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

Sjur K. Dyrkolbotn and Truls Pedersen

Abstract We propose a formal approach to the argumentative theory of reason, combining argumentation theory and modal logic in a novel way. We argue that the resulting framework can be used to model important mechanisms identified by the theory, including how confirmation bias and other modes of reasoning typically considered fallacious may in fact serve an important argumentative purpose that can also give rise to classically sound conclusions through the process of social deliberation. We go on to suggest that the argumentative theory is based on an understanding of intelligent reasoning and rationality that sees these notions as irreducibly social, and that the argumentative theory itself provides a possible starting point in the search for new theoretical foundations based on this understanding. Moreover, we suggest that formal logic can be used also as a means to investigate foundational issues, and we sketch the development of an axiomatic approach to the study of rational deliberative interaction.

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 skeptic, however, it also begs the following question: what is the "I" in such a line of thought? The armchair philosopher might be too submerged in 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, this question is seen as crucial [Mead, 1967].

Sjur K. Dyrkolbotn
Durham Law School, Durham University, e-mail: s.k.dyrkolbotn@durham.ac.uk

Truls Pedersen
Department of Information Science and Media Studies, University of Bergen

The best known theories from 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. 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 addressed 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 interest in interdisciplinary research [Benthem, 2008, Verbrugge, 2009, 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.

We will argue that in order to provide adequate formal foundations for rational interaction we must depart from such reductionist assumptions. In search of an alternative approach, we look to the argumentative theory of reason, introduced in [Mercier and Sperber, 2011]. The key idea there is that reasoning evolved to facilitate efficient argumentation, and not necessarily to help us arrive at logically correct forms of inference, or in making reasonable decisions.

This can have important implications for logical modeling of rational interaction, and we sketch the development of a general logical framework that enables us to capture the idea formally. Our overreaching aim is to argue that a formal approach to the argumentative theory of reason can provide interesting new insights, particularly regarding the social nature of identity, reasoning and rational interaction.

The structure of the paper is as follows. In Section 2 we present the argumentative theory, focusing on its implications for our understanding of agency. We argue that the notion of argumentation that is at work challenges existing traditions in argumentation theory, particularly formal theories, and we go on to present an alternative formalization which looks at argumentative structures in a new light, as

the basis upon which agents' subjective interpretations of the world are formed. In short, we introduce the *argumentative agent*, and we argue that he should be studied further.

In Section 3 we go on to present a logical formalism for studying argumentative deliberation, interaction between argumentative agents that serve to generate novel interpretations of the world. Since our purpose in this paper is to focus on main ideas rather than technical details, we present a very simple formal framework, containing only some very basic constructions which can be developed further using existing tools from modal logic. We show through examples that this is sufficient to allow us to capture essential aspects of the mechanisms addressed in [Mercier and Sperber, 2011], and we also sketch the development of an axiomatic approach to deliberative rationality. In Section 4 we discuss future work and offer a conclusion.

2 Argumentative agents: A semantics for individual reasoning based on argumentation

The argumentative theory of reason is formulated on the basis of reassessing a large amount of experimental evidence which appears to suggest that human reasoning evolved to facilitate efficient argumentation. From the point of view of an individual, it seems that the most significant purpose of reason is not in helping him to arrive at logically correct forms of inference, but to maximize his chance of winning arguments. This, according to the theory, explains why humans so often reason in a way that is regarded as fallacious. According to the argumentative theory, reasoners sometimes commit fallacies because they are useful to them in the context of argumentation. The most obvious example is confirmation bias, the tendency to look disproportionately for reasons to support existing beliefs. It is only natural that agents who seek to win arguments will tend to focus on looking for reasons to support their own positions, rather than looking for reasons to reject them.

More generally, the argumentative theory suggests that in order to model reasoning in a descriptively accurate way, we must take into account that agents are *argumentative*. In terms of rational choice terminology, one might express this by saying that agents' utility functions are heavily influenced by their desire to do well in arguments, often to the detriment of sound classical reasoning. In fact, this is the first formal insight inspired by the argumentative theory, namely that we might need to consider argumentative utility functions, which allow us to recognize the intrinsic value that agents tend to attach to success in argumentation.

While this insight is interesting and worthy of further attention, it is not our main focus in this paper. Rather, we wish to point our and explore a second insight stemming from argumentative theory, regarding the foundations for our understanding of rationality and sound reasoning as concepts. It is striking, as pointed out in [Mercier and Sperber, 2011], how deliberation can often lead to classically sound outcomes even if each individual reasoner is argumentative and unreasonable. This,

to us, is the crucial mechanism that is flagged in the argumentative theory, one which also serves to explain why argumentative reasoning has proved so successful for the human race, even if it regularly leads to unsound, or even absurd, results, when people reason in isolation.

Following up on this, we believe that the argumentative theory suggests that we need to consider rationality principles that do not target individual reasoners at all, but rather the deliberative processes that they may come to take part in. This, we believe, is a call for further study of socially emergent normative notions of rationality, and in the following we will attempt to show that it can be facilitated using formal logic.

Towards a formal account we first need to be precise about how to represent the basic building blocks of such a theory, and we will start with the agents themselves. In this regard, it is important to note that the argumentative theory involves a notion of argumentation which is conceptually distinct from that found in traditional argumentation theory, as it has been developed in modern times following the influential work of [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 at the individual level more descriptively, and to take seriously the fact that reasoning appears to have evolved as a mechanism to facilitate efficient argumentation.

