On a Formal Connection between Truth, Argumentation and Belief

Sjur Dyrkolbotn

Durham Law School, Durham University, UK
s.k.dyrkolbotn@durham.ac.uk

Abstract. Building on recent connections established between formal models used to study truth and argumentation, we define logics for reasoning about them that we then go on to axiomatize, relying on a link with three-valued Łukasiewicz logic. The first set of logics we introduce are based on formalizing so called skeptical reasoning, and our result shows that a range of semantics that are distinct for particular models coincide at the level of validities. Then, responding to the challenge that our logics do not capture credulous reasoning, we explore modal extensions, leading us to introduce models of three-valued belief induced by argument. We go on to take a preliminary look at some formal properties of this framework, offer a conjecture, then conclude by presenting some challenges for future work.

1 Introduction

There are close formal connections between argumentation, truth and kernels in directed graphs. This was first observed in [11] (truth and kernels) and [13] (kernels and argumentation), and all three were considered together in [18], where a formal link to Łukasiewicz three-valued logic was noted, see also [42]. Here we build on this work, first by characterizing the propositional validities of skeptical argumentation, and then by proposing a modal extension that allows us to capture credulous reasoning. This leads us to introduce three-valued models of belief induced by argumentation frameworks, and we offer a preliminary investigation of their properties.

The structure of the paper is as follows. In Sect. 2 we survey the connections between kernel theory and models used to study truth and argumentation, and we introduce basic concepts and notation. Then in Sect. 3 we introduce various argumentation-based logics for reasoning about these models, using Ł_3 to provide axiomatizations of their validities. In Sect. 4 we develop modal extensions of these logics. We argue that three-valued $KD45$ logic is a good candidate for reasoning about truth and argumentation, and we conjecture that the classes of serial models induced from argumentation frameworks under preferred and semi-stable semantics respectively are in fact canonical for this logic. In Sect. 5 we offer a conclusion.

2 Truth, Kernels and Argumentation

To set the stage for the novel work we present in later sections, we now give a summary of the connections that have been established previously. This serves both as necessary background and further motivation for the questions we address later.

M. Colinet et al. (Eds.): ESSLLI 2012/2013, LNCS 8607, pp. 69–90, 2014.

2.1 Truth

The search for truth is often carried out on the basis of an implicit and imprecise understanding of what exactly constitutes it. In most academic fields, there are methodological principles and procedural safe-guards in place to ensure that accepted results are indeed truthful, but these are dependent on context and differ from field to field. In philosophy, on the other hand, the search for a unified theory of truth has a long tradition and it is often carried out analytically, by assuming as few primitive notions as possible and abstracting away from contextual factors to the greatest possible extent. Indeed, a large body of work is devoted to the *formal* study of truth, much of which is based on logically examining this Aristotelian principle[1], an approach due to [39]:

T: *A statement is true if, and only if, what it says is the case*

As a theory of truth this might seem like an uninformative truism, but hard philosophical problems arise already at this level. The paradigmatic example is the *liar statement*: "this statement is false". If it is true, then it must be false according to T, since this is what it says. On the other hand, if it is false, then we must conclude, again by T, that it is true. This is problematic, particularly to those who think truth and falsehood are mutually exclusive, as one would expect from how these notions are ordinarily used.

For formal theories of truth, semantic paradoxes such as the liar occupy pride of place. Indeed, they are the first obstacle that arises, even for the most rudimentary theoretical accounts.[2] This presents us with a surprising problem: either something is wrong with the rules of classical logic, or else something is wrong with the intuitively obvious principle T.

To address this problem using logic, it is common to formalize T in some system of predicate logic, with truth as a predicate. Much work has been carried out in this tradition, often focusing on the question of how to modify T to arrive at a theory which does not lead to paradoxes such as the liar. However, according to some philosophers, most notably Kripke, the paradoxes do not serve to demonstrate fault with principle T, they merely show that truth is *partially defined* [25].

It has long been accepted wisdom that referential patterns play a crucial role in the emergence of paradox. The liar, for instance, is viciously circular, explicitly negating itself. It is tempting to depict referential patterns using graph-structures. For instance, the liar can be pictured as a directed graph with a single vertex pointing to itself:

Much formal work on truth makes use of graphs as pictures, an idea that was first developed formally in [3]. Here, non-wellfounded sets are used to define the semantics for self-referential statements, and graphs are used to depict such sets, following the work of [1]. More recently, it has been observed that graph-structures can also be used

[1] See Aristotle's *Metaphysics*, 1011b, 26.

[2] In this paper we think of a paradox as a contradiction which we arrive at from premises that we think are uncontroversial. This only captures what Quine calls the falsidical paradoxes [33], but one might argue that the other kind he proposes - the veridical ones - are not really paradoxes at all, but merely surprising *facts*.

to represent the semantic content of statements directly, as an alternative to a more traditional formulation in predicate logic. This idea is due to [11], who noted that one might as well interpret edges in a directed graph (digraph) as negations and branching as conjunction.

Towards formalization, assume we have a collection Π of atoms, thought of as statement names, including a constant $\mathbf{1}$ denoting some arbitrary true statement. Then for any index set I, a *truth-theory* of cardinality $|I|$ is a collection $\bigcup_{i\in I}\{x_i \leftrightarrow \bigwedge\{\neg x \mid x \in X_i\}\}$ where $x_i \in \Pi, X_i \subseteq \Pi$ for all $i \in I$. A truth-theory is *finitary* if X_i is finite for all $i \in I$. Truth-theories encode instantiations of the principle T, applied to concrete sets of statements referring to each other.[3] The reader might worry that the form assumed for formulas appearing on the right of an equivalence is overly restrictive, but in [5] it was shown that truth-theories provide a normal form for propositional theories, so the format is in fact fully general.[4]

We can now *define* paradox, saying that a truth-theory is paradoxical just in case it is classically inconsistent [18]. This captures the liar: the truth-theory $\{p \leftrightarrow \neg p\}$ is obviously an inconsistent theory. Truth-theories might seem trivial and uninteresting, but the connection we can set up with digraphs makes them very useful. In the next subsection, we will argue for this in some depth, showing how the combinatorial perspective provides a great template for further exploration of when principle T becomes problematic.

2.2 Kernels

A directed graph over Π is a set $N \subseteq \Pi \times \Pi$ of directed edges. When $(x, y) \in N$, we write $y \in N(x)$ and $x \in N^-(y)$ (so N^- is the converse of N) and we extend this notation to sets, e.g., such for $X \subseteq \Pi$ we have $N(X) = \bigcup_{x\in X} N(x)$. We say that a digraph is finite if N is finite and *finitary* if $N(x)$ is finite for all $x \in \Pi$. The set $\Pi(N)$ is used to denote $\{x \mid N(x) \cup N^-(x) \neq \emptyset\}$, the set of atoms that stand in a relation to some other atom in N. Moreover, a digraph N' is said to be a *subdigraph* of N if $N' = \{(x, y) \in N \mid x, y \in \Pi(N')\}$.

The connection between truth-theories and digraphs can now be expressed in two simple equations. First, for all digraphs N we let $sinks(N)$ denote the set of atoms without outgoing edges. Then we form the corresponding truth-theory defined as follows:

$$\mathsf{T}(N) = \bigcup_{x\in\Pi(N)\setminus sinks(N)} \{x \leftrightarrow \bigwedge_{y\in N(x)} \neg y\} \cup \{\bigcup_{x\in sinks(N)} x \leftrightarrow \mathbf{1}\} \tag{2.1}$$

Conversely, if T is a truth-theory indexed by I we define the digraph N_T:

$$N_T = \bigcup_{i\in I}\{(x_i, x) \mid x \in X_i \setminus \{\mathbf{1}\}\} \tag{2.2}$$

[3] We remark that their concreteness means that we might as well omit explicit representation of truth as a predicate, e.g., not bother to write $T(p) \leftrightarrow \neg T(q) \wedge \neg T(r)$ (interpreting p, q, r as constants).

