21.7	88.8	16.5	4.04	< 0.001	0.38	0.81
Anonymity between tasks	24.9	20.1	63.5	34.3	6.41	< 0.001	0.55	1.37
Anonymity within tasks	71.5	18.6	91.4	14.5	5.93	< 0.001	0.52	1.19
Void precedence	47.3	24.7	73.0	32.0	4.29	< 0.001	0.40	0.88
Maximality	38.8	25.0	57.2	27.2	4.42	0.001	0.33	0.70

Table 3: Descriptive statistics for normative principles scales in two groups

	1	2	3	4	5	6	7	8	9
1. Independence	0.46								
2. Anonymity between tasks	0.14	0.29							
3. Anonymity within tasks	0.40^{**}	0.68^{**}	0.76						
4. Void precedence	0.53^{**}	0.23	0.29	0.68					
5. Maximality	0.46^{**}	−0.16	0.06	0.51^{**}	0.73				
6. Control tasks	0.43^{**}	$0.36*$	0.49^{**}	0.47^{**}	−0.08	0.37			
7. Cognitive reflection	0.23	0.07	0.19	0.21	0.05	0.20	0.74		
8. Need for Cognition	−0.17	−0.17	−0.06	0.02	0.02	−0.04	−0.11	0.32	
9. Faith in intuition	−0.10	−0.04	−0.07	−0.23	−0.21	0.04	0.05	−0.03	0.84

Table 4: Correlations in the no-graph group. Cronbach's αs are presented on a diagonal line. $^{*}p < 0.05$, $^{**}p < 0.01$.

	1	2	3	4	5	6	7	8	9
1. Independence	0.40								
2. Anonymity between tasks	0.46^{**}	0.76							
3. Anonymity within tasks	0.49^{**}	0.67^{**}	0.81						
4. Void precedence	0.54^{**}	0.67^{**}	0.64^{**}	0.90					
5. Maximality	0.38^{**}	0.43^{**}	0.44^{**}	0.65^{**}	0.78				
6. Control tasks	0.26	0.32^{*}	0.19	0.47^{**}	0.32^{*}	0.47			
7. Cognitive reflection	0.28^{*}	0.24	0.27^{*}	0.37^{**}	0.18	0.20	0.82		
8. Need for Cognition	0.10	0.02	0.10	0.14	0.19	0.27^{*}	0.05	0.50	
9. Faith in intuition	0.32^{*}	0.04	0.11	0.13	0.13	0.11	0.02	0.34^{*}	0.89

Table 5: Correlations in the graph group. Cronbach's αs are presented on a diagonal line. $^{*}p < 0.05$, $^{**}p < 0.01$.

them). They could also be presented with all possible combinations of pairs of arguments in each task and instructed to decide if one of the two arguments is stronger and, if yes, which one. Robustness of results could also be tested by varying instruments and material. Future research might also include additional tasks that would allow for research of other principles, or which would examine reasoning on more complex structures of arguments [16, 19, 24, 25]. It would also be interesting to examine if solely graphical representation of arguments structure (without the text) would lead to different results. Finally, the different populations, beside university students, should be also included in further studies in order to extend the generalisability of the obtained results.

Given the empirically proven usefulness of the graphical representation, one can ask the question of how to use this fact in practice. Suppose that the people are discussing a question online in the framework of e-democracy (e.g. whether to build a swimming pool, a park, or a railway station). We know that they will have a more insightful discussion and a better mutual understanding if they have access to the graph corresponding to their discussion. 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

[1] Leila Amgoud and Jonathan Ben-Naim. 2013. Ranking-Based Semantics for Argumentation Frameworks.. In *Scalable Uncertainty Management - 7th International Conference, SUM 2013, Washington, DC, USA, September 16-18, 2013. Proceedings*. 134–147. https://doi.org/10.1007/978-3-642-40381-1_11

[2] Leila Amgoud and Jonathan Ben-Naim. 2018. Evaluation of arguments in weighted bipolar graphs. *Int. J. Approx. Reason.* 99 (2018), 39–55. https://doi.org/10.1016/j.ijar.2018.05.004

[3] Leila Amgoud, Jonathan Ben-Naim, Dragan Doder, and Srdjan Vesic. 2017. Acceptability Semantics for Weighted Argumentation Frameworks.. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 56–62. https://doi.org/10.24963/ijcai.2017/9

[4] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. 2011. An introduction to argumentation semantics. *Knowledge Eng. Review* 26, 4 (2011), 365–410. https://doi.org/10.1017/S0269888911000166

[5] Pietro Baroni, Dov M. Gabbay, Massimiliano Giacomin, and Leendert van der Torre (Eds.). 2018. *Handbook of formal argumentation*. College Publications, Erscheinungsort nicht ermittelbar.