How can we formalize the argumentative agent, when argumentation is understood in this way? In the following we suggest that we can make use of argumentation frameworks, first 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 and Simari, 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]. The work of [Brewka et al., 2011], on the other hand, 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 more fine-grained formalism of abstract dialectical frameworks [Brewka and Gordon, 2010]. For our work here, it is also important to note that much recent work focuses on providing logical foundations for the theory [Grossi, 2010b, Grossi, 2010a, Arieli and Caminada, 2013, Caminada and Gabbay, 2009], which we can draw on when we develop multi-agent extensions.[1]

---

[1] 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 [Caminada et al., 2011, Caminada and Pigozzi, 2011]. It is even more clearly felt in the survey of the field given in [Rahwan and Simari, 2009].

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 the internal workings of the agents themselves. In particular, we will use argumentation frameworks to model the agents' interpretations of semantic meaning. In terms of each individual agent, using terminology from cognitive science, this 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]. It also makes good sense with respect to the argumentative theory; since agents reason to win arguments, it seems natural to assume that they also tend to represent semantic information in argumentative terms.

Moreover, the theory of argumentation frameworks provide a flexible framework for modeling many different ways in which individual agents may choose to reason, using one among the many semantics that have been developed for reasoning about such frameworks. Hence we make no commitment to a given set of reasoning rules or principles – the argumentative agent is defined by the fact that he maintains an argumentative interpretation of the world, not by the fact that he reasons about it in a given way.

2.1 Argumentation frameworks, agents and semantic views

Given a set of semantic atoms Π, which we will tend to think of 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. We can depict E as a directed graph, giving a nice visualization of how the atoms in Π are related as arguments, see Figure 1 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 $E^n(x) = \Pi \setminus E^+(x) \cup E^-(x)$. We use $\Pi(E) = \{x \in \Pi \mid E^n(x) = \Pi\}$ to denote the set of atoms from Π that does not appear in any attack from E.

Fig. 1 An AF E such that $\Pi(E) = \{p, q, q', p'\}$

Given an AF E, the purpose of an argumentation semantics is to identify, in terms of E, the collection of sets of arguments that can be accepted if they are taken together, typically called *extensions*, see e.g., [Wu et al., 2009]. 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. It is natural to represent semantics for AFs using three-valued assignments.

In particular, given an AF E and an extension $A \subseteq \Pi$, the associated three-valued assignment is $\mathsf{c}_A : \Pi \to \{1, 0, \frac{1}{2}\}$ given by

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

This representation corresponds to an intuitive reading where arguments in A are regarded as true/successful, arguments attacked by one of these are regarded as false/failed, while all other arguments are taken to have an undecided semantic status. The three-valued representation of extensions also facilitate elegant definitions of various argumentation semantics, based on thinking of semantics as prescribing collections of three-valued assignments to AFs. In this way, it also becomes natural to reason about AFs using three-valued logic, an idea that has been explored in some recent work [Dyrkolbotn and Walicki, 2013, Arieli and Caminada, 2013, Dyrkolbotn, 2013]. This will be exploited in the coming sections, as we will rely on three-valued Łukasiewicz logic when we evaluate complex formulas over given AFs.

In Figure 2 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 2 are all based on the same idea, but adds other requirements that are less obviously reasonable. In the following we will let ε refer to some generic semantics defined in Figure 2.

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

Fig. 2 Various semantics, defined for any $E \subseteq \Pi \times \Pi$

Towards a logic for reasoning about AFs we will start from the simple propositional language $\mathscr{L}$, defined by the grammar below.

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

where $p \in \Pi$. We can then define the inductive extension of three-valued assignments to arbitrary formulas from $\mathscr{L}$ as in Łukasiewicz three-valued logic, as detailed below.

$$\overline{c}(p) = c(p) \text{ for } p \in \Phi$$
$$\overline{c}(\neg\alpha) = 1 - \overline{c}(\alpha) \tag{1}$$
$$\overline{c}(\alpha \to \beta) = min\{1, 1 - (\overline{c}(\alpha) - \overline{c}(\beta))\}$$

This evaluation behaves like classical logic on the semantic values $\{0,1\}$. For an intuition of how the third value is dealt with, notice that the definition ensures that if $c(p) = \frac{1}{2}$, we have $\overline{c}(p \leftrightarrow \neg p) = 1$. In particular, the claim that p is equivalent with its own negation is true just in case p has value $\frac{1}{2}$. This means, in particular, that the logic allows us to express that some argument, or collection of arguments , have this value. This is what makes it so natural to use Łukasiewicz logic to evaluate complex formulas, as opposed to some other three-valued logic, such as Kleene's logic.

We are now prepared to offer our definition of an argumentative agent. To this end, let $\mathscr{A}$ be a set of agent names. Then a *view* for agent $a \in \mathscr{A}$ is an AF $V_a \subseteq \Pi \times \Pi$. It encodes his interpretation of the semantic relationship between the arguments under consideration. We also define an *argumentative state* as a tuple $(V_a)_{a\in\mathscr{A}}$, 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 be extended in future work, by nested application of the ideas we develop in the next sections.

Now, given an agent $a \in \mathscr{A}$ 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 $\mathscr{L}^\blacklozenge$

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

where $\alpha \in \mathscr{L}$ and $a \in \mathscr{A}$.