[4] This means that work on this formalism also has potential importance to the study of boolean satisfiability, as explored in [41].

When does N correspond to a non-paradoxical truth-theory? It is straightforward to verify that an assignment $f : \Pi(N) \to \{1, 0\}$ satisfies $\mathsf{T}(N)$ under classical logic if, and only if, we have the following for all $x \in \Pi(N)$:

$$f(x) = 1 \Leftrightarrow f(y) = 0 \text{ for all } y \in N(x) \tag{2.3}$$

Translating this into the language of directed graphs it follows that N is non-paradoxical if, and only if, it admits some set $K \subseteq \Pi(N)$ such that:

$$N^-(K) = \Pi(E) \setminus K \tag{2.4}$$

As observed in [11], sets satisfying the above equation are known as *kernels* in graph theory. They were introduced by Von Neumann and Morgenstern to provide an abstract solution concept in cooperative game theory [40], and have attracted quite some theoretical interest, see [7] for an overview of the field. The connection with kernels means that the problem of paradox can be addressed graph-theoretically. In particular, let us write $Kr(N)$ for the set of kernels in a digraph N. Then the problem of paradox can be rephrased as follows: for what N do we have $Kr(N) \neq \emptyset$?

Many results on this have been obtained in kernel theory, most of which provide sufficient conditions for the existence of kernels [14,15,21]. Sufficient conditions actually tend to ensure something stronger than existence of kernels, namely *kernel perfectness*: the existence of kernels in all induced subdigraphs.

The first non-trivial result that was established states that a finitary digraph with no odd-length cycle is kernel perfect, due to [38].[5] The original proof is rather complicated, but was greatly simplified by the introduction of the notion of a *semikernel* [30]. A semikernel in a digraph N is a set $S \subseteq \Pi$ such that:

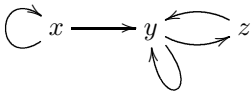
$$N(S) \subseteq N^-(S) \subseteq \Pi \setminus S \tag{2.5}$$

In other words, S is a semikernel if everything it points to is outside it and points back into it.[6] In particular, if S is a semikernel in N then it is a kernel in the subdigraph induced by $N(S) \cup S$. In other words, S witnesses to the fact that by restricting attention to this set and the statements it refers to, paradox can be avoided.

Given a digraph N, we use $Lk(N)$ to denote the set of all semikernels in N. Notice that $\emptyset \in Lk(N)$ for any N, and that the loop does not have any non-empty semikernel. A digraph can have a non-empty semikernel without having a kernel, however, as illustrated by the following digraph N:

[5] This does not hold for infinitary digraphs, the standard example being Yablo's paradox $\bigcup_{i \in \mathbb{N}} \{(i, j) \mid j > i\}$ [43]. The study of conditions applying to infinitary digraphs is harder and less progress has been made (but see [28,20,34,5]).

[6] In truth-theory terms, all statements negated by S are outside and in turn negate at least one member of S. Notice that if we assume such a collection to consists only of true statements, their truth can be verified by a constructive form of circular reasoning: all statements they negate are indeed false since they in turn negate a statement assumed to be true. In particular, the assuming their truth is perfectly consistent, not a paradox.



(2.6)

Since they are self-negating, neither x nor y can be in a kernel. But then as x only negates y it follows that x could not possibly negate a member of any kernel, so $Kr(N) = \emptyset$. But we have $Lk(N) = \{\{z\}\}$, since z both negates and is negated by y. The technical importance of semikernels stems from the following result.

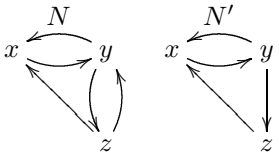
Theorem 2.7 *[30] A digraph N is kernel perfect if, and only if, every non-empty induced subdigraph of N admits a non-empty semikernel.*

In light of this result, it is possible to establish conditions that ensure kernel perfectness by showing that they ensure existence of non-empty semikernels for every non-empty induced subdigraph. Since semikernels are formulated locally on the digraph, this can be very helpful, and it is the approach followed in most work in kernel theory. The following theorem summarizes the most significant results. Recall that a *chord* on a cycle is an edge connecting two non-consecutive vertices.

Theorem 2.8 *For all digraphs N, we have that $Kr(N) \neq \emptyset$ if every odd cycle in N has one of the following*

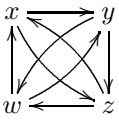
1. *at least two symmetric edges [14],*
2. *at least two crossing consecutive chords [15] or*
3. *at least two chords with consecutive targets [21].*

As an example, consider the digraph N depicted on the left below. It has a kernel, and this is ensured by all points of Theorem 2.8.



(2.9)

In fact, N has two kernels: $\{x\}$ and $\{y\}$. We notice that one of these, $\{x\}$, is also a kernel in N'. But this does not follow from any of the results from Theorem 2.8. This is interesting because it suggests that some simple cases are not covered, motivating further work. However, a natural conjecture stating that a digraph has a kernel if every odd cycle has *one* reversible edge is not true, as witnessed by the following digraph:[7]



(2.10)

[7] It holds for the special case of a single odd cycle, however: take the target of some symmetric edge, skip two vertices, and from then on take every other vertex as you move along the cycle. You end up with a kernel, the only kernel admitted by this digraph.

The problem is that the odd cycles in this digraph interact in ways that make it impossible to solve them all simultaneously. This problem of *compatibility* is the essence of what makes the search for sufficient conditions both interesting and difficult. Semikernels and inductive arguments to establish kernel perfectness is the standard way to address it, but we mention that a new approach was recently introduced in [17]. This paper introduced the following notion, which is useful for proving sufficient conditions for the existence of kernels in digraphs that are not kernel-perfect.

Definition 2.11 *A* solver *for a digraph N is a sequence of induced subdigraphs and semikernels $\langle N_i, S_i\rangle_{1\leq i\leq n}$ such that:*

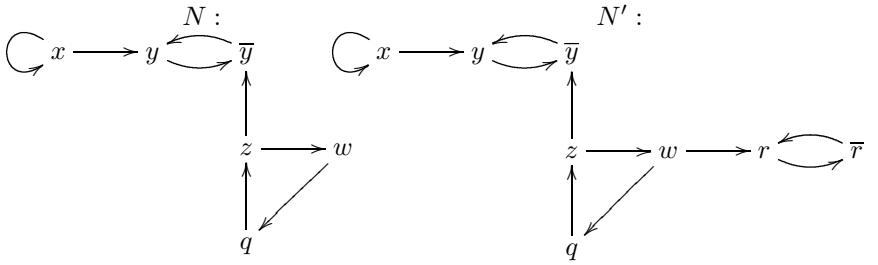
1. $N_1 = N$
2. S_i *is a semikernel in* N_i *for all* $1 \leq i \leq n-1$
3. $N_{i+1} = N_i \setminus (S_i \cup N^-(S_i))$ *for all* $1 \leq i \leq n-1$
4. S_n *is a kernel of* N_n.

Solvers are useful because of the following result.