[6] Philippe Besnard and Anthony Hunter. 2008. *Elements of Argumentation*. MIT Press. https://mitpress.mit.edu/books/elements-argumentation

[7] Elise Bonzon, Jérôme Delobelle, Sébastien Konieczny, and Nicolas Maudet. 2016. A Comparative Study of Ranking-Based Semantics for Abstract Argumentation.. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. 914–920. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12465

[8] Martin W. A. Caminada, Walter Alexandre Carnielli, and Paul E. Dunne. 2012. Semi-stable semantics. *J. Log. Comput.* 22, 5 (2012), 1207–1254. https://doi.org/10.1093/logcom/exr033

[9] Federico Cerutti, Marcos Cramer, Mathieu Guillaume, Emmanuel Hadoux, Anthony Hunter, and Sylwia Polberg. 2021. Empirical Cognitive Studies About Formal Argumentation. In *Handbook of Formal Argumentation*, Guillermo R. Simari Dov Gabbay, Massimiliano Giacomin and Matthias Thimm (Eds.). Vol. 2. College Publications.

[10] Federico Cerutti, Nava Tintarev, and Nir Oren. 2014. Formal Arguments, Preferences, and Natural Language Interfaces to Humans: an Empirical Evaluation. In *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014) (Frontiers in Artificial Intelligence and Applications, Vol. 263)*, Torsten Schaub, Gerhard Friedrich, and Barry O'Sullivan (Eds.). IOS Press, 207–212. https://doi.org/10.3233/978-1-61499-419-0-207

[11] Marcos Cramer and Mathieu Guillaume. 2018. Directionality of Attacks in Natural Language Argumentation.. In *Proceedings of the fourth Workshop on Bridging the Gap between Human and Automated Reasoningco-located with the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018), Stockholm, Schweden, July 14, 2018*. 40–46. http://ceur-ws.org/Vol-2261/paper7.pdf

[12] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artif. Intell.* 77, 2 (1995), 321–358. https://doi.org/10.1016/0004-3702(94)00041-X

[13] F. H. van Eemeren and R Grootendorst. 2004. *A systematic theory of argumentation the pragma-dialectical approach*. Cambridge University Press, Cambridge. OCLC: 1198276136.

[14] Shane Frederick. 2005. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19, 4 (Nov. 2005), 25–42. https://doi.org/10.1257/089533005775196732

[15] Sarah Alice Gaggl and Stefan Woltran. 2013. The cf2 argumentation semantics revisited. *J. Log. Comput.* 23, 5 (2013), 925–949. https://doi.org/10.1093/logcom/exs011

[16] Abdelraouf Hecham, Madalina Croitoru, and Pierre Bisquert. 2017. Argumentation-Based Defeasible Reasoning For Existential Rules.. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*. 1568–1569. http://dl.acm.org/citation.cfm?id=3091364

[17] Luca Iandoli, Ivana Quinto, Anna De Liddo, and Simon Buckingham Shum. 2014. Socially augmented argumentation tools: Rationale, design and evaluation of a debate dashboard. *International Journal of Human-Computer Studies* 72, 3 (2014), 298–319. https://doi.org/10.1016/j.ijhcs.2013.08.006

[18] Takayuki Ito and Toramatsu Shintani. 1997. Persuasion among Agents: An Approach to Implementing a Group Decision System Based on Multi-Agent Negotiation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97, Nagoya, Japan, August 23-29, 1997, 2 Volumes*. Morgan Kaufmann, 592–599.

[19] Sanjay Modgil and Henry Prakken. 2014. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation* 5, 1 (2014), 31–62. https://doi.org/10.1080/19462166.2013.869766

[20] Rosemary Pacini and Seymour Epstein. 1999. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology* 76, 6 (1999), 972–987. https://doi.org/10.1037/0022-3514.76.6.972

[21] Alison R. Panisson. 2017. Argumentation Schemes and Enthymemes in Multi-agent Systems. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*, Kate Larson, Michael Winikoff, Sanmay Das, and Edmund H. Durfee (Eds.). ACM, 1849–1850. http://dl.acm.org/citation.cfm?id=3091467

[22] Sylwia Polberg and Anthony Hunter. 2018. Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. *Int. J. Approx. Reason.* 93 (2018), 487–543. https://doi.org/10.1016/j.ijar.2017.11.009

[23] Ariel Rosenfeld and Sarit Kraus. 2016. Providing Arguments in Discussions on the Basis of the Prediction of Human Argumentative Behavior. *ACM Trans. Interact. Intell. Syst.* 6, 4 (2016), 30:1–30:33. https://doi.org/10.1145/2983925

[24] Francesca Toni. 2014. A tutorial on assumption-based argumentation. *Argument & Computation* 5, 1 (2014), 89–117. https://doi.org/10.1080/19462166.2013.869878

[25] Bruno Yun. 2019. *Argumentation techniques for existential rules. (Techniques d'argumentation pour les règles existentielles)*. Ph.D. Dissertation. University of Montpellier, France. https://tel.archives-ouvertes.fr/tel-02197405

[26] Bruno Yun, Srdjan Vesic, and Madalina Croitoru. 2020. Ranking-Based Semantics for Sets of Attacking Arguments.. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. 3033–3040. https://aaai.org/ojs/index.php/AAAI/article/view/5697

[27] Bruno Yun, Srdjan Vesic, and Madalina Croitoru. 2020. Sets of Attacking Arguments for Inconsistent Datalog Knowledge Bases.

[28] Bruno Yun, Srdjan Vesic, and Nir Oren. 2020. Representing Pure Nash Equilibria in Argumentation. *CoRR* abs/2006.11020 (2020). arXiv:2006.11020 https://arxiv.org/abs/2006.11020