When we wish to reason about a single AF, we will omit reference to the single agent who instantiates it, and write simply $\blacklozenge\alpha$. Given an argumentative state $\mathscr{B} = (V_a)_{a\in\mathscr{A}}$, we can now define truth for formulas from $\mathscr{L}^\blacklozenge$ inductively as follows, for all formulas ϕ

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

For instance, assume we have a single agent a and that his view, V_a, is given by the AF in Figure 1. Then it is easily verified that the following claims holds for all semantics ε.

- $V_a \models_\varepsilon \blacklozenge q$
- $V_a \models_\varepsilon \neg\blacklozenge p \wedge \blacklozenge\neg p$
- $V_a \models_\varepsilon \blacklozenge(\neg q \wedge q')$
- $V_a \models_\varepsilon \blacksquare(\neg q \to (p \leftrightarrow \neg p))$

Having access to a formal logical language for talking about individual views of agents, we can now explore formally what we mean by classically sound reasoning. We already noted that the evaluation of complex formulas agrees with classical logic

on the boolean values. So if the argumentation semantics that the agent uses to reason only returns two-valued assignments, disallowing assignments that assign $\frac{1}{2}$ to some arguments, it might be tempting to say that he reasons classically.

However, the core requirements underlying the semantics from Figure 2 also seem to capture some crucial aspects of what we mean by classical soundness. First, and most importantly, they all require us to disallow assignments that assign 1 to two arguments that are in conflict. This is intuitively speaking a classical soundness principle, an expression of the law of non-contradiction. However, if we accept this requirement as part of our notion of what it means to reason classically, we immediately encounter situations where the underlying interpretation of the semantic reality makes this impossible. For instance, there is no way to reason classically about an argument that attacks itself. Under all semantics disallowing the assignment of 1 to two arguments in conflict such an argument must by necessity obtain the value $\frac{1}{2}$.

In terms of logic, we have $\blacksquare(x \leftrightarrow \neg x)$ expressing that x is necessarily equivalent to its own negation. In this case, classical reasoning breaks down already due to a problematic argumentative interpretation of reality, not due to any shortcoming of the principles used to reason about it. In fact, there are strong formal reasons for thinking that classical reasoning about AFs is captured by the stable semantics [Dyrkolbotn and Walicki, 2013]. Recall that the stable semantics collects all those labellings that are admissible and also satisfy the requirement that the set of arguments assigned 1 attacks all other arguments, c.f., Figure 2. Hence it also always ensures that every argument is assigned a boolean value, and in this way its classical nature becomes apparent. If we now wish to consider separately the class of those argumentative states for which classical reasoning is possible, we can do so using a formal definition which takes this to be the class of states where each agents' views admits a non-empty set of stable labellings.

The necessary failure of classical reasoning in the context of certain problematic interpretations of reality can now be characterized in terms of graph structures for which no stable assignment exists. In fact, the combinatorial problem of when an AF admits a stable set has been analyzed in graph theory since the 50s, by researchers using a different terminology, in the field of *kernel theory*. The core result from this field, due to [Richardson, 1953], immediately implies that any AF which has no directed odd cycle of attack admits a non-empty set of stable assignments. This result was also rediscovered in [Dung, 1995], but kernel theory offers many additional results and techniques, see for instance [Galeana-Snchez and Neumann-Lara, 1984]. These result can now all be understood as giving conditions which ensure the possibility of imposing classical standards of reasoning on agents' interpretations of the world.

While formalizing the argumentative agent already opens up interesting new ways of looking at classical reasoning and rationality, the main challenge that remains is to find a way to formally introduce an appropriate kind of multi-agent interaction. The formal framework should allow us to study deliberative processes which are based on the agents' interpretations of the world, and which can result in aggregated views that are potentially different from all of them. One of the questions we can then address is the question of what principles we need, at the social level,

to make sure that classical reasoning can be regained through deliberation. Before we move on to tackle this challenge, we give an extended example.

2.2 Extended example: Rain in Bergen

In this section, we consider in depth 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 represents the claim that it will rain, while $\bar{r}$ represents 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.

$$r \underset{\longrightarrow}{\longleftarrow} \bar{r} \qquad r \overset{a,b}{\underset{a,b}{\longleftarrow}} \bar{r}$$

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 $\{\bar{r}\}$ can be taken as true, for all semantics defined in Figure 2. 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 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 (1) to (3), as depicted in Figure 3.

$$(1):\qquad\qquad (2):\qquad\qquad (3):$$
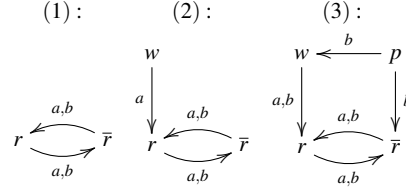
Fig. 3 An operational account of the rain debate

Certainly, this is a neat and uncontroversial depiction of the actual utterance made. Moreover, looking at the AF in (3), forgetting about the agent labels, we conclude that p, r must be assigned the value true, so that $\bar{r}, w$ are both defeated and become false. Hence we conclude that agent b won the debate. But is this really the rational way of looking at things? Applying an argumentation semantics directly to the actual exchange between a and b, it would seem so. But intuitively speaking, it seems problematic. 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 audience, a weather report might carry some weight, and so might a puddle, but neither seem particularly conclusive.

