

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

No Author Given

## 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 it has a certain appeal: "I think, so I am, so I might as well go on thinking". For the skeptic, 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 come to occupy a 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 and sound reasoning 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 im-

portant aspects of the economic, legal, environmental and computational worlds that contain 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 this paper, we argue that this challenge can be taken on 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).

However, while much recent work in applied logic has been devoted to modeling 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. In search of an alternative approach, we start from the argumentative theory of reason, introduced in (Mercier and Sperber 2011). We consider its implications for logical modeling, and we sketch the development of a general logical framework that enables us to capture key aspects of the theory. Our overreaching aim is to argue that a formal approach to the argumentative theory can provide interesting insights into the nature of rational interaction, and that further work in this vein can serve as a fruitful meeting point between different fields, and make a valuable contribution to the search for new, social, foundations for notions such as rationality and intelligence.

The structure of the paper is as follows. In Section 2 we present the necessary background on the argumentative theory of reason. We note that the notion of argumentation used to formulate this theory is conceptually distinct from the notion of argumentation relied on in the theory of argumentation frameworks, as studied in artificial intelligence. We discuss the differences and common elements between the two notions, and we conclude we can only use argumentation frameworks to arrive at logics for representing the argumentative theory if we conceptualize argumentation frameworks as subjective representations of semantic content, on the basis of which deliberation can take place. We go on to argue that a fundamental question raised by the argumentative theory, which can then be analyzed by formal tools, is the question of how argumentative deliberation works, and how it can sometimes result in desirable outcomes, even if the individual participants are not classically sound in their own reasoning. 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 theoretical 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, to suggest that more work is needed in this area, to map the different notions of rationality we can define in various languages of different complexity and expressive power.

In Section ?? we discuss the limitations of our own approach, and suggest directions for future work, and in Section 4 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. Such models only provide a starting point, however, and the real challenge becomes how to describe deliberative processes that can take place on the basis of the agents' diverging interpretations of the world.

We return to this question in Section ??, after we have formalized our definition of the argumentative agent, using argumentation frameworks.

## 2.1 Argumentation frameworks, agents and semantic views

Given a set of semantic atoms  $\Pi$ , which we tend to think of as names of arguments, an argumentation framework (AF) over  $\Pi$  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  $\Pi$  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  $=^n (\Pi \setminus E^+(x) \cup E^-(x))$ .

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  $c_A : \Pi \rightarrow \{1, 0, \frac{1}{2}\}$  such that

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

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 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.

$$\begin{aligned}
\text{Admissible: } a(E) &= \{c \in \mathbf{C}(E) \mid E^-(c^1) \subseteq c^0\} \\
\text{Complete: } c(E) &= \{c \in \mathbf{C}(E) \mid c^1 = \{x \in \Pi \mid E^-(x) \subseteq c^0\}\} \\
\text{Grounded: } g(E) &= \{\bigcap c(E)\} \\
\text{Preferred: } p(E) &= \{c_1 \in a(E) \mid \forall c_2 \in a(E) : c_1^1 \not\subseteq c_2^1\} \\
\text{Semi-stable: } ss(E) &= \{c_1 \in a(E) \mid \forall c_2 \in a(E) : c_1^{\frac{1}{2}} \not\subseteq c_2^{\frac{1}{2}}\} \\
\text{Stable: } s(E) &= \{c \in a(E) \mid c^{\frac{1}{2}} = \emptyset\}
\end{aligned}$$

**Fig. 1** Various semantics, defined for any  $E \subseteq \Pi \times \Pi$

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.

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

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

$$\begin{aligned}\bar{c}(p) &= c(p) \text{ for } p \in \Phi \\ \bar{c}(\neg\alpha) &= 1 - \bar{c}(\alpha) \\ \bar{c}(\alpha \rightarrow \beta) &= \min\{1, 1 - (\bar{c}(\alpha) - \bar{c}(\beta))\}\end{aligned}\tag{1}$$

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

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

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  $\mathcal{L}^\Diamond$  inductively as follows, for all formulas  $\phi$

$$\begin{aligned}\mathcal{B} \models_\varepsilon \Diamond_a \alpha &\text{ if there is } c \in \varepsilon(V_a) \text{ s.t. } \bar{c}(\alpha) = 1 \\ \mathcal{B} \models_\varepsilon \neg\phi &\text{ if not } \mathcal{B} \models_\varepsilon \phi \\ \mathcal{B} \models_\varepsilon \phi \wedge \psi &\text{ if } \mathcal{B} \models_\varepsilon \phi \text{ and } \mathcal{B} \models_\varepsilon \psi\end{aligned}\tag{2}$$

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.