Theorem 2.12 ([17]) *A digraph has a kernel iff it has a solver.*

Using solvers, it is possible to show that a range of various conditions is sufficient to ensure the existence of kernels in digraphs that are not kernel perfect. The conditions are rather complicated to state, so we omit them here. However, we think this work deserves to be mentioned because it identifies new heuristics we can follow when attempting to map the logical consequences of truth-theories.

As an example, consider the following two digraphs below:



In both N and N', there are two odd cycles: (x, x) and (z, w, q, z). It is tempting to look at $y, \overline{y}$ for a possible resolution. There is a problem, however, namely that they can only solve one of the sequences in question. In both N and N', we have semikernels $\{y\}$ and $\{\overline{y}, w\}$, corresponding to whether we use them to solve (x, x), or use them to solve (z, w, q, z). In N, this is where the story ends – it is not possible to resolve both, and we conclude that $Kr(N) = \emptyset$. In N', on the other hand, it *is* possible to solve both, but only if you solve (x, x) first by choosing y. This, in particular, no longer precludes solving (z, w, q, z), since it is possible to choose r and obtain the kernel $\{y, r, z\}$, as predicted also by Theorem 2.6 from [17].

Examples such as these show how sufficient conditions for existence of kernels can be interpreted as describing circumstances under which the truth of some statements can

lead to the resolution of problematic referential patterns. It is tempting, in particular, to think that y and r must be regarded as true *because* they resolve odd cycles. It seems necessary to accept their truth not because they cannot be refuted, but because accepting them is needed in order to resolve problems with (implicit) self-negation affecting other parts of the network. A basic intuition in much work on truth has been that semantic judgments should conform to classical logic to the greatest possible extent.

This involves accepting that some statements are true not because of what they say about the world, but because of what other statements say about them. Such statements are different from both truth-tellers and liars. They are not paradoxes and they are not undetermined. Rather, they must be assigned a unique value to resolve referential patterns that would otherwise become problematic. For instance, consider the following sentence A: "this sentence and the truth-teller B are both false". If B is a standard truth teller, stating "this sentence is true", it seems that B *must* be regarded as true in this referential network, since otherwise A becomes paradoxical. If we are committed to the idea that truth satisfies the property that paradox is avoided whenever possible, it seems to follow that our conclusion that B *must* be true is sufficient, in such a case, to conclude also that it *is* true.

2.3 Argumentation

The desire to arrive at some general notions of what counts as a logically correct argument seems to arise naturally in all human societies. If there is interaction there is argument, and some preliminary agreement on what is required for an argument to count as *successful* is of great importance, if nothing else then for pragmatic reasons.[8]

Following Frege and the formal turn in logic, the study of argumentation was largely seen as distinct from the formal study of correct reasoning. At best, it belonged to the informal branch. The search for logical perfection would famously flounder over results on incompleteness and undecidability, however, and since then the trend has been turning. Following the increasing popularity of non-classical logics, in particular defeasible logics [36,31], argumentation and logic have moved closer to each other.

This development took a particularly interesting turn with the seminal work of [16], who established a nice formal connection between argumentation on the one hand and non-monotonic reasoning and logic programming on the other.[9] Since then, abstract argumentation has attracted much attention, particularly in the AI-community [35]. The theoretical part of this work centers around the following question: Given some collection of arguments and some model of their content, how do we judge which arguments we should accept?

The novel move made in [16] was to rely on directed graphs as models, often referred to as argumentation frameworks (AFs) in this context. In argumentation theory, the

[8] In recent work from cognitive science it is even suggested that human reasoning may have evolved primarily because it proved useful in the context of argumentation [27].

[9] We also mention [12], a less cited work that did not involve the concept of argumentation, but which nevertheless has close connections to Dung's work. In particular, this work was the first, of which we are aware, to observe the close connection between kernel theory, logic programming and default logic.

atoms Π are thought of as arguments, and edges are thought of as attacks between them, such that e.g., $(p, q) \in N$ expresses that p is attacking q. By using digraphs to model the content of arguments, it becomes possible to give a range of argumentation semantics using intuitive graph constructions.

Given an AF N the task of such a semantics is to identify sets of arguments that can be held successfully together, typically called *extensions* in the literature. Most semantics are based on the intuition that a set of arguments should be internally consistent and able to defend itself against attack from other arguments. Different semantics differ about the details, but they all share the same overall aim: they give an answer, for any $p \in \Pi$, whether p should be accepted in the argumentative scenario represented by N. In particular, they all have the same signature, they are defined as an operator ϵ which takes an N and returns a set of sets $\epsilon(N) \subseteq 2^{\Pi}$.

To the best of our knowledge, all the semantics that have been studied share the property that arguments in an acceptable set should be free of internal conflict. Formally, for all semantics ϵ, all AFs N and all $A \in \epsilon(N)$, we have $N^-(A) \subseteq \Pi \setminus A$: no two arguments in A attack each other.

At first sight it seems we are working with a binary notion of acceptance: for a given argument, it is accepted or it is not. However, a moment's thought will show that this perspective fails to do justice to the nature of the structure (N, ϵ) in two important ways. First, there is the question of whether it is correct to say that p is accepted on N under ϵ when there *exists* some $A \in \epsilon(N)$ such that $p \in A$, or whether we should require $p \in A$ for *all* such A. Both notions of acceptance have been studied, and the former is typically dubbed *credulous* acceptance while the latter is referred to as *skeptical*.[10]

The second sense in which acceptance is not a binary notion has to do with the structure of N. In particular, given any $A \in \epsilon(N)$ the status of p with respect to A can be any of the following:

$$\begin{array}{ll} 1 : p \in A & 2 : p \in N(A) \\ 3 : p \in \Pi \setminus (A \cup N(A)) & \end{array} \tag{2.13}$$

Notice that by conflict-freeness of A, it follows that if $p \in N(A)$ then $p \notin A$. Hence when the focus is on the status of individual arguments, we might as well view $\epsilon(E)$ as a set of partitions of Π into three disjoint sets or, equivalently, as a collection of so called *(Caminada) labellings*, functions $\mathsf{c} : \Pi \to \{1, 0, \frac{1}{2}\}$ such that for all $x \in \Pi$:

$$\mathsf{c}(x) = 0 \Leftrightarrow \exists y \in N^-(x) : \mathsf{c}(y) = 1 \tag{2.14}$$

For any AF E we let $\mathsf{cf}(N)$ be the set of all labellings for E, and we define $\mathsf{c}^1 = \{x \in \Pi \mid \mathsf{c}(x) = 1\}, \mathsf{c}^0 = \{x \mid \mathsf{c}(x) = 0\}$ and $\mathsf{c}^{\frac{1}{2}} = \{x \in \Pi \mid \mathsf{c}(x) = \frac{1}{2}\}$. This defines a semantics for argumentation such that for all N, we regard $A \subseteq \Pi$ as acceptable if there is some $\mathsf{c} \in \mathsf{cf}(N)$ such that $\mathsf{c}^1 = A$.[11] In applications of argumentation theory,

[10] See [35, p. 32]. The terminology goes back to Dung [16], who in turn borrowed it from non-monotonic logic, where it is used to describe two notions of entailment, corresponding to existential and universal quantification over the possible extensions of a theory, see e.g., [22, p. 398].

[11] Hence it is not hard to see that values assigned by labellings correspond to the three points of (2.13) whenever we restrict attention to conflict-free sets of accepted arguments. Notice, in particular, that $p \in \mathsf{c}^0 \Leftrightarrow p \in N^+(\mathsf{c}^1)$ and $p \in \mathsf{c}^{\frac{1}{2}} \Leftrightarrow p \in \Pi \setminus (\mathsf{c}^1 \cup \mathsf{c}^0)$.

this is usually considered too permissive, and a range of various restrictions has been considered, each giving rise to a new semantics, the most well-known of which are defined in Fig. 1.