We believe the problem is with the representation we gave in Figure 3, and the fact that it failed to include information about the agents' view of each other's utterances. In a perfect world, this might not matter, since 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 common ground, and common ground 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
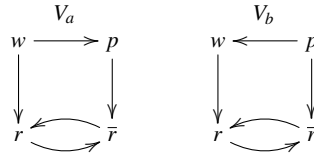
Fig. 4 Two views on rain in Bergen

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, since we now use AFs to encode subjective views, we are in a position to reject the AF that was used to encode the actual exchange of arguments. This AF encodes an interpretation that takes the attack (p, w) to encode correct semantic information about p and w. But according to his view, agent a disagrees with agent b about the presence of this attack, so it is unwarranted to take it for granted that (p, q) will come to influence the outcome of deliberation. Indeed, in order to even begin talking about the outcome we need first some aggregated view on the semantic meaning of the arguments involved. Depending on what this interpretation is taken to look like, the outcome will be different, and while it might be unclear how such an interpretation should be constructed, it seems clear that it needs to be based on the agents' views, not on the actual events that took place during the debate.

3 Argumentative deliberation: A formalization using modal 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 from the views of the individual agents. We want to study the *effect* of deliberation on semantic meaning, which, in our opinion, is the crucial question that needs to be addressed in the search for new foundations for rational interaction.

Towards formalization, we will first define the set of all possible interpretations that may result from deliberation. These will be the states of our models, and we will represent them using AFs. At this stage there is only one requirement that we require such states to meet, namely that they are all based on the views of the agents. This is encoded in the following definition.

Definition 1. Given an argumentative state $\mathscr{B}$, we say that $q \subseteq \Pi \times \Pi$ is a *deliberative state* for $\mathscr{B}$ if

$$\bigcap_{a \in \mathscr{A}} V_a \subseteq q \subseteq \bigcup_{a \in \mathscr{A}} V_a \tag{3}$$

We collect all deliberative states for $\mathscr{B}$ in the set $\mathsf{D}(\mathscr{B})$.

The requirement that states must be based on the agents' views means that we do not allow deliberation to result in states 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.[2] Having defined the set of states, we can now define the notion of a *deliberative model*.

Definition 2. Given a deliberative state $\mathscr{B}$, a deliberative model for $\mathscr{B}$ is a tuple (Q, R) such that

- $Q \subseteq \mathsf{D}(\mathscr{B})$ is a set of deliberative states for $\mathscr{B}$.
- 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 $\mathscr{B}$. If $(q_1, q_2) \in Q$ 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. We abstract away from the deliberative 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, say a vote, to reach a decision about some argument. In this paper, we will keep things simple and leave the exact content of events unspecified.

As an example of a deliberative model, consider the framework in Figure 5. Here, the argumentative state is problematic from the point of view of classical logic. In particular, we have $\mathscr{B} \models_\varepsilon \neg \blacklozenge_a x \wedge \neg \blacklozenge_a \neg x$ under all ε from Figure 2, arising from the fact that in agent a's view, 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 $\mathscr{B} \models_\varepsilon \blacklozenge_b y \wedge \blacklozenge_b \neg y$. 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 2, 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 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

---

[2] Interestingly, this does not mean that a possible unanimity regarding the semantic status of an argument is necessarily reflected in the view aggregated by deliberation, not even when all agents reason according to the same semantics. If the agents differ in their account of *why* an argument should be accepted, in particular, deliberation might lead to its rejection. We will formalize a scenario like this in Section 3.1, Example 3.

$V_a:$ $x \longrightarrow y$, $V_b:$ $x \rightleftarrows y$

$q_0: x$ y , $q_1:$ $x \longrightarrow y$, $q_2:$ $x \rightleftarrows y$
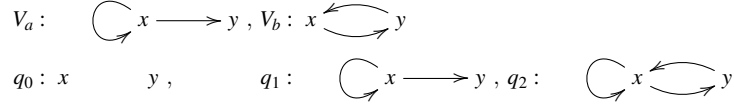
Fig. 5 A deliberative model (Q,R) over $\mathscr{B} = (V_a, V_b)$ with $R = \{(q_0,q_1),(q_1,q_2)\}$

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. We will focus on making conceptual points, so we consider the following simple language $\mathscr{L}_1$, which adds to $\mathscr{L}^{\blacklozenge}$ a modality for talking about the current deliberative state and the one-step possibilities in deliberative models.

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

where $\alpha \in \mathscr{L}$. We now use $\blacklozenge$ as a modality to talk about the semantic status of arguments at a deliberative states, while the agent-indexed modalities still apply to the agents' views in the argumentative state. The definition of satisfaction for $\mathscr{L}_1$ on deliberative models is then defined analogously to classical modal logic.

Definition 3. Given an argumentation semantics ε, an argumentative state $\mathscr{B}$ and a corresponding deliberative model (Q,R), the truth of $\phi \in \mathscr{L}_1$ on (Q,R) at $q \in Q$ is defined inductively as follows (we omit the boolean cases).

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

We use the shorthand $\Box \phi := \neg \Diamond \neg \phi$ as usual. Let us consider the model from Figure 5 as an example. Then it is easy to verify that $\mathscr{B},(Q,R),q_0 \models_\varepsilon \Box\Box\blacksquare\neg x$, 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 $\mathscr{B},(Q,R),q_0 \models_\varepsilon \Diamond\Diamond\blacklozenge\neg x$. 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 deem 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 Section 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 Section 3.2, 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, that we can then explore using the expressive power of more complex modal languages.