## 2.2 Extended example: Rain in Bergen

We consider a simple example which motivates the need for introducing subjective views, suggesting also some shortcomings of a traditional approach to argumentation in the context of multi-agent deliberation. We assume given two agents  $a, b$  who argue about whether it will rain in Bergen today, where  $r$  represent the claim that it will rain, while  $\bar{r}$  represent the claim that it will not. Let us first assume that none of the agents actually argue in favor of their positions. However, assuming that the disagreeing parties recognize that they disagree, the positions themselves play an argumentative role, and the AF shown on the left below represents the initial state of disagreement. So far, in particular, the model appears to be an uncontroversial objective representation of the state of affairs. There is not yet any discernible need for introducing subjective views.

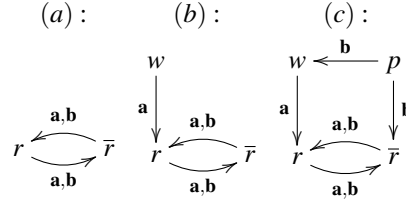


On the right above, we show a simple agent-indexed AF which illustrates a naive attempt at introducing agency to the initial AF. We label the two attacks, from  $r$  to

$\bar{r}$  and from  $\bar{r}$  to  $r$ , by both  $a$  and  $b$  to encode that they are *common* to the agents. That is, both agents acknowledge that these attacks are present – they agree that they disagree.

In this case, the rational outcome, the position that *should* be adopted, is unclear. It seems, in particular, that both  $r$  and  $\bar{r}$  are acceptable, since no further arguments have been made. This, indeed, also follows from the admissibility criterion in pure argumentation theory – both  $\{r\}$  and  $\{\bar{r}\}$  are acceptable, for all semantics defined in Figure 1. Admissibility also informs us that we cannot include *both*  $r$  and  $\neg r$ , since then we would get a set with an internal conflict; It cannot both rain and not rain, so agents  $a$  and  $b$  cannot both be right.

Let us now assume that  $a$  and  $b$  begin to argue in favor of their claims. We keep it simple and consider only two further steps of debate: first **a** introduces the argument that the weather report says that it will not rain, and then **b** counters this by announcing that he has seen a puddle on the pavement, suggesting that it has already been raining today. Let us call the arguments provided by the weather report and the puddle  $w$  and  $p$  respectively. Then, noting that  $w$  is an argument used by **a** against  $r$  and that  $p$  is an argument used by **b** as a retort against  $w$  and also, let us assume, directly against  $\bar{r}$ , a naive representation in the traditional spirit would be to view the debate as progressing from (a) to (c), as depicted below.



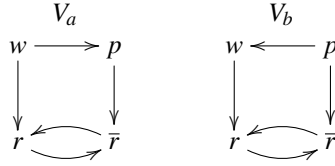
Certainly, this is a neat and uncontroversial depiction of the actual utterance made. It is tempting to ask, then: Who won the debate? Looking at the depictions and thinking of (c) above as a normal AF, forgetting about the agent labels, we conclude that  $p, r$  are the successful arguments, and that  $\bar{r}, w$  are both defeated. So **b** won the debate – the rational consensus is that it will rain in Bergen? Applying admissibility directly to the actual exchange between  $a$  and  $b$ , it would seem so. Intuitively, however, this conclusion seems entirely unwarranted. Why would a puddle be conclusive evidence, and stronger evidence than a weather report? No, it seems that a clear answer of who won the debate cannot be provided at all; for an objective bystander, a weather report might carry some weight, and so might a puddle, but neither appear particularly conclusive. This is where we recognize the shortcoming of the operational depiction given above, where we focused on the actual utterances, and the order in which they were made.

