Arguably argumentative: A formal approach to the argumentative theory of reason

January 12, 2014

Abstract

1 Introduction

The idea that thought entails existence, or at least presupposes it, is a conceptual pillar of analytic philosophy. It is hard to disagree that is has a certain appeal: "I think, so I am, so I might as well go on thinking". For the sceptic, however, it also begs the following question: what is the "I" in such a line of thought? The armchair philosopher might be too busy with his thoughts to worry about it, but in the field of social psychology, particularly in the tradition going back to the work of George Herbert Mead and the Chicago school, it has well established pride of place (Mead 1967).

The best known theories developed in this field all agree that any "I" is essentially a social construction; you exist as a thinker only because you interact with other thinkers. In particular, social interaction is partly constitutive of self, not merely emergent from it. Consequently, reason itself is emergent from social contact, and the view that individual rationality is the basis for rational interaction must be rejected. Rather, interaction and reasoning should be seen as mutually dependent notions, on the basis of which more subtle notions of rationality can be explored.

This idea is both convincing and powerful, and it is becoming increasingly important to many different fields of research, including economy, law, biology and artificial intelligence (Blume and Durlauf 2001; Dworkin 1986; Waal and Ferrari 2010; Benthem 2011; Ossowski 2013). In all these research areas, there is a trend towards viewing rationality as fundamentally embedded in a social context. Importantly, this context is seen as important not only because people are social and tend to interact, but also because *who* they are, *what* they want, and *why* they want it, tends to depend on how they engage with each other and their environment. Hence the individual – the *agent* in the context of formal models – is himself in need of more subtle analysis, in terms of the same structures that are used to describe important aspects of the economic, legal, environmental and computational world that contains him.

To accommodate this point of view across different domains, we are in need of better theoretical foundations, allowing us to investigate the relationship between reasoning and interaction, taking into account that they are co-dependent and co-evolving. In paper, we address this challenge, and we do so using formal logic, drawing on tools and techniques developed in the context of multi-agent systems. The connection between various branches of social science and formal logic and computer science has received much attention in recent years, and it has led to a surge of interesting interdisciplinary research (Wooldridge 2009; Benthem 2008; Verbrugge 2009; Ditmarsch, van der Hoek, and Kooi 2007; Parikh 2001).

While much recent work in applied logic has been devoted to modelling agency and interaction, the standard starting point is still that agents reason in strict adherence to some common standard of correctness, specified by some given formal logic. Also, it is typically assumed that rational interaction emerges from the fact that agents are individually rational in some appropriate sense, for instance because they seek to maximize some given utility function. In this paper, we argue that in order to provide adequate formal foundations for rational interaction we need to depart from such reductionist assumptions. We point to the argumentative theory of reason, introduced in (Mercier and Sperber 2011), as an alternative approach, and we sketch a *formal* representation of basic elements of this theory. We show, in particular, how existing tools and techniques in contemporary logic allows us to formulate systems of dynamic logic for multi-agent argumentation which can be used to encode and explore key theoretical aspects, as well as facilitate modelling of concrete systems.

The structure of the paper is as follows. In Section 2 we present the necessary background on the argumentative theory of reason and the distinct notion of argumentation relied on in the theory of argumentation frameworks, as studied in artificial intelligence. We discuss the differences between these two notions of argumentation, and we argue that in order to use argumentation frameworks to arrive at logics for representing the argumentative theory, we need to conceptualize argumentation frameworks as subjective representations of semantic content, on the basis of which deliberation can take place. We argue that a fundamental question raised by the argumentative theory, which can then be analysed by formal logical tools, is the question of how argumentative deliberation works, and how it can sometimes create a common representation, a *consensus* among participants. We propose, in particular, that the argumentative theory implicitly relies on, and suggests further study of, *social* rationality constraint – imposed at the deliberative level, and formulated with respect to the outcome of deliberation. Moreover, we argue that these constraints are *not* reducible to corresponding notions of rationality that applies to individual reasoners, who are instead characterized by a distinct form of *argumentative rationality*, in that they seek primarily to maximize their influence, to win as many arguments as possible.