3.1 Using deliberative logic to model argumentative deliberation

In [Mercier and Sperber, 2011], 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 Section 2.2. There we argued that instead of directly modeling the actual exchange of arguments, the deliberative events that took place, we should start from a representation of the agents' view of the arguments. These were given in Equation 4, and we argued that these views were consistent with the deliberative event under consideration. Now we can try again to model the deliberation that took place, not by making a naive description of the actual utterances, but more abstractly by seeing it as a traversal on a deliberative model. To do this, we will consider the possible effect that each utterance could have on the deliberative state, given a starting point where r and $\bar{r}$ are in recognized mutual opposition to each other. We recall that the first event was that agent a pointed out the weather report, and that the second event was agent b pointing to the puddle. This leads us then naturally to consider the deliberative model (Q,R), depicted in Figure 6.

What do we get from this model? Let us assume, for an intuitive presentation of the scenario, that the model describes a situation where deliberation is based on
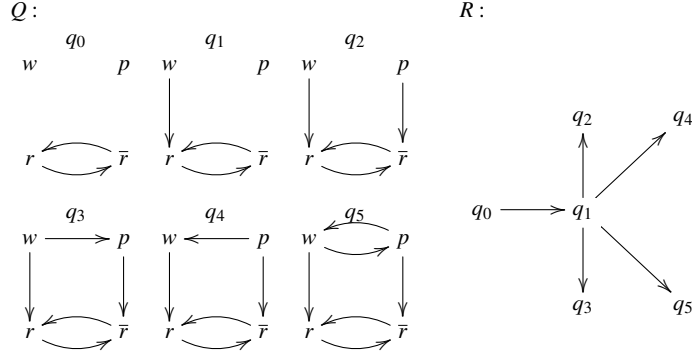
Fig. 6 The deliberative model (Q,R) based on the argumentative state $\mathscr{B}$ defined in Figure 4

searching for agreement about what attacks *not* to include. That is, unless the agents agree on something else we always end up in the deliberative state which is formed by taking the union of the agents' view. Then, if disagreement runs deep, the Bergen rain debate will end in state q_5. Here it is not hard to see that the question of whether or not it will rain remains unsettled – 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 open in the social group $\{a,b\}$, both can make up his own mind and disagreement may persist. However, if the agents are willing to consider a consensus, then they can settle on either r or $\bar{r}$, by moving to state q_3 or q_4 respectively. 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 agreement on including only those attacks that are present in both views. This means that they end in state q_2, and here both w and p are regarded as successful, meaning that *both r and $\bar{r}$* becomes defeated and impossible to accept.[3]

   This consensus outcome is interesting since it seems plausible that in an actual debate, this is what one would get if the agents respect each others' input. Hence 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 looses on grounds that he believes to be unreasonable. More interesting still, it is 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

---

[3] In this paper we only sketch a framework that permits us to logically examine spaces of possible outcomes, such as those identified by (Q,R). 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.

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.

*Example 2 (Two wrongs that make a right).* Consider the following argumentative state:
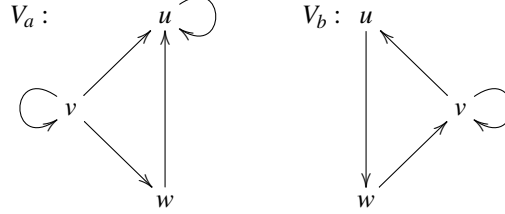
Here the two agents both have an inconsistent view on the semantic elements, as the reader may easily verify. Moreover, the agents agree that v attacks itself. But even so, deliberation can result in consistency being regained. In particular, the AF depicted below is a deliberative state for (V_a, V_b) and it is easily seen to be classically consistent, under the evaluation $\{v \mapsto 0, w \mapsto 1, u \mapsto 0\}$
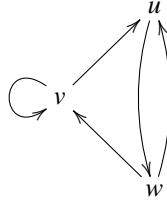
In fact, we can say more about deliberation based on (V_a, V_b). It is true, in particular, that we can evaluate the three arguments classically if and only if we end up in a deliberative state where w is understood to attack both v and u. How to ensure that our deliberation leads to this result is unclear, but at least we can recognize it as a possible way in which, for this argumentative state, two wrongs could in fact come to make it right.

Example 3 (An agreement that disagrees with itself). While deliberation can some-times take us from an inconsistent argumentative state to deliberative states that admit classical evaluation, the direction of deliberation can also go in the other di-rection. Consider, for instance, the following views
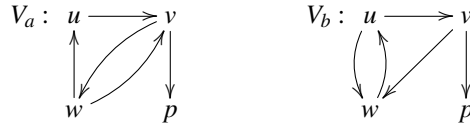
For $\mathscr{B} = (V_a, V_b)$ we have $\mathscr{B} \models_\varepsilon \blacksquare_a(\neg v \wedge p) \wedge \blacksquare_b(\neg v \wedge p)$. That is, the agents both have an interpretation of the world such that v comes out false and p comes out

true under all labellings permitted by any of the semantics from Figure 2. Moreover, both views can be evaluated classically, as the reader may easily verify. But the agents disagree about something else, namely the status of u and w. In particular, we have $\mathscr{B} \models_{\varepsilon} \blacksquare_a(w \wedge \neg u) \wedge \blacksquare_b(\neg w \wedge u)$. This disagreement could spell disaster for deliberation taking place on the basis of $\mathscr{B}$. It is easy to see, in particular, that the following state q could result through deliberation.

$$
\begin{array}{ccc}
u & \longrightarrow & v \\
\uparrow & \diagdown & \downarrow \\
w & & p
\end{array}
$$
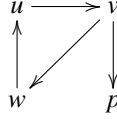
In this case there are *no* classical evaluations available. That is, for any deliberative model containing q, we have $\mathscr{B}, (Q, R), q \models \neg \blacklozenge x \wedge \neg \blacklozenge \neg x$ for all $x \in \{o, v, w, p\}$. Hence deliberation, in this case, took unproblematic views and produced a problematic outcome. How to avoid this mechanism in general is a difficult problem that deserves further study, but we can already conclude that our model allows us to capture how deliberation can sometimes end up undermining agreements about the status of arguments, and can also lead to the introduction of *new* inconsistencies. Indeed, this too is a mechanism that is intuitively clear. For instance, if one man happens to be right, for all the right reasons, while everyone else either disagrees with his conclusion or his reasons, then deliberation typically lead to suboptimal outcomes. This is important to remember, all the while the other, more positive potential in deliberation, is essentially just an expression of the same mechanism. The whole is more than the sum of its parts, and in order to explore and evaluate it we need new tools and principles that sees it as such.