The problem, we believe, is that in this representation we failed to include information about the agents' view on each other's utterances. In a perfect world, this might not matter – all debate might eventually be settled conclusively by brute empirical fact, such as observing actual rain as opposed to consulting weather reports and puddles. However, in such a world, deliberation would certainly not be very



interesting, and it is not how it tends to play out in the real one. Rather, a debate involves crucially a search for consensus, and consensus depends crucially on how agents perceive the statements made by others, as they reflect on the totality of the debate. This is why we need to be explicit about subjective views, and always ask for a representation of how each individual agent interprets the semantic meaning of all those claims that are relevant to the scenario at hand.

In particular, what is missing in the naive model of the Bergen rain debate is some account of how agent  $a$  views puddles, and what agent  $b$  thinks of weather reports. Let us assume that their views on this are in fact the following.



It seems clear that these views are consistent with the actual exchange of arguments described earlier, and that they might easily come to result in deliberative events that appear in this form. We notice, moreover, that according to the views depicted here, both  $b$  and  $a$  acknowledge that their respective arguments for and against rain are correct, yet they also both think that their own argument is stronger, in that it attacks also the other agent's argument, but not vice versa. This might result, for instance, because the weather report gives  $a$  reason to doubt that  $b$  is telling the truth about the puddle, while seeing the puddle gives  $b$  reason to doubt the relevance of the weather report. Importantly, it might not be rational of  $a$  and  $b$  to disagree about how their arguments are related, but that they would do so is nevertheless consistent with the fact that reasoners often tend to display confirmation bias, putting more weight on evidence in support of their own beliefs.

Importantly, if we now look at the AFs encoding their views rather than the AF encoding the actual exchange of arguments, there is no longer any intuitive reason to think that the outcome of deliberation should permit us to conclude that it rains in Bergen. This, in particular, is a conclusion that relies on an interpretation that recognizes the attack  $(p, w)$  to encode correct semantic information about  $p$  and  $w$ . But his view shows that agent  $a$  disagrees with agent  $b$  in this regard, and so we reach the more appropriate conclusion that it is unwarranted to assume that  $(p, q)$  will come to influence the outcome of deliberation. In order to even begin talking about this outcome, in particular, we need first to specify an aggregated view on the semantic meaning of the arguments involved. Depending on how this interpretation is constructed, the outcome will be different. A dynamic logic based on this starting point, focusing on the development of an aggregated interpretation of meaning, is what we are aiming at. We already see how subjective views are needed to facilitate its development.

### 3 Argumentative deliberation: Towards formalization using dynamic argumentation logic

Given a basis which encodes agents' views of the arguments, we are interested in the possible ways in which agents can deliberate, and how deliberation can serve to create new, socially defined, interpretations of the arguments, interpretations that are aggregated in a non-trivial way from the views of the individual agents. We are interested, in particular, in characterizing and studying the *effect* of deliberation on semantic meaning. This, in our opinion, is the crucial question that needs to be addressed in the search for new foundations for rational interaction.

The argumentative theory, for instance, makes the assertion that unsound reasoning on the individual level, motivated by agents' desire to win arguments, can lead to sound results when embedded in a suitable deliberative context. The challenge, then, is to attempt to characterize such suitable contexts, and to define conditions of deliberation that helps promote certain normative standards of soundness that can not, and should not, be imposed at the individual level.

This is a challenge that points beyond the work already done in [?], and we believe a possible criticism against the argumentative theory, as it stands today, is that it fails to make clear how the deliberative process can take unsound reasoning and produce a sound outcome. What mechanisms are in play here, and what standards of soundness can we apply to them? Clearly, they cannot be reduced to mechanisms and standards that we should apply to reasoners individually, but need to be formulated in terms of the deliberative circumstances themselves.

This is clear from the reasoning used to support the argumentative theory, for instance when it is pointed out that groups of people tend to perform "better" in reasoning tasks when each individual is challenged by people that have views which diverge from his own. That this is so has been established in much empirical work, but how are we to make sense of it as a normative claim, and how can we express it in a theoretical and formally precise setting? This is the question we embark on now, developing a dynamic logic framework for defining and studying various notions of deliberative soundness and rationality.