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

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

First, the *admissible* semantics [16] is obtained by restricting attention to conflict-free labellings c for which all those arguments that attack c^1 are in turn attacked by c^1. Hence the semantics captures the intuition that a set of acceptable arguments should be able to defend itself against attacks. The *complete* semantics [16] adds a further restriction, which captures the intuition that all arguments that are not disputed should be accepted. Hence, in addition to conflict-freeness it is also required that c^1 is equal to the set of those arguments that it defends. The *grounded* semantics [16] encodes a skeptical attitude, since it prescribes a unique labeling, namely the smallest complete labeling. This labeling always exists and is computable in linear time, starting from the labeling where all arguments are assigned $\frac{1}{2}$ and then iteratively labeling arguments by the boolean values, starting with those that are not attacked by any argument. The least fixed point of such a process will be the set of acceptable arguments under the grounded semantics, as explained in [16] and [8] for the labeling formulation.

The *preferred*, *semi-stable* and *stable* semantics all capture variants of the intuition that labellings should not only be admissible, but also allow us to reach a definite conclusion about the status of as many arguments as possible. According to the preferred semantics, which was first defined in terms of extensions rather than labellings [16], this amounts to maximizing the number of accepted arguments. According to the semi-stable semantics [10], it amounts to maximizing the number of boolean-valued arguments, while according to the stable semantics it amounts to requiring that no argument whatsoever is assigned the value $\frac{1}{2}$. This, however, is sometimes impossible, making the stable semantics the only one that sometimes fails to produce a labeling. This happens, for instance, on the AF $\{(x, x)\}$, corresponding to the liar statement, where all the other semantics admit $\{(x, \frac{1}{2})\}$ as the only permissible labeling.

The semantics are all defined as labellings, but (2.14) establishes an obvious one-to-one correspondence between a set of labellings and a set of extensions (sets of arguments assigned 1). Hence in the following we will allow ourselves to switch freely between these two representations, without introducing redundant notation to distinguish them.

In Fig. 2 we give two AFs, F and F', that serve as examples. In F, every argument is attacked by some argument, and from this it follows that we have $g(\mathsf{F}) = \emptyset$, i.e., the

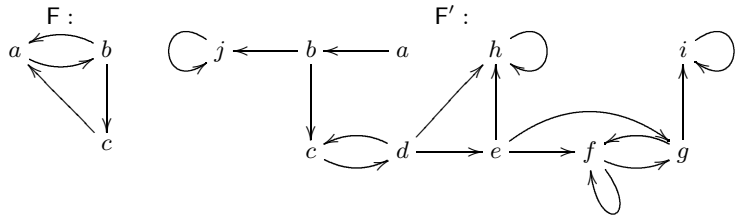


Fig. 2. Two argumentation frameworks

grounded extension is the empty set. The non-empty conflict-free sets are the singletons $\{a\}, \{b\}$ and $\{c\}$, but we observe that a does not defend itself against the attack it receives from c (since there is no attack (a, c)), and that c does not defend itself against the attack it receives from b. So the only possible non-empty admissible set is $\{b\}$. It is indeed admissible; b is attacked only by a and it defends itself, attacking a in return. In fact, since b also attacks c, the set $\{b\}$ is the unique stable set of this framework. It follows that $s(\mathsf{F}) = p(\mathsf{F}) = ss(\mathsf{F}) = \{\{b\}\}$ and $a(\mathsf{F}) = c(\mathsf{F}) = \{\emptyset, \{b\}\}$.

For a more subtle example, consider F'. The first thing to notice here is that we have an unattacked argument a, so the grounded extension is non-empty. In fact, the framework is such that all semantics from Figure 1 behave differently. It might look a bit unruly, but there are many self-attacking arguments that can be ruled out immediately (since they are not in any conflict-free sets), and it is easy to verify that the extensions of F' under the different semantics are the following:

$$g(\mathsf{F}') = \{a\}, s(\mathsf{F}') = \emptyset, ss(\mathsf{F}') = \{\{a, d, g\}\}$$
$$a(\mathsf{F}') = \{\emptyset, \{a\}, \{a, c\}, \{a, c, e\}, \{d\}, \{a, d\}, \{a, d, g\}, \{d, g\}\}$$
$$p(\mathsf{F}') = \{\{a, d, g\}, \{a, c, e\}\}, c(\mathsf{F}') = \{\{a\}, \{a, d, g\}, \{a, c, e\}, \{a, d\}\}$$

It is easy to see that the semantics for argumentation are closely connected to kernels and semikernels of digraphs. In particular, let $\overline{N} = \{(x, y) \mid (y, x) \in N\}$, so that $\overline{N}$ is the digraph obtained from N by reversing the direction of all edges. Then it is trivial to verify the following for all AFs N, $A \subseteq \mathit{\Pi}$.

$$A \in a(N) \Leftrightarrow A \in Lk(\overline{N}) \;\&\; A \in s(N) \Leftrightarrow A \in Kr(\overline{N}) \tag{2.15}$$

This connection was first observed in [13] but does not appear to have received much attention in the literature on argumentation. However, it follows from it that much work done in kernel theory, highly theoretical in nature, can be applied in argumentation theory. All the results mentioned in Sect. 3 detail circumstances when AFs admit non-empty stable sets, and the proofs are also mostly constructive, and identify scenarios where such sets can be computed quickly.[12] In particular, the connection to kernel theory gives us a taxonomy of different case types and different forms of inconsistency. We

[12] The decision problems in argumentation tend not to be tractable, and except for the grounded semantics, even computing the set of extensions is hard [35, Part I, Chap. 5]. Hence it is worth noting that many proofs from kernel theory provide computational information about *how* to argue in order to make sure that a given argument turns out to be accepted. For instance, the notion of a minimal semikernel, used in [17], can be understood in this way.

think combinatorial techniques developed in graph theory can be very helpful in future work that aims to shed light on the patterns underlying successful argumentation.

In the other direction, we note that while kernel theory can be understood as focusing on the question of classical consistency, corresponding to the existence of stable sets, argumentation theory has developed semantics which aim to facilitate reasoning about scenarios where classical consistency cannot be achieved. To assess these semantics from a philosophical perspective on truth, and a technical perspective on digraphs, seems like a very fruitful avenue for future research.

One crucial question concerns the logical foundations of these various semantics, and there has recently been quite some work devoted to this, most of which focuses on finding neat ways to *define* argumentation semantics, see [24,23] which relies on modal logic, and [2] which uses quantified boolean formulas.[13] While we think this work is interesting, we note that there has so far been a shortage of logics designed to permit reasoning *about* AFs, and to study meta-logical properties.

We can certainly attempt to use logics that are expressive enough to define various semantics in the object-language, but such an approach easily runs the risk of complicating matters to the extent that interesting results become hard or impossible to obtain. In particular, it will typically require us to use (fragments of) very powerful logics that may not admit any straightforward axiomatization, if they are decidable at all. In the next section we propose another route, focusing on extending the connection between propositional logic and semantics formulated on digraphs. In particular, we show that Łukasiewicz logic can be used to reason about AFs, and that a strong correspondence can be established for skeptical reasoning, whereby the validities of argumentation coincide for all non-stable semantics defined in Fig. 1. In particular, we show that they are all axiomatized by Wajsberg's rules for Łukasiewicz's three-valued logic.