3.2 The search for formal characterizations of rational argumentative interaction

Having set up a logical framework for describing argumentative deliberation, we are now also in a position to begin exploring formal representations of principles of rational argumentative interaction. We can think of such principles as restrictions on the class of deliberative models we consider permissible, allowing us to distinguish the "good" scenarios from the "bad" ones in a formally precise way. This can be done either semantically, in terms of direct mathematical definition of the appropriate classes of models, or axiomatically, by using formulas and schemata from some formal language to express key properties that we believe can serve to characterize good deliberation.

We have already seen how deliberative logic allows us to model scenarios where the outcome is classically sound even if all individual views are inconsistent. The requirement that deliberation should be organized in such a way that it *always* functions in this way 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. In particular, a requirement to the effect that the outcome of deliberation should always be classically consistent must by necessity 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 share an inconsistent interpretation of the world are obvious examples. The fact that deliberation alone cannot ensure a consistent outcome in such cases seems clear, and it is an insight that we can now express formally. First we must define an appropriate formalization of classical soundness at the level of deliberation, and how to do this is not obvious. We will discuss some subtleties regarding this later, but for now let us simply consider the following intuitive axiom schema for classically rational deliberation.

$$\blacklozenge \neg(\phi \leftrightarrow \neg\phi) \tag{4}$$

If we require it to be true on all models, in all points, we stipulate that for all claims we can express about the model, it should always be possible to reason about this claim in such a way that it is not considered to be equivalent to its own negation. The restriction is perhaps too weak, but at least it appears like a reasonable minimal requirement. We note, moreover, that the scheme is *not* valid 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 $\mathscr{B}$, there are *no* corresponding deliberative models such that Schema 4 holds. Therefore, if we impose it as an axiom of deliberation, we also restrict the class of permissible argumentative states, meaning that it does not provide 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 imply 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. 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.

An example of a liberal principle could for instance be $\neg\square(\neg p \wedge p)$, expressing that the relations used to form deliberative models must be serial, in just the same way as in classical modal logic. In the context of deliberation it would be the principle of that deliberation is open-ended, that there is always a deliberative next step

(although at some point it might just be an endless repetition of previously visited states). A more subtle example, involving deliberative interactions, is the principle $\blacklozenge\phi \to \Diamond\blacklozenge\phi$ which expresses that if something is true in a deliberative state it should also be true in all following states, encoding commitment to previous outcomes. This is not a restriction on the kinds of relations that are allowed to exist between deliberative states, but rather a restriction on how deliberation is allowed to unfold from the argumentative state. Intuitively speaking, it restricts the kinds of deliberative events we allow. It is easy to see that it is liberal, however, since a single state without successors will always satisfy it. Note that if we add a loop to this state, it witnesses to the liberality of the principle which requires seriality plus commitment to previous outcomes; a debate that never ends and always strictly increases the set of truths it produces.

For an example of an idealistic principle, it is enough to point out that Schema 4 is idealistic since it excludes certain argumentative states. In fact, we can provide a simple characterization of those argumentative states that are permitted. Let us say that an argumentative state $\mathscr{B}$ satisfies an axiom schema if there is some deliberative model based on this argumentative state for which the schema is true in all states. Moreover, let us say that an argumentative state $\mathscr{B} = (V_a)_{a\in\mathscr{A}}$ is finite if $\Pi(V_a)$ is finite for all $a \in \mathscr{A}$. Then we have the following result.

Theorem 1. *For all semantics ε from Figure 2, a finite argumentative state $\mathscr{B} = (V_a)_{a\in\mathscr{A}}$ satisfies Schema 4 if, and only if, there is some deliberative state q for $\mathscr{B}$ such that $s(q) \neq \emptyset$*

Proof. $\Rightarrow$) We let $A = \bigcup_{a\in\mathscr{A}} \Pi(V_a)$ and form the conjunction $\phi = \bigwedge_{x\in A}(x \vee \neg x)$. We notice that for all three-valued assignments c we have $\bar{\mathsf{c}}(\phi) \in \{1, \frac{1}{2}\}$ and that $\bar{\mathsf{c}}(\phi) = \frac{1}{2}$ if, and only if, there is some $x \in A$ such that $\mathsf{c}(x) = \frac{1}{2}$. Let q be some deliberative state for $\mathscr{B}$ such that $\blacklozenge\neg(\phi \leftrightarrow \neg\phi)$ is true on q under ε. Hence there is some $\mathsf{c} \in \varepsilon(q)$ such that $\mathsf{c}(x) \in \{1, 0\}$ for all $x \in A$. Since all other atoms from Π are not involved in attacks from q, it follows that $s(q) \neq \emptyset$.
$\Leftarrow$) Choosing some arbitrary $\mathsf{c} \in s(q)$ suffices to choose the claim for all ε defined in Figure 2, since for all such ε we have $s(E) \subseteq \varepsilon(E)$ for all AFs E.