We propose to think of deliberation as a process which consists in moving between possible interpretations of the meaning of arguments, where the idea is that each move is caused by some deliberative event, the exact nature and structure of which we leave unspecified the initial suggestion presented here. The intention is that each possible interpretation represents a possible aggregation of the views of the involved agents. In particular, given an argumentative state  $\mathcal{B}$ , we say that  $q \subseteq \Pi \times \Pi$  is a *deliberative state* for  $\mathcal{B}$  if

$$\bigcap_{a \in \mathcal{A}} V_a \subseteq q \subseteq \bigcup_{a \in \mathcal{A}} V_a \quad (3)$$

We collect all deliberative states for  $\mathcal{B}$  in the set  $D(\mathcal{B})$  and we use  $\Pi(q) = \{x \in \Pi \mid q^n(x) \neq \Pi\}$  to denote the set of arguments that appear in some attack from  $q$ . That is,  $\Pi(q)$  contains the arguments that are not neutral with respect to all other arguments,

according to the AF  $q$ . The constraint imposed by the definition of a deliberative state appears natural and hard to dispute. Indeed, it encodes the following principle about deliberation, which appears safe to assume in most, if not all, contexts:

If some information regarding the semantic relationship between two arguments is included in a deliberative state, the correctness of this information is endorsed by at least one agent.

In particular, we do not allow deliberation to result in interpretations that deviate from interpretations that are held unanimously by the agents. If everyone agrees on the meaning of an argument, the argument has this meaning, no matter how deliberation proceeds. As we will see, this does not mean that a possible unanimity regarding the acceptance status of an argument is necessarily reflected in the view aggregated by deliberation. For instance, even if all agents agree that  $p$  should be accepted, it is quite possible that  $p$  will not be accepted after deliberation. If the agents differ in their account of *why*  $p$  should be accepted, in particular, deliberation might lead to the rejection of  $p$ . This in itself is interesting, and it suggests that subtle questions and phenomena arise when we attempt to be more precise about our normative claims regarding deliberative rationality.

Towards formal precision, we now define the core notion of a *deliberative Kripke model*

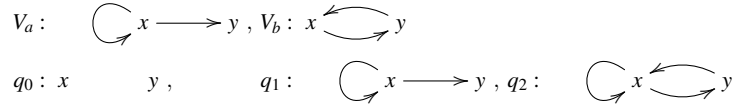
**Definition 1.** Given a deliberative state  $\mathcal{B}$ , a deliberative Kripke model for  $\mathcal{B}$  is a tuple  $(Q, R)$  such that

- $Q \subseteq D(\mathcal{B})$  is a set of deliberative states for  $\mathcal{B}$
- $R$  is a relation  $R \subseteq Q \times Q$

The idea is that the relation  $R$  encodes a process of deliberation based on the views in  $\mathcal{B}$ . If  $(q_1, q_2) \in R$  the intuition is that there is some event that can take place in the deliberative state  $q_1$  so that the aggregated views of the arguments is updated, taking us to the deliberative state  $q_2$ . In the first instance, we abstract away from events that can induce such a link, but this could be some agent presenting his point of view, or it could be some joint effort to reach a decision about some argument. The latter type of event will be encoded in Subsection ?? where we take it as the basis for defining the class of open deliberations, which we propose as a possible candidate for a rationality constraint at the social level. For now, we are content with leaving the exact content of event unspecified.