3 Logics for Reasoning about Argumentation and Truth

In this section we will talk about digraphs using the following language $\mathcal{L}$:

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

where $p \in \Pi$. Since argumentation semantics are formulated in terms of three-valued labellings, we already have in place a corresponding interpretation of atomic formulas from $\mathcal{L}$. The semantic value of p, in particular, is one among $\{1, 0, \frac{1}{2}\}$. This is not a novel proposal, merely a logical reformulation of what is already commonplace in the literature on argumentation, see e.g., [35, Chapter 2]. However, we will now extend labellings inductively to provide a three-valued interpretation of the whole language $\mathcal{L}$. This involves a new construction, but it is easy enough to motivate once we consider the intended reading of formulas in $\mathcal{L}$.

We will think of $\mathcal{L}$ as containing meta-arguments addressing the semantic status that arguments *should* obtain in an AF. The connectives are read intuitively as follows: the

[13] For completeness, we also mention [19,9] which develops similar ideas by exploiting (other) ways to define argumentation semantics in modal logic, and [42], which relates argumentation to three-valued labellings for logic programing.

formula $\neg\phi$ is the argument that ϕ should be rejected, while the argument $\phi \to \psi$ is the argument that it should be at least as easy to accept ψ as it is to accept ϕ. On such a reading it seems clear that for all AFs N and all semantics ϵ, the following inductive definition of $\overline{\mathsf{c}} : \epsilon(N) \times \mathcal{L} \to \{1, 0, \frac{1}{2}\}$ appropriately extends any $\mathsf{c} \in \epsilon(N)$ to any $\phi \in \mathcal{L}$:

$$\overline{\mathsf{c}}(\phi) = \begin{cases} \mathsf{c}(\phi) \text{ if } \phi = p \in \mathit{\Pi} \\ 1 - \overline{\mathsf{c}}(\psi) \text{ if } \phi = \neg\psi \\ min\{1, 1 - (\overline{\mathsf{c}}(\psi_1) - \overline{\mathsf{c}}(\psi_2))\} \text{ if } \phi = \psi_1 \to \psi_2 \end{cases} \tag{3.1}$$

To illustrate the definition, assume we have a labeling $\mathsf{c} = \{p \mapsto 0, q \mapsto \frac{1}{2}\}$. In this case, it is intuitively clear that the argument that it should be at least as easy to accept q as it is to accept p is itself acceptable. This is also the outcome prescribed by (3.1), since $\overline{\mathsf{c}}(p \to q) = min\{1, 1 - (\mathsf{c}(p) - \mathsf{c}(q))\} = min\{1, 1.5\} = 1$. For a different example, suggesting also that formulas of $\mathcal{L}$ should not be read as stating that ϕ *is* accepted, consider the same meta-argument when the labeling is $\mathsf{c} = \{p \mapsto 1, q \mapsto \frac{1}{2}\}$. In this case, it seems clear that we cannot accept the argument that q is at least as easy to accept as p. However, since the status of q is undetermined, we cannot reject the argument that this *should* be the case. Hence it seems that the meta-argument itself should be regarded as undetermined, which, indeed, is what (3.1) ensures. To further illustrate that this analysis is appropriate, consider an AF N which admits *two* labellings, $\mathsf{c}_1 = \mathsf{c}$ and $\mathsf{c}_2 = \{p \mapsto 1, q \mapsto 1\}$. In this case, it is still not correct to say that $p \to q$ is acceptable on N, but it is also wrong to say that it has been rejected, since it only fails to be acceptable when q is undetermined, and is acceptable in all other cases. In fact, it seems that $p \to q$ not only should be accepted, but *must* be accepted, since neither of the two possible assignments entitle us to reject it.

This distinction introduces a modal flavor to $\mathcal{L}$, and in the list below we give some useful expressions along with three definable non-trivial modalities.[14]

- $\top := p \to p$ where $p \in \mathit{\Pi}$ is arbitrary.
- $\bot := \neg\top$.
- $\phi \vee \psi := (\phi \to \psi) \to \psi$.
- $\phi \wedge \psi := \neg(\neg\phi \vee \neg\psi)$.
- $\phi \leftrightarrow \psi := (\phi \to \psi) \wedge (\psi \to \phi)$.
- $\Box\phi := \neg(\phi \to \neg\phi)$ (meaning ϕ is accepted).
- $\Diamond\phi := \neg\phi \to \phi$ (meaning ϕ is not rejected).

[14] The observation that they are definable in terms of $\{\neg, \to\}$ was made by Tarski in 1921, who was Łukasiewicz's student at the time, see [26, p. 167]. Our preferred reading of these modalities is slightly non-standard. In particular, we will often think of truth normatively as providing permissions and/or obligations to accept claims as being true. According to the T-scheme, a permission to accept ϕ arises only when ϕ is the case, i.e., when it has the value 1. On the other hand, the T-scheme also implies that a permission to reject ϕ arises only when it is not the case, i.e., when ϕ has value 0. Moreover, we think of truth as prescribing the *norm* that a statement should be either accepted as true or rejected as false. The paradoxes show that this is sometimes impossible, and hence we think of the value $\frac{1}{2}$ as signifying that one has an obligation to accept (or reject), yet no permission to do so. Hence, our modal reading lets us think of semantic paradoxes as a form of normative conflict. In future work, we would like to explore this point of view further.

– $\triangle\phi := \phi \leftrightarrow \neg\phi$ (meaning ϕ is neither accepted nor rejected).

To better understand the behavior of the modal operators, consider the unpacking of the inductive definition of $\overline{\mathsf{c}}$ for these formulas, shown below and easily established against (3.1), for any labeling c.

$$\overline{\mathsf{c}}(\phi \vee \psi) = max\{\overline{\mathsf{c}}(\phi), \overline{\mathsf{c}}(\psi)\} \qquad \overline{\mathsf{c}}(\phi \wedge \psi) = min\{\overline{\mathsf{c}}(\phi), \overline{\mathsf{c}}(\psi)\}$$

$$\overline{\mathsf{c}}(\Box\phi) = \begin{cases} 1 \text{ if } \overline{\mathsf{c}}(\phi) = 1 \\ 0 \text{ otherwise} \end{cases} \qquad \overline{\mathsf{c}}(\Diamond\phi) = \begin{cases} 0 \text{ if } \overline{\mathsf{c}}(\phi) = 0 \\ 1 \text{ otherwise} \end{cases}$$

$$\overline{\mathsf{c}}(\triangle\phi) = \begin{cases} 1 \text{ if } \overline{\mathsf{c}}(\phi) = \frac{1}{2} \\ 0 \text{ otherwise} \end{cases}$$

Notice that all the modal expressions have the property that they evaluate to boolean values. Intuitively, this is reasonable: if someone says of some argument "that argument has been accepted" he is not reiterating it, but claiming that it has as a matter of fact been accepted. This, unlike the acceptability of the argument itself, seems natural to interpret in boolean terms.

With an extension of labellings to formulas in place, it is straightforward to associate a formal logic with every argumentation semantics. In particular, let $\mathcal{AF}$ denote the set of all AFs over Π. Then we define a class of argumentation logics as follows.