This theorem is a simple illustration of the kinds of results we can obtain when we begin to formalize deliberative principles using logic. It also illustrates the subtleties involved, and how classical concepts might take on new and surprising forms. Notice, for instance, how the case of two wrongs that make a right, considered in Example 2, follows from the result. So according to Schema 4, such a collection of views will be permitted, even if each view is not in itself classically consistent. Rather than forbidding the views, we must instead exclude some deliberations that they can give rise to.

How to do this can be a tricky question, and the language of $\mathscr{L}_1$ might be too weak to deliver principles for deliberation that allow us to get very far in this regard. Consider for instance a scenario where deliberation proceeds in a step-wise fashion, such that one argument is considered at a time starting from the deliberative state which contains no attacks. This protocol itself seems perfectly reasonable, but it

now becomes highly unreasonable to require classical consistency at every state. It makes more sense to stipulate weaker notions such that it should *eventually* hold, after deliberation has had a chance to work. In general, the ability to express that something holds eventually is an important addition the expressive power of out logical language, which at once allows us to consider a whole range of interesting questions.

In the following we merely sketch some of these, intended to serve as an illustration of the great potential for using stronger modal languages to reason about deliberative models. In particular, we define the modality $\Diamond^*\phi$, intuitively to be read as saying "after finitely many steps, ϕ becomes true". Formally, we let $\Diamond^n\phi$ denote $\overbrace{\Diamond\Diamond\ldots\Diamond}^{n}\phi$ and define satisfaction for $\Diamond^*\phi$ inductively as follows

$$\mathscr{B},(Q,R),q \models_\varepsilon \Diamond^*\phi \text{ if there is } n \in \mathbf{N} : \mathscr{B},(q,R),q \models_\varepsilon \Diamond^n\phi \qquad (5)$$

We also define $\Box^*\phi := \neg\Diamond^*\neg\phi$. This then expresses "always ϕ". With these constructs in hand we can express many subtly different properties of deliberation, some of which might be seen as candidates for rationality principles.

Let us assume that ϕ expresses some principle which we take to define "good" states in a normative theory. For instance, ϕ could be an instance of Schema 4. However, even if we believe that ϕ captures some essential normative requirement on the outcome of deliberation, it is not at all clear that we should require *all* states in a deliberative model to be good states. Indeed, it can often seem more natural to think of deliberation as a process that complies with a norm of rationality just in case it is capable of taking us from bad to good states. In this case, we are cheating it we try to implement our normative theory by simply forbidding the bad states. Instead, we might want to restate our principle ϕ in one of the following ways, as a requirement on what it should be possible to achieve through deliberation, from the current state, rather than a requirement on what the current state itself should be like.

- $\Box\phi$; all deliberative events take us to a state where ϕ is true.
- $\Diamond\phi$; there is at least one event taking us to a state where ϕ is true.
- $\Diamond^*\phi$; there is a chain of events such that ϕ eventually becomes true.
- $\Diamond^*\Box\phi$; there is a chain of events taking us to a state where every event will make ϕ true.
- $\Diamond^*\Box^*\phi$; there is a chain of events taking us to a state where no further chain of events can make ϕ false.
- $\Box^*\Diamond^*\phi$; for every chain of events, there is a way to continue this chain so that ϕ eventually becomes true.
- $\Box^*\Diamond^*\Box^*\phi$; for every chain of events, there is a continuation so that ϕ eventually becomes true, and remains true forever.

These are just some simple examples of notions that can be formulated in temporal logic. What makes them interesting to us is that in the present setting they express properties of deliberation. In the context of temporal logic they are quite

well understood, but when ϕ also involves occurrences of modalities such as $\blacklozenge$, expressing argumentative properties of the states, new questions arise, both technical and philosophical. One of the most interesting concerns the question of expressive power, and in future work we hope to consider various temporal and fixed-point languages and to investigate what they allow us to say about deliberative models. Moreover, we think it will be very fruitful to combine this technical work with addressing the philosophical challenge of developing a better foundational understanding of the nature of argumentative deliberation. To formulate and study normative principles of deliberative rationality axiomatically is a particularly interesting aspect of this work, and we believe the preliminary investigation carried out here shows its promise.

4 Conclusion

We have presented a formal approach to the argumentative theory of reason and argued that it should be studied further, also using logical tools. To us, the argumentative theory is particularly interesting because it suggests an alternative approach to rationality, one which sees it as irreducibly social and emergent from deliberation. As such we think it can contribute important insight to the search for new theoretical foundations for rationality, foundations that could prove relevant to a range of different fields, including artificial intelligence, which increasingly focuses its attention on social processes and the study of rational and intelligent interaction.

For future work we hope that the simple formal framework developed in this paper can provide a point of departure for the development of more sophisticated technical tools. In the first instance, these should be developed so that they allow clear and informative mathematical modeling of important aspects of argumentative deliberation. But we also think that such formalisms can facilitate a highly interesting axiomatic approach to deliberative rationality, allowing formal logical investigation of foundational issues, assisted by the use of temporal and fixed-point modal languages interpreted on argumentative structures. On the technical side, we think the study of the expressive power of various languages that allow interactions between computational and argumentative modalities is highly interesting and should be considered further.