As an example of a deliberative model, consider the framework in Figure 3. Here, the argumentative state is problematic from the point of view of classical logic. In particular, we have  $\mathcal{B} \models_{\varepsilon} \neg \blacklozenge_a x \wedge \neg \blacklozenge_a \neg x$  under all  $\varepsilon$  from Figure 1, encoding that for agent  $a$ , the argument  $x$  attacks itself and is not defeated. Hence it cannot be regarded as either true or false without leading to contradiction, and agent  $a$  is prevented from reaching any classically sound conclusions about the status of either argument (since he also perceives  $x$  to attack  $y$ ). The agent  $b$ , on the other hand, has the view that  $x$  and  $y$  are in opposition to each other; If one of them is accepted the other must be rejected and vice versa, but he has no information which suggests choosing one over the other. In particular, we have  $\mathcal{B} \models_{\varepsilon} \blacklozenge_b y \wedge \blacklozenge_b \neg y$ . Hence from his point of

view, the semantic status of  $x$  and  $y$  remains unclear. Through deliberation, however, it is possible to arrive at a definite outcome which also resolves the inconsistency that  $a$  believes to be present at  $x$ . One such scenario is depicted in Figure 1, where deliberation starts with the empty framework over  $\{x, y\}$  and then proceeds by agent  $a$  first putting forth his point of view, resulting in  $q_1$ , and then continuing with agent  $b$  adding to this his own understanding, which results in the deliberative state  $q_2 = V_a \cup V_b$ . Here there is no problem, and the status of  $x$  and  $y$  has been definitely resolved, since  $y$  must be accepted and then  $x$  will be defeated, under all semantics from Figure 1, including classical logic, as encoded by the stable semantics.



**Fig. 2** A deliberative model  $(Q, R)$  over  $\mathcal{B} = (V_a, V_b)$  with  $R = \{(q_0, q_1), (q_1, q_2)\}$

This is an example of a scenario where everything runs smoothly and there is no controversy. In particular, both agents uncritically accept adding each others' points of view to the aggregated deliberative state, resulting in the union of their views emerging as the final outcome of deliberation. Things might not be so simple, however, and it is the more complicated scenarios that can benefit the most from logical modeling. It could be, for instance, that agent  $a$  has reservations about agent  $b$ 's interpretation of  $y$  as an argument that also attacks  $x$ . If we are unsure about agent  $a$ 's stance in this regard, or, more generally, unsure about whether deliberation based on the views of agents  $a$  and  $b$  will eventually return a state where the  $(y, x)$ -attack is included, we can model this by introducing branching in the deliberative model. In particular, we could introduce a reflexive loop at  $q_1$ , to indicate the possibility that  $b$ 's perspective might come to be rejected. Then we have a branching deliberative model, and while it is still *possible* to resolve the problems with original argumentative state, deliberation can then also fail to do so.

To talk about deliberative models, allowing us to distinguish and identify situations such as these, we can use existing modal languages of varying expressive power. In this paper, we will stick to simple languages to illustrate the conceptual points, and we consider first the following simple language  $\mathcal{L}_1$ , which simply adds to  $\mathcal{L}^\diamond$  a modality for talking about one-step possibilities in deliberative models.

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

where  $\alpha \in \mathcal{L}$ . The definition of satisfaction for  $\mathcal{L}_1$  on deliberative models is then defined analogously to classical modal logic.

**Definition 2.** Given an argumentation semantics  $\varepsilon$ , an argumentative state  $\mathcal{B}$  and a corresponding deliberative model  $(Q, R)$ , the truth of  $\phi \in \mathcal{L}_1$  on  $(Q, R)$  at  $q \in Q$  is defined inductively as follows

- $\mathcal{B}, (Q, R), q \models_\varepsilon \blacklozenge \alpha$  if  $\exists c \in \varepsilon(q) : \bar{c}(\alpha) = 1$

- $\mathcal{B}, (Q, R), q \models_{\varepsilon} \blacklozenge_a \alpha$  if  $\exists c \in \varepsilon(\mathcal{V}_a) : \bar{c}(\alpha) = 1$
- ...
- $\mathcal{B}, (Q, R), q \models_{\varepsilon} \blacklozenge \phi$  if there is  $q' \in Q$  s.t.  $(q, q') \in R$  and  $\mathcal{B}, (Q, R), q' \models_{\varepsilon} \phi$

We use the shorthand  $\Box\phi := \neg\blacklozenge\neg\phi$  as usual. Let us consider the model from Figure 3 as an example. Then it is easy to verify that  $\mathcal{B}, (Q, R), q_0 \models_{\varepsilon} \Box\Box\blacksquare\neg x$ , expressing how two steps of deliberation will necessarily suffice to resolve  $a$ 's semantic problems with  $x$  in this scenario, leading us to conclude  $\neg x$  at the social level. However, if we add a reflexive edge  $(q_1, q_1)$  to this model, to encode uncertainty about whether agent  $b$ 's view will survive deliberation, we obtain only the weaker  $\mathcal{B}, (Q, R), q_0 \models_{\varepsilon} \blacklozenge\blacklozenge\neg x$ . It is still *possible* that the problems at  $x$  are resolved, but this is no longer necessarily so.