In Section 3 we introduce *deliberative Kripke frames*, a versatile formal semantics based on modal logic which gives us access to an abstract view of argumentative deliberation, well suited for further exploration of core theoreti-

cal aspects. We provide some examples of semantic modeling facilitated by this formalism, and we go on to present a simple modal language to reason about argumentative social processes. We then motivate what we believe to be the main challenge for future work: how to characterize interesting notions of social rationality using theories in modal logic. We present a few preliminary suggestions in this regard, but argue that more work is needed to explore different theories, in languages of different complexity and expressive power.

In Section 4 we discuss the limitations of our own approach, and suggest directions for future work, and in Section 5 we conclude.

2 Argumentative agents: Towards a semantics for individual reasoning based on argumentation

The argumentative theory of reason is formulated on the basis of a vast amount of experimental evidence, and its core idea is that the notion of argumentation can serve a foundational role in cognitive science. The theory holds that human reasoning evolved to facilitate efficient argumentation, and that the function of reason is not in arriving at logically correct forms of inference but to contribute to social interaction in such a way as to maximize the positive effect of deliberation. This, for instance, can serve to shed light on why humans so often reason in a way that most theories would judge to be fallacious. According to the argumentative theory, they sometimes do so because fallacies can be useful in the context of argumentation.

It is important to note that the theory involves a notion of argumentation which is conceptually distinct from that found in traditional argumentation theory, going back to (Toulmin 2003) (first edition from 1958). In this research tradition, the theories developed to account for argumentation tend to be highly normative, focusing on recognizing and categorizing fallacies and on designing argumentation schemes and models that are meant to facilitate sound and rational reasoning, particularly regarding what arguments we should accept in a given scenario. The argumentative theory, on the other hand, asks us to look at human reasoning only as an element of more complex social processes that may or may not have outcomes that we judge desirable. In particular, to define more interesting normative forms of rationality, the argumentative theory suggests specifying them with respect to deliberative processes, not with respect to the reasoning processes taking place inside individuals.

Individual reasoning, on the other hand, is understood descriptively, in terms of how people reason, but in such a way that the theory explicitly tackles the normatively laden question of *why* people reason the way they do. According to the argumentative theory, reason developed in order to facilitate successful argumentation, and the evidence we have about the nature of human reasoning supports the hypothesis that people reason in order to maximize their chances of exerting influence in the context of argumentative deliberation; They reason

in order to win arguments.

This also explains why people often make "mistakes" when they reason, and why they often make decisions that are not optimal, or even rational, in terms of a classical normative understanding. But as the argumentative theory explains, the outcome of deliberation can still resemble what traditional accounts of rationality deems to be desirable outcomes of individual reasoning. Hence it might be that established normative ideas about reason still have a role to play. They should be formulated differently, however, with respect to social processes. This latter point is not explicitly made in [**?**], but the presentation given there is highly suggestive of it, as most of the examples and arguments used in favor of an argumentative view of reason relies on showing how the "quality" of individual reasoning – understood in a classical sense – improves when the social conditions are favorable. We believe one of the most interesting questions raised by the argumentative theory is how to be more precise about the way in which constraints imposed at a social level can replace or at least support notions of individual rationality as a basis for exploring intelligent interaction.

In the following we will attempt to shed light on it by the use of multi-agent logic, and the first step is to identify the appropriate notion of agency. In particular, we need to provide a formalization of the *argumentative agent*, the agent who reasons in order to win arguments. To do this, we will make use of argumentation frameworks, introduced in (Dung 1995). These are simple mathematical objects, essentially directed graphs, which facilitate the investigation of a whole range of interesting semantics (Baroni and Giacomin 2007).

The theory of argumentation frameworks has been influential in the context of artificial intelligence (Rahwan 2009). It is capable of capturing many different semantic notions, including semantics for multi-valued and non-monotonic logics, logic programs and games (Dung 1995; Dyrkolbotn and Walicki 2013). More recently, the work in (Brewka, Dunne, and Woltran 2011) shows how argumentation frameworks can be used to provide a faithful (and computationally efficient) representation also of semantics that are formulated with respect to the much more fine-grained formalism of abstract dialectical frameworks (Brewka and Woltran 2010). For our project, it is also important to note that much recent work focuses on providing logical foundations the theory (Grossi 2010a; Grossi 2010b; M. W. A. Caminada and Gabbay 2009; Arieli and Caminada 2013; Dyrkolbotn and Walicki 2013; Dyrkolbotn 2013).