Definition 3.2 *For all argumentation semantics ϵ, we define $\models_\epsilon \subseteq \mathcal{AF} \times 2^{\mathcal{L}}$ such that for all $N \in \mathcal{AF}, \phi \in \mathcal{L}$:*

$$N \models_\epsilon \phi \textit{ if, and only if } \forall \mathsf{c} \in \epsilon(N) : \overline{\mathsf{c}}(\phi) = 1$$

We write $\models_\epsilon \phi$ just in case $N \models_\epsilon \phi$ for all *$N \in \mathcal{AF}$, in which case we say that ϕ is* valid *in the logic ϵ.*

Intuitively, we think of $N \models_\epsilon \phi$ as encoding that it is true that the meta-argument ϕ should be skeptically accepted on N, according to ϵ. To illustrate the behavior of some argumentation logics, consider the AF from Fig. 3.

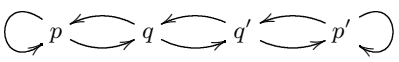


Fig. 3. An AF E such that $\Pi(E) = \{p, q, q', p'\}$

Below we list some truths about this N, under various logics corresponding to semantics from Figure 1.

$N \models_s \bot$ since $s(N) = \emptyset$
$N \models_g \triangle v$ for all $v \in \{p, q, q', p'\}$
$N \models_x \Box q \vee \Box\neg q$ for $x \in \{p, ss, s\}$
$N \not\models_x \Box q \vee \Box\neg q$ for $x \in \{a, c, s\}$
$N \models_x \triangle p \rightarrow \neg q$ for $x \in \{p, ss, s\}$
$N \models_x \triangle p \rightarrow \Diamond\neg q$ for $x \in \{a, c, g, p, ss, s\}$

The first point illustrates that the stable semantics is in a special position since it requires boolean labellings. In particular, for all AFs that do not admit such a labeling, skeptical reasoning gives rise to deductive explosion – *all* arguments are skeptically acceptable. This behavior is captured and extended by the corresponding logic, which judges every formula to be true on such AFs. Next, let us turn the last two points in the list. They express variants of the intuition that in the scenario described by N, it is acceptable to argue that it is as easy to reject q as it is to leave p undetermined. For the preferred and semi-stable semantics this is true since the only labeling which leaves p undetermined involves rejecting q. For the remaining non-stable semantics, it could be that *both* p and q are undetermined, meaning that rejecting q is harder than leaving p undetermined. However, the weaker form expressed in the last formula is true for all semantics.

Having formally defined logics based on argumentation semantics, we are ready to formally investigate the question of characterizing the validities of ϵ, the set of formulas ϕ such that $\models_\epsilon \phi$.

3.1 The Validities of Propositional Argumentation

Out of all the formalisms we consider in this paper, three-valued Łukasiewicz logic, Ł$_3$, has the longest history. It was introduced by the Polish logician Jan Łukasiewicz in the 1920s and is still studied both theoretically and from the point of view of applications. It is standardly defined for the language $\mathcal{L}$ and the semantics can be provided using three-valued functions $\rho : \Pi \to \{1, 0, \frac{1}{2}\}$, see e.g., [29]. These functions are extended to provide an interpretation for any $\phi \in \mathcal{L}$ in exactly the same way as detailed in (3.1), and the difference between Ł$_3$ and the argumentation logics arising from Definition 3.2 is that in Ł$_3$, a model is a single three-valued function ρ, not an AF which defines a *set* of such functions. Moreover, *any* three-valued function counts as a model, regardless of whether or not it is possible to induce it by an argumentation framework. Let $Ł = \{1, 0, \frac{1}{2}\}^{\Pi}$ denote all functions from Π to $\{1, 0, \frac{1}{2}\}$. Then we can give the following formal definition.

Definition 3.3 *The logic Ł$_3$ is defined as* $\models \subseteq 2^{\mathcal{L}} \times \mathcal{L}$ *such that for all* $\Phi \in 2^{\mathcal{L}}, \psi \in \mathcal{L}$

$$\Phi \models \psi \Leftrightarrow \forall \rho \in Ł : \big((\forall \phi \in \Phi : \overline{\rho}(\phi) = 1) \Rightarrow \overline{\rho}(\psi) = 1\big)$$

When $\Phi = \emptyset$ *we write* $\models \psi$ *and say that* ψ *is valid.*

The following deduction system is sound and complete for Ł$_3$, see e.g., [29]:

Axioms

1. $\phi \to (\psi \to \phi)$
2. $(\phi \to \psi) \to ((\phi \to \gamma) \to (\psi \to \gamma))$
3. $(\neg\psi \to \neg\phi) \to (\phi \to \psi)$
4. $((\phi \to \neg\phi) \to \phi) \to \phi$

Inference rule

– Modus ponens:

$$\frac{\phi \to \psi \qquad \phi}{\psi}\ \text{(MP)}$$

Given some set Φ, we let $\Phi \vdash \phi$ denote that ϕ can be derived in this reasoning system from the premises in $\Phi \subseteq \mathcal{L}$. In case Φ is empty we write simply $\vdash \phi$ and say that ϕ

is a theorem of Ł. Soundness and completeness of the system can then be expressed as follows (see [29] for a proof of general completeness for $Ł_3$).

$$\Phi \models \phi \Leftrightarrow \Phi \vdash \phi \tag{3.4}$$

Notice that the standard deduction theorem, $\phi \vdash \psi \Leftrightarrow \vdash \phi \rightarrow \psi$, fails for $Ł_3$. However, the following restricted version is easy to verify.

$$\phi \vdash \psi \Leftrightarrow \vdash \phi \rightarrow (\phi \rightarrow \psi) \tag{3.5}$$

We now show that all non-stable semantics from Fig. 1 give rise to the same validities as $Ł_3$. The most straightforward route to such a result would be to show, for each semantics, that every $\rho : \Pi \rightarrow \{1, 0, \frac{1}{2}\}$ is included in the set of labellings for *some* AF. This, however, does not hold. Consider, in particular, the assignment $\rho : \Pi \rightarrow \{1, 0, \frac{1}{2}\}$ defined by $\rho(p) = 0$ for all $p \in \Pi$. It is easy to see that it never obtains, for any of the semantics in Fig. 1. In particular, there can be no AF in which all arguments are rejected, since no argument would then be left to successfully attack them.

But for all non-stable semantics, it is not hard to show that there is an argumentation framework that induces it under this semantics. Then since all formulas from $\mathcal{L}$ contain only finitely many atoms, our result follows. For an arbitrary function $f : X \rightarrow Y$, let $f|_A = \{(x, y) \in f \mid x \in A\}$ denote its restriction to $A \subseteq X$. Then the sketch above can be formalized as follows.