This toy example illustrates that with the machinery now in place we can formally describe how deliberation can sometimes turn individual views that classical logic and traditional notions of rationality deems problematic into deliberative states that are classically consistent. This is a mechanism that is stressed as being crucial in the argumentative theory of reason, and in Subsection 3.1 we provide some more examples of how deliberative models provides a specification formalism capable of representing it, allowing us also to express and explore this mechanism using logic. However, we can also be more ambitious on behalf of the logical approach, seeking to move beyond mere modeling of concrete instances, towards formalization of theoretic concepts and principles. We provide a preliminary exploration of this line of inquiry in Subsection ??, arguing that it shows great promise, but also that it suggests development of more subtle logical tools which can allow us to introduce more structure to our semantics, and that can give us access to the expressive power of more complex modal languages.

### 3.1 Using deliberative model to represent concrete instances of argumentative interaction

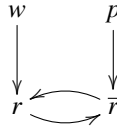
In [?], one of the primary claims concerns confirmation bias, the mechanism by which reasoners disproportionately tend to favor reasons that support previous beliefs rather than challenge them. According to the authors, this bias is not necessarily an example of flawed reasoning since it has an argumentative function that can serve to enhance the positive effects of deliberation. This claim is supported by empirical evidence, and in this section we show how scenarios where cognitive bias plays such a constructive role can be represented by deliberative models and reasoned with using modal logic. Following this, we go on to consider some more examples which we believe illustrate that as an approach to modeling, the formal framework suggested in this paper appears to be both flexible and expressive.

*Example 1 (Rain in Bergen revisited).* We return to the Bergen rain example, considered in depth in Subsection ?? as an illustration of how traditional, non-subjective,

approaches to argumentation can not do justice to the form of argumentative deliberation considered in the argumentative theory of reason.

indeed, this is also witnessed by our example; it is this "irrational" stance that serves to distinguish the views of **a** and **b**, thereby making non-trivial consensus possible without violating truthfulness.

So what do we get? Well, if disagreement runs deep and blocks consensus, then, eventually, the result will be the deliberative state that is simply the union of their two views. Then it is not hard to see that everything consistent is a possible outcome – both  $\{w, \bar{r}\}$  and  $\{r, p\}$  are admissible in the resulting AF. In this case, then, debating only served to establish the social fact that the question of whether it will rain in Bergen is still a thorny one in the social group  $\{a, b\}$ . However, if the agents are willing to consider a consensus, then they can settle on either  $r$  or  $\bar{r}$ , and, moreover, they can also choose to conclude, in agreement, that the available evidence is *insufficient* to draw any conclusion. This in particular, is the outcome resulting from the following deliberative state, which emerges from debate if the agents are prudent and reach the consensus that the intersection of their views should be adopted.



Here, both  $w$  and  $p$  are regarded as successful, meaning that *both*  $r$  and  $\bar{r}$  becomes defeated and impossible to accept.<sup>1</sup> This particular consensus outcome is particularly interesting since it seems plausible that in an actual debate, this is what one would get. More importantly still, it might be the outcome we *should* get. After all, it represent both a clear conclusion, and also a "fair" one in light of the evidence, where neither agent loses on grounds that he believes to be unreasonable. More interesting still, it is, as we mentioned, a consequence of the model that this outcome is only achievable because the agents display confirmation bias with respect to their own arguments; logically, there is little doubt that the two arguments, pulling in opposite directions, attack each other. Yet from the fact that each agent underestimates his opponent's arguments beyond what is rationally warranted, a situation is created whereby deliberation may result in a non-trivial, reasonable interpretation that produces an unambiguous and fair outcome to end the disagreement regarding whether or not it will rain in Bergen today. There is, as usual in Bergen, no way of knowing for sure.