In our opinion, this makes argumentation frameworks highly suited as a technical starting point towards logics for argumentative deliberation. However, we propose to make use of them in a novel way, not to model actual argumentation scenarios, but to model agents' interpretations of semantic meaning – in argumentative terms – of the propositions that are up for debate. In terms of each individual agent, using terminology from cognitive science, it places the argumentation framework at the informational level of cognitive processing, where previous work have already shown that logical tools can have a particularly crucial role to play, also serving to shed new light on established truths arrived at through empirical work, see e.g., (Stenning and van Lambalgen 2005).

While much work on multi-agent argumentation has already been carried out

in a formal and semi-formal context, we note that this work is mostly based on a traditional view of argumentation theory. For instance, we think this is implicit in recent formal work such as that of (M. Caminada, Pigozzi, and Podlaszewski 2011; M. Caminada and Pigozzi 2011) and even more so in the survey of the field given in (Rahwan 2009). In our opinion, however, this view is inappropriate when attempting to formalize the argumentative theory. The problem is that the representation of the argumentation scenario is fixed and not open to dispute and dynamic change, except with respect to the question of how it should be evaluated. But to model the argumentative theory, we need to depart from this starting point, since it is crucial that the basic representation of the surrounding semantic reality is itself a subjective construction, distinctly produced in each individual agent. This is why we use argumentation frameworks as models of the agents' internal view of the relevant arguments and how they are related.

In the following subsection we develop this idea in formal detail, starting with the necessary preliminaries regarding argumentation frameworks.

2.1 Argumentation frameworks, agents and semantic views

Given a set of atoms Π – acting as names of arguments – an argumentation framework (AF) over Π is a relation $E \subseteq \Pi \times \Pi$. Intuitively, an element $(x, y) \in E$ encodes the fact that arguments x attacks argument y and we can depict E as a directed graph, giving a nice visualization of how the atoms in Π are related as arguments, see Figure **??** for an example. We introduce the notation $E^+(x) = \{y \in \Pi \mid (x, y) \in E\}, E^-(x) = \{y \in \Pi \mid (y, x) \in E\}$ and $=^n (\Pi) \backslash E^+(x) \cup E^-(x)$.

Given an AF E, the purpose of an argumentation semantics is to identify, using the structure of E, the collection of sets of arguments that can be accepted if taken together, typically called *extensions*, see e.g., [**?**] For instance, if $E = \{(p, q), (r, p)\}$, then the semantics might prescribe $\{r, q\}$ as a set that can be accepted, since r defends q against the argument made by p and r is not in turn attacked. There are many different argumentation semantics, each catering to a different set of intuitions about what should be required for a given set of arguments to count as acceptable.

Given an AF E, it is natural to represent an extension $A \subseteq \Pi$ as a three-valued assignment $\mathsf{c_A} : \mathsf{\Pi} \to \{\mathsf{1}, \mathsf{0}, \frac{\mathsf{1}}{\mathsf{2}}\}$ such that

$$\mathsf{c_A(x)} = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \in E^+(A) \\ \frac{1}{2} \text{ otherwise} \end{cases}$$

Then a possible intuitive reading is that arguments in A are regarded as true propositions, arguments attacked by one of these are regarded as false proposition, while all other arguments are taken to correspond to propositions which have an undecided semantic status. The three-valued representation of extensions also lead to an alternative view on argumentation semantics, due to [**?**], which takes a semantics to be a collection of three-valued assignments. In this way, a semantics for argumentation can be reasoned about using three-valued

Admissible:	$a(E) = \{\mathsf{c} \in \mathbf{C}(\mathsf{E}) \mid \mathsf{E}^{-}(\mathsf{c}^{1}) \subseteq \mathsf{c}^{0}\}$
Complete:	$c(E) = \{\mathsf{c} \in \mathbf{C}(\mathsf{E}) \mid \mathsf{c}^{1} = \{\mathsf{x} \in \Pi \mid \mathsf{E}^{-}(\mathsf{x}) \subseteq \mathsf{c}^{0}\}\}$
Grounded:	$g(E) = \{\bigcap c(E)\}$
Preferred:	$p(E) = \{\mathsf{c}_1 \in \mathsf{a}(\mathsf{E}) \mid \forall \mathsf{c}_2 \in \mathsf{a}(\mathsf{E}) : \mathsf{c}_1^{1} \not\subset \mathsf{c}_2^{1}\}$
Semi-stable:	$ss(E) = \{\mathsf{c}_1 \in \mathsf{a}(\mathsf{E}) \mid \forall \mathsf{c}_2 \in \mathsf{a}(\mathsf{E}) : \mathsf{c}_1^{\frac{1}{2}} \not\supset \mathsf{c}_2^{\frac{1}{2}}\}$
Stable:	$s(E) = \{\mathsf{c} \in \mathsf{a}(\mathsf{E}) \mid \mathsf{c}^{\frac{1}{2}} = \emptyset\}$