Theorem 3.6 *For all semantics $\epsilon \in \{a, c, g, p, ss\}$ and all formulas $\phi \in \mathcal{L}$ we have*

$$\models \phi \Leftrightarrow \models_\epsilon \phi$$

Proof. ($\Rightarrow$) Follows trivially from Definition 3.2 since all labellings are three-valued assignments.
($\Leftarrow$) Let $\epsilon \in \{a, c, g, p, ss\}$ be arbitrary and assume $\models_\epsilon \phi$. By Definition 3.2 this means that for all AFs N and all $\mathsf{c} \in \epsilon(N)$ we have $\overline{\mathsf{c}}(\phi) = 1$. Let $\rho : \Pi \rightarrow \{1, 0, \frac{1}{2}\}$ be arbitrary. Then all we need to conclude the proof is to show that $\overline{\rho}(\phi) = 1$. Clearly, the value of $\overline{\rho}(\phi)$ only depends on $\rho|_{\Pi(\phi)}$ – the values assigned to arguments that appear in ϕ. Hence we are done if we can show that there is an AF N with some $\mathsf{c} \in \epsilon(N)$ for which $\mathsf{c}|_{\Pi(\phi)} = \rho|_{\Pi(\phi)}$, since then $\overline{\rho}(\phi) = \overline{\mathsf{c}}(\phi) = 1$ will follow from $\models_\epsilon \phi$. To construct such an AF, we let $r \in \Pi \setminus \Pi(\phi)$ be some argument not appearing in ϕ. Then the following AF will prove the claim, for any $\epsilon \in \{a, c, g, p, ss\}$:

$$N = \{(r, x) \mid x \in \Pi(\phi) \text{ and } \rho(x) = 0\} \cup$$
$$\{(x, x) \mid x \in \Pi(\phi) \text{ and } \rho(x) = \frac{1}{2}\}$$

It is easy to verify that the only non-empty labeling in $\epsilon(N)$ is c, defined for all $x \in \Pi$ as follows:

$$\mathsf{c}(x) = \begin{cases} \rho(x) \text{ if } x \in \Pi(\phi) \\ 1 \text{ otherwise} \end{cases}$$

Hence we obtain $\mathsf{c}|_{\Pi(\phi)} = \rho|_{\Pi(\phi)}$ as desired and this concludes the proof.

We obtain the following as a simple corollary.

Corollary 3.7 *For all semantics* $\epsilon \in \{a, c, g, p, ss\}$ *and all formulas* $\phi \in \mathcal{L}$*, we have*

$$\models_\epsilon \phi \Leftrightarrow \vdash \phi$$

For the stable semantics, it follows already from the correspondence between kernels and truth-theories (and the fact that the latter provide a normal form for propositional theory) that the stable validities are exactly those of propositional logic. Hence if we use $\models_b$ to denote logical consequence in classical propositional logic, we can complete the picture as follows.

Theorem 3.8 *For the stable semantics and all formulas* $\phi \in \mathcal{L}$*, we have*

$$\models_\epsilon \phi \Leftrightarrow \models_b \phi$$

We think that the axiomatizations provided here are important observations regarding the theoretical foundations of argumentation, and we also believe they can be useful in practical applications and further developments of argumentation theory. If we allow users of this theory to make use of Ł_3 in order to reason about AFs, it will permit them to make more subtle claims about their properties, allowing also the precise formal study of the acceptability of such meta-arguments. Indeed, we have identified a reasoning system for establishing validity of such arguments, allowing us to identify patterns of reasoning about AFs that can *always* be relied on. We remark that other reasoning systems have also been developed for Łukasiewicz logic, and these may be more efficient in practice than using Wajsberg's calculus, see [6].

Before we conclude, we consider the question of what happens when we interpret truth-theories using Łukasiewicz logic. Is the correspondence to AFs preserved? In [18] is was shown that complete labellings for AFs are three-valued models of the corresponding truth-theory and vice versa. This means that for the complete semantics we can use truth-theories to simulate the behavior of an AF, in place of the explicit encoding of the labeling as provided in the proof of Theorem 3.6. The advantage of doing this is that truth-theories corresponding to an AF can be computed quickly, in linear time by naive application of (2.1). Hence for the complete semantic it holds that the search for extensions in AFs is reducible in linear time to the problem of determining satisfiability of theories in Ł_3.

If we switch to classical logic, this gives us a linear time *equivalence*, since we can decide satisfiability of arbitrary propositional theories by studying the kernel problem in associated digraphs. This no longer holds for Ł_3, for any of the argumentations semantics from Fig. 1. To see this, note that *no* truth-theory is inconsistent in Ł_3. In particular, the grounded labeling (which is also complete), witnesses to this.[15] Hence the behavior or truth-theories under Ł_3 is fundamentally different from the behavior of such

[15] Also, the reader can easily verify that this assignment takes linear time to compute, by inductively inducing values from unattacked arguments and assigning $\frac{1}{2}$ to all remaining ones, as described, e.g., in [18].

theories under classical logic: truth in Ł$_3$ is a consistent notion, while in classical logic it is not.[16]

In the next section we consider modal reasoning about truth and argumentation, leading to the study of what propositions can rationally be believed on the basis of semantic information that it is possible to encode in a digraph.

4 Rational Belief on the Basis of Argument – A Modal Extension

The significance of our results so far is limited by the fact that we only cater to skeptical reasoning about AFs. A meta-argument is true if it holds for *all* acceptable labellings, and we lack the resources to express that a given argument can be credulously accepted (that there *exists* some acceptable labeling for which it is true).[17]

This is a shortcoming that we can address by modalizing our approach to skeptical reasoning, so that credulous reasoning arises as its dual. Notice that taking the truth-functional dual of Łukasiewicz logic, by letting $\frac{1}{2}$ count as a designated value, will not suffice.[18] Credulous acceptance of ϕ involves quantifying over all labellings under a given semantics, asking if ϕ evaluates to 1 in *one of them*. Hence, no truth-functional approach will give us what we want. However, if we think of labellings as possible worlds, we can capture credulous reasoning using a Kripkean approach.[19] In particular, we can associate to any AF a corresponding three-valued Kripke model, as follows:

Definition 4.1 *Given a semantics ϵ, an AF N and a set of states Q,*

- *An* evaluation frame *over Q is a function $V : Q \to \{1, 0, \frac{1}{2}\}^{\mathcal{I}}$, mapping states to labellings.*
- *For any evaluation frame V, the associated* Kripke model *is a tuple $\mathcal{M}(\epsilon, N, V) = (Q, V, R)$ where $R \subseteq Q \times Q$ such that for all $q, q' \in Q$:*

$$(q, q') \in R \Leftrightarrow V(q') \in \epsilon(N)$$

[16] We omit lengthy discussion of "revenge" issues, the worry that "stronger" paradoxes always tend to undermine attempts at regaining consistency in this way (for a collection on papers on revenge, see [4]). However, we mention that one strategy for countering revenge objections in the present context is to follow [25] who argued that the gap, the value $\frac{1}{2}$, should not to be seen as a semantic value at all, but merely as an expression of truth's partiality (so that, for instance, saying of a sentence in a gap that it is "not true" is akin to a category mistake, all the while truth as a concept does not apply to that sentence, i.e., it is like saying "the cheese is not true").

[17] In terms of truth, we are only able to address the truths that are *necessary* given the truth-theory; the mere *possible* truths, those that are contingent on the world beyond principle T, can not be talked about.

[18] Such a logic would bring us into paraconsistent territory, resulting in a system that stands to Łukasiewicz logic as Priest's LP stands to Kleene's three-valued logic [32].

[19] Importantly, we do not here ask for modal logics that encode the AF as such. This has been done already [24,9], resulting in logics where one talks directly about the structure of the digraph, rather than its meaning under a given semantics. What we want, rather, is to form three-valued meta-arguments that mix the credulous and skeptical modes of reasoning about AFs.

The definition builds models where all the states pointed to are required to correspond to acceptable labellings under an argumentation semantics. Intuitively, they are doxastic models such that the plausible states are taken to be those that cannot rationally be excluded on the basis of the argumentation semantics applied to the underlying AF. Indeed, notice that all relations R will automatically come out as both transitive and euclidian ($K45$ relations). Moreover, for non-stable ϵ we can use the axiom $\blacklozenge(p \to p)$ to restrict attention to serial relations, obtaining three-valued $KD45$ Kripke models.

In the present context, $\blacklozenge(p \to p)$ intuitively amounts to restricting attention to models where at least one state can be rationally entertained on the basis of the underlying AF. For the stable semantics this will lead us to discard some AFs, since some of them give rise to no rational beliefs under this semantics (only paradox). For non-stable ϵ, on the other hand, the fact that $\epsilon(N) \neq \emptyset$ ensures that $KD45$ models always exist.