*Example 2 (Two wrongs that make a right).*

*Example 3 (An agreement that disagrees with itself).*

<sup>1</sup> The first step of our project will focus on logically examining spaces of possible rational outcomes, such as that identified here. We remark, however, that a natural next step is to try to investigate which one of these would actually result from cooperation, given some assumptions about the faculties of the agents involved, and depending on how arbitration takes place inside coalitions.

### 3.2 *The search for formal characterization of rational argumentative interaction*

The requirement that deliberation should be organized in such a way that it *always* function like this might then suggest itself as a good candidate for a normative notion of rationality imposed at the social level. It is a very strong notion, however, and it is potentially problematic also because it is not in fact wholly social – a requirement to the effect that the outcome of deliberation should always be classically consistent *must* also involve restriction on what individual views we permit agents to endorse. This is easy to see intuitively. The case of a system with a single agent who believes something absurd, for instance, or a system with many agents where all of them agree on a contradiction, are obvious examples. The fact that deliberation alone cannot ensure consistency in such cases seems hard to dispute, and it also implies that classical consistency might not be a very helpful rationality norm, when focus now shifts to the social level. This, in particular, is true if we take seriously the argumentative theory, and attempt to create standards that can be applied to *actual* reasoners, even if the circumstances are such that an outcome in complete adherence to classical principles cannot be achieved.

This is an insight that we can now formalize in terms of logic. In particular, we can formalize it as a *social* rationality principle, a claim about deliberative models. There might be many different ways of doing this, but for now let us consider the following intuitive candidate axiom for classically rational deliberation:

$$\models_{\varepsilon} \blacklozenge\phi \vee \blacklozenge\neg\phi \quad (4)$$

It says that for all argumentative states, at all states in any corresponding deliberative model, it should be the case that every formula can either be accepted or rejected. This, we recall, is the conditions under which the underlying argumentation framework describes a classically consistent interpretation of the semantic atoms in the model. It is easy to see that this axiom is *not* true on the class of all deliberative models. Hence it captures a non-trivial principle, a genuine restriction on deliberation. However, we also notice that for some argumentative states  $\mathcal{B}$ , there are *no* corresponding deliberative models such that schema 4 holds. Hence if we impose it as an axiom of deliberation, we also restrict the class of permissible argumentative states, meaning that it is not a purely deliberative approach to rationality. This, in particular, is the formal expression of the intuition that constraints on deliberation alone is not enough to ensure classical consistency in all circumstances.

This recognition does not in itself suggest that schema 4 should be discarded. Rather, we think it is interesting to investigate further in it what ways this and similar schemata can still permit *more* variety in the reasoning patterns of individuals than what is allowed under normative theories that presuppose classical reasoning at the individual level. In fact, in Theorem ?? below, we characterize the argumentative states for which scheme 4 is possible to satisfy, noting also that it includes argumentative states in which *all* the agents are individually irrational.

However, we believe that while classical logic can inspire standards of rationality that are more subtle, and easier to achieve, at the social than at the individual level, we should also look to other, less restrictive, principles. Moreover, the discussion above suggests that we are now in a position to define the following two different kinds of social rationality principles, which can help us to provide more structure to future inquiries.

- Liberal principles: Rationality constraints that do not force us to restrict the set of argumentative states that we consider possible.
- Idealistic principles: Rationality constraints that require us to restrict the set of possible argumentative states.

Given a choice of logical language, we can also search for formalizations of these two kinds of principles, and we can explore them using logical tools. A liberal principle definable by  $\mathcal{L}_1$ , for instance, is some collection  $\mathcal{E} \subseteq \mathcal{L}_1$  of formulas (possibly infinite), such that for all  $\phi \in \mathcal{E}$ , for all deliberative states  $\mathcal{B}$ , there is a model  $(Q, R)$  for  $\mathcal{B}$  such that for every state  $q \in Q$  we have

$$\mathcal{B}, (Q, R), q \models \phi$$

## 4 Conclusion