Figure 1: Various semantics, defined for any $E \subseteq \Pi \times \Pi$

logic, an idea that has been explored in some recent work [**?**, **?**, **?**]. This will be exploited in the coming sections, as we will rely on three-valued Łukasiewicz logic when we reason statically, i.e., either about a given agents' subjective view or the current deliberative state.

In Figure 1 we provide definitions of the most commonly known semantics based on argumentation frameworks. The logic introduced in the next section is parameterized by an argumentation semantics and the choice of such a semantics will not crucial be for our analysis in this paper. We note, however, that the *admissible* semantics encode what seems to be minimal criteria of acceptability of arguments. Intuitively, it requires that an acceptable set must be free from internal conflict and that it must also be able to defend itself against all attacks. The other semantics in Figure 1 are all based on the same idea, but adds other requirements that are less obviously reasonable. In the following we will assume only that whatever our argumentation semantics returns as an acceptable set, it is always also acceptable under the admissible semantics.

Towards the definition of an argumentative agent, let $\mathcal{A}$ be a set of agent names. Then a *view* for agent $a \in \mathcal{A}$ is an AF $V_A \subseteq \Pi \times \Pi$. It encodes his interpretation of the semantic relationship between the arguments under consideration, specified in keeping with the idea that his reasoning is based on an *argumentative* understanding of meaning. Then an *argumentative state* is a tuple $(V_a)_{a \in \mathcal{A}}$, associating a view with each agent. In this paper, we will assume for simplicity that the argumentative state remains the same throughout the course of deliberation, so that the views of the agents are not themselves subject to revision as the debate unfolds. This, however, can easily be extended by application of the dynamic framework developed in the next section.

To reason about AFs we will use a simple propositional language $\mathcal{L}$, with negation and implication, as defined by the grammar below.

$$\alpha := p \mid \neg\alpha \mid \alpha \to \alpha$$

where $p \in \Pi$. Then we can define static argumentative truth following Łukasiewicz three-valued logic, by defining extensions of $\mathsf{c} : \Pi \to \{1, 0, \frac{1}{2}\}$ inductively as follows

$$\begin{array}{l} \overline{\mathsf{c}}(p) = \mathsf{c}(\mathsf{p}) \text{ for } \mathsf{p} \in \Phi \\ \overline{\mathsf{c}}(\neg\alpha) = 1 - \overline{\mathsf{c}}(\alpha) \\ \overline{\mathsf{c}}(\alpha \to \beta) = min\{1, 1 - (\overline{\mathsf{c}}(\alpha) - \overline{\mathsf{c}}(\beta))\} \end{array} \quad (1)$$

Then, given an agent $a \in \mathcal{A}$ with a view V_A, we can add the modality $\blacklozenge_a$ to perform (boolean) meta-reasoning about the acceptance status of arguments on AFs, under some arbitrary semantics ε. In particular, we get the following multi-agent language $\mathcal{L}^{\blacklozenge}$

$$\phi := \blacklozenge_a \alpha \mid \neg\phi \mid \phi \wedge \phi$$

where $\alpha \in \mathcal{L}$ and $a \in \mathcal{A}$. Given an argumentative state $\mathcal{B} = (V_a)_{a\in\mathcal{A}}$, we define truth for formulas from $\mathcal{L}^{\blacklozenge}$ inductively as follows, for all formulas ϕ

$$\begin{array}{l} \mathcal{B} \models_\varepsilon \blacklozenge_a \alpha \text{ if there is } \mathsf{c} \in \varepsilon(\mathsf{V_a}) \text{ s.t. } \overline{\mathsf{c}}(\alpha) = 1 \\ \mathcal{B} \models_\varepsilon \neg\phi \text{ if not } \mathcal{B} \models_\varepsilon \phi \\ \mathcal{B} \models_\varepsilon \phi \wedge \psi \text{ if } \mathcal{B} \models_\varepsilon \phi \text{ and } \mathcal{B} \models_\varepsilon \psi \end{array} \tag{2}$$

The crucial challenge that remains, and which will be addressed in the next section, is to find a mechanism for formally introducing an appropriate kind of multi-agent interaction and dynamics, suitable for representing the argumentative theory. This is the question we address in the following section.