The study of deliberative models is the study of how representations of meaning and communicative events together give rise to branching structures of deliberation, where what is truly sound and rational might be hidden somewhere in a combinatorial object, a recurring pattern, or else not emerge at all except in the limit. Wherever it is hiding, and whatever its true nature might be, we believe the pursuit itself has great value, and that it should continue. In the end, perhaps the search for truth is itself a deliberative imperative, the normative power of which is derived not from the fact that it may some day settle, but from the fact that it must always be carried on.

References

[Arieli and Caminada, 2013] Arieli, O. and Caminada, M. W. A. (2013). A QBF-based formalization of abstract argumentation semantics. *Journal of Applied Logic*, 11(2):229 – 252.

[Baroni and Giacomin, 2007] Baroni, P. and Giacomin, M. (2007). On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(1015):675700.

[Benthem, 2008] Benthem, J. v. (2008). Logic and reasoning: Do the facts matter? *Studia Logica*, 88(1):6784.

[Benthem, 2011] Benthem, J. v. (2011). *Logical Dynamics of Information and Interaction*. Cambridge University Press.

[Blume and Durlauf, 2001] Blume, L. and Durlauf, S. (2001). The interaction-based approach to socioeconomic behavior. In Durlauf, S. and Young, P., editors, *Social Dynamics*. Brookings Institutions Press.

[Brewka et al., 2011] Brewka, G., Dunne, P. E., and Woltran, S. (2011). Relating the semantics of abstract dialectical frameworks and standard AFs. In *IJCAI*, pages 780–785.

[Brewka and Gordon, 2010] Brewka, G. and Gordon, T. F. (2010). Carneades and abstract dialectical frameworks: A reconstruction. In *Proceedings of the 2010 conference on Computational Models of Argument: Proceedings of COMMA 2010*, page 312, Amsterdam, The Netherlands, The Netherlands. IOS Press.

[Caminada and Pigozzi, 2011] Caminada, M. and Pigozzi, G. (2011). On judgment aggregation in abstract argumentation. *JAAMAS*, 22(1):64102.

[Caminada et al., 2011] Caminada, M., Pigozzi, G., and Podlaszewski, M. (2011). Manipulation in group argument evaluation. In Walsh, T., editor, *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI-11)*. AAAI Press/International Joint Conferences on Artificial Intelligence.

[Caminada and Gabbay, 2009] Caminada, M. W. A. and Gabbay, D. M. (2009). A logical account of formal argumentation. *Studia Logica*, 93(2-3):109145.

[Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321357.

[Dworkin, 1986] Dworkin, R. (1986). *Laws Empire*. Harvard University Press.

[Dyrkolbotn, 2013] Dyrkolbotn, S. (2013). The same, similar, or just completely different? equivalence for argumentation in light of logic. In Libkin, L., Kohlenbach, U., and Queiroz, R., editors, *Logic, Language, Information, and Computation*, volume 8071 of *Lecture Notes in Computer Science*, pages 96–110. Springer Berlin Heidelberg.

[Dyrkolbotn and Walicki, 2013] Dyrkolbotn, S. and Walicki, M. (2013). Propositional discourse logic. *Synthese*, pages 1–37. [Online first publication, http://dx.doi.org/10.1007/s11229-013-0297-x].

[Galeana-Snchez and Neumann-Lara, 1984] Galeana-Snchez, H. and Neumann-Lara, V. (1984). On kernels and semikernels of digraphs. *Discrete Mathematics*, 48(1):6776.

[Grossi, 2010a] Grossi, D. (2010a). Argumentation in the view of modal logic. In *ArgMAS*, pages 190–208.

[Grossi, 2010b] Grossi, D. (2010b). On the logic of argumentation theory. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1 - Volume 1*, AAMAS '10, page 409416, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Mead, 1967] Mead, G. H. (1967). *Mind, Self and Society*. University of Chicago Press, Chicago, 3 edition.

[Mercier and Sperber, 2011] Mercier, H. and Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(02):5774.

[Ossowski, 2013] Ossowski, S., editor (2013). *Agreement Technologies*, volume 8 of *Law, Governance and Technology Series*. Springer.

[Parikh, 2001] Parikh, R. (2001). Social software. *Synthese*, 132:2002.

[Rahwan and Simari, 2009] Rahwan, I. and Simari, G. R., editors (2009). *Argumentation in arti-ficial intelligence*. Springer.

[Richardson, 1953] Richardson, M. (1953). Solutions of irreflexive relations. *The Annals of Mathematics, Second Series*, 58(3):573590.

[Stenning and van Lambalgen, 2005] Stenning, K. and van Lambalgen, M. (2005). Semantic interpretation as computation in nonmonotonic logic: The real meaning of the suppression task. *Cognitive Science*, 29(6):919960.

[Toulmin, 2003] Toulmin, S. (2003). *The Uses of Argument*. Cambridge University Press, 2 edition. First edition from 1958.

[Verbrugge, 2009] Verbrugge, R. (2009). Logic and social cognition the facts matter, and so do computational models. *J. Philosophical Logic*, 38(6):649–680.

[Waal and Ferrari, 2010] Waal, F. B. M. d. and Ferrari, P. F. (2010). Towards a bottom-up perspective on animal and human cognition. *Trends in Cognitive Sciences*, 14(5):201 – 207.

[Wu et al., 2009] Wu, Y., Caminada, M. W., and Gabbay, D. M. (2009). Complete extensions in argumentation coincide with 3-valued stable models in logic programming. *Studia Logica*, 93(2-3):383–403.