3 Argumentative deliberation: Towards formalization of rational interaction using dynamic argumentation logic

Given a basis which encodes agents' view of the arguments, we are interested in the possible ways in which agents can deliberate to reach *agreement* on how arguments are related. That is, we are interested in the set of all AFs that can plausibly be seen as resulting from a *consensus* regarding the status of the arguments in Π. What restrictions is it reasonable to place on a consensus? It seems that while many restrictions might arise from pragmatic considerations, and be implemented by specific protocols for "good" deliberation in specific contexts, there are few restrictions that can be regarded as completely general. For instance, while there is often good reason to think that the position held by the majority will be part of a consensus, it is hardly possible to stipulate an axiomatic restriction on the notion of consensus amounting to the principle of majority rule. Indeed, sometimes deliberation takes place and leads to a single dissenting voice convincing all the others, and often, these deliberative processes are far more interesting than those that transpire along more conventional lines. However, it seems reasonable to assume that whenever *all* agents agree on how an argument p is related to an argument q, then this relationship is part of any consensus. This, indeed, is the only restriction we will place on the notion of a consensus; that when the AF F is a consensus for *basis*, it must satisfy the following *faithfulness* requirement.

- *For all $p, q \in \Pi$, if there is no disagreement about p's relationship to q (attack/not attack), then this relationship is part of* F

This leads to the following definition of the set $\Upsilon(\mathcal{B})$, which we will call the set of *complete assents* for $\mathcal{B}$, collecting all AFs that are faithful to $\mathcal{B}$.

$$\Upsilon(\mathcal{B}) = \left\{ \mathsf{F} \subseteq \Pi \times \Pi \;\middle|\; \bigcap_{a \in \mathcal{A}} V_a \subseteq \mathsf{F} \bigcup_{a \in \mathcal{A}} V_a \right\} \tag{3}$$

An element of $\Upsilon(\mathcal{B})$ represents a possible consensus among agents in $\mathcal{A}$, but it is an *idealization* of the notion of assent, since it disregards the fact that in practice, assent tends to be *partial*, since it results from a dynamic process, emerging through *deliberation*. Indeed, as long as the number of arguments is not bounded we can *never* hope to arrive at complete assent via deliberation. We can, however, initiate a process by which we reach agreement on more and more arguments, in the hope that this will approximate some complete assent, or maybe even be *robust*, in the sense that there is *no* deliberative future where the results of current partial agreement end up being undermined. Complete assent, however, arises only in the limit.

In practice, we can only every analyze this limit by looking at smaller parts of the whole, some partial consensus that has been obtained through deliberation. Hence we define a *deliberative state* as a tuple $q = (q_S, q_E)$ such that $q_S \subseteq \Pi$ and

$$\bigcap_{a \in \mathcal{A}} V_a|_{q_S} \subseteq q_E \subseteq \bigcup_{a \in \mathcal{A}} V_a|_{q_S} \tag{4}$$

So a deliberative state q is an AF q_E such that all attacks of q_E are between arguments of q_S. Moreover, q is faithfully generated from the views of the agents; q_E only contains semantic information that is present in at least one agents' view. Now we are ready to define deliberative Kripke models.

Definition 3.1. *Given a basis $\mathcal{B}$, a deliberative Kripke model with respect to $\mathcal{B}$ is a tuple (Q, R) where Q is a set of deliberative states for $\mathcal{B}$ and $R \subseteq Q \times Q$ is a relation on Q.*

We will reason about deliberative Kripke models using the following language.

$$\phi \quad ::= \blacklozenge\alpha \;\mid\; \blacklozenge_a\alpha \mid \neg\phi \mid \phi \wedge \phi \mid \langle p \rangle \phi \mid \Diamond\phi$$

where $p \in \Pi$, $\alpha \in \mathcal{L}$ and $a \in \mathcal{A}$.

4 Discussion and future work

5 Conclusion