To reason about three-valued Kripke models, we use a modal language with implication $\mathcal{L}_{\blacklozenge}$:

$$\phi ::= p \mid \neg\phi \mid \phi \to \phi \mid \blacklozenge\phi$$

where $p \in \Pi$. Truth can now be defined standardly, by first defining an appropriate three-valued evaluation of formulas at states. In particular, for all $M = \mathcal{M}(\epsilon, N, V)$ we define an associated three-valued labeling $\overline{M} : Q \times \mathcal{L}_{\blacklozenge} \to \{1, 0, \frac{1}{2}\}$ as follows:

$$\overline{M}(q,\phi) = \begin{cases} \epsilon(q)(\phi) \text{ if } \phi = p \in \Pi \\ 1 - \overline{M}(q,\psi) \text{ if } \phi = \neg\psi \\ min\{1, 1 - (\overline{M}(q,\psi_1) - \overline{M}(q,\psi_2))\} \text{ if } \phi = \psi_1 \to \psi_2 \\ \overline{M}(q,\phi) = max_{q' \in R(q)}\{\overline{M}(q',\psi)\} \text{ if } \phi = \blacklozenge\psi \end{cases} \tag{4.2}$$

Definition 4.3 *For all $M = \mathcal{M}(\epsilon, N, V)$ and all $q \in Q$, if $\overline{M}(q,\phi) = 1$, we write $M, q \models \phi$ and say that ϕ is true at q on M. If $M, q \models \phi$ for all $q \in Q$ we write $M \models \phi$ and say that ϕ is true on M, while if $\mathcal{M}(\epsilon, N, V) \models \phi$ for all N, V, we write $\models_\epsilon \phi$ and say that ϕ is valid under ϵ.*

For an example, consider again the AF N from Fig. 3. Assume someone claims the following: "if your beliefs are based on N it should be at least as hard to believe that you must reject an argument as it is to disbelieve that you should accept it". Quite a mouthful, but also meaningful, as it expresses absence of a certain kind of normative conflict about what meta-arguments to accept. In terms of $\mathcal{L}_{\blacklozenge}$, we can represent the claim as follows: $\blacklozenge(p \to p) \to (\blacksquare\Diamond\neg p \to \blacklozenge\neg p)$, for all $p \in \Pi$. It is not hard to see that it holds for N. This follows, in particular, from the fact that all atoms that can be assigned $\frac{1}{2}$ by an admissible labeling can be assigned 0 by some other such labeling. However, a stronger principle, making the same claim about *formulas* rather than atoms, fails on N. In particular, we do not have $\blacklozenge(p \to p) \to (\blacksquare\Diamond\phi \to \blacklozenge\phi)$ on any serial model induced by N. This is witnessed by the formula $\phi = \neg p \vee \neg p'$, since any admissible assignment must assign $\frac{1}{2}$ to at least one of p, p', meaning that while $\Diamond\phi$ evaluates to 1 in every state corresponding to an admissible labeling, the formula $\neg\phi$ evaluates to $\frac{1}{2}$ in all such states (hence is believed to be harder to accept).

In fact, we can prove a general result about this kind of normative conflict in argumentation assessment.

Proposition 4.4 *For all ϵ, N, V, we have that if N is finite then $\mathcal{M}(\epsilon, N, V) \models \blacksquare\Diamond\phi \rightarrow \blacklozenge\phi$ if, and only if, for all $q \in Q$, $R(q) \cap s(\epsilon) \neq \emptyset$*

Proof. To prove ($\Rightarrow$) we assume towards contradiction that $R(q)$ contains no q' such that $V(q')$ is boolean-valued (meaning, in particular, that no q' corresponds to a stable set in N). Since N is finite it follows that $\mathit{\Pi}(N)$ is finite as well. Hence the formula $\bigwedge_{p\in\mathit{\Pi}(N)}\{p \vee \neg p\}$ is in $\mathcal{L}_{\blacklozenge}$ and it evaluates to $\frac{1}{2}$ at all $q' \in R(q)$. It follows that $\blacksquare\Diamond \bigwedge_{p\in\mathit{\Pi}(N)}\{p \vee \neg p\}$ evaluates to 1 at q while $\blacklozenge \bigwedge_{p\in\mathit{\Pi}(N)}\{p \vee \neg p\}$ evaluates to $\frac{1}{2}$, contradicting the assumption that $\blacksquare\Diamond\phi \rightarrow \blacklozenge\phi$ is true at q for all ϕ. For ($\Leftarrow$), let $q' \in R(q)$ be such that $V(q')$ is a stable labeling for N. Notice first that $\blacksquare\Diamond\phi$ can not evaluate to $\frac{1}{2}$ at q since $\Diamond\phi$ is always boolean-valued. Moreover, the case when it evaluates to 0 is trivial. So assume it evaluates to 1. Then it follows that $\Diamond\phi$ evaluates to 1 at q'. Since $V(q')$ is boolean-valued, we conclude that ϕ evaluates to 1 as well, so q' witnesses to the fact that $\blacklozenge\phi$ evaluates to 1 at q.

This result is only an example of the potential for making interesting use of modal reasoning about argumentation semantics.[20] In future work we would like to explore characterizations such as these in more depth. However, the most obvious meta-logical question raised by Definition 4.1 is the question of finding a sound and complete reasoning system. This question can be approached by checking if every three-valued $KD45$ model admits a modally equivalent model that is induced by an AF. If this can be established, modulo some argumentation semantics, it follows that the validities of modal reasoning about AFs under this semantics coincide with those of regular three-valued $KD45$.

Preliminary work suggests that such a result holds for the preferred and semi-stable semantics for argumentation. It seems, in particular, that under preferred and semi-stable semantics we can induce any set of three-valued labellings with finite domain using an appropriately constructed AF. We plan to work out the implications of this for modal reasoning about AFs in a future paper, where we will also consider the matter of completeness and canonicity with respect to other argumentation semantics.

5 Conclusion

We started from the study of truth and went on to establish connections to argumentation and belief, through formal equivalences between models used to study these notions. The link to kernel theory and Łukasiewicz logic was emphasized, and we made use of the latter to provide axiomatizations of the skeptical validities arising from formal argumentation. We then proposed a modal extension, where credulous and skeptical forms of reasoning are captured as dual modalities. The semantics was provided by a special

[20] We remark that Proposition 4.4 does not hold for infinite AFs. Consider for instance the AF $N = \bigcup_{i\in\mathbb{N}}\{(p_i, p_i), (q_i, p_i), (r_i, q_i), (q_i, r_i), (z, q_i), (r_i, z)\}$. It is not hard to verify that for all finite subsets $P \subseteq \mathit{\Pi}$ there is an admissible labeling for N, $\mathtt{c}_P$, that is boolean-valued on P. Let Q be the set of all finite subsets of $\mathit{\Pi}$ and consider the Kripke model $M = \mathcal{M}(\epsilon, N, V)$ with V defined by $V(q) = \mathtt{c}_q$ for all $q \in Q$. It is easy to verify that $\blacksquare\Diamond\phi \rightarrow \blacklozenge\phi$ is true on M, even though N does not admit any stable labeling.

class of three-valued Kripke frames, those that can be induced from AFs using an argumentation semantics. We conjectured that the classes obtained under the preferred and semi-stable semantics are canonical for three-valued $KD45$ models.

In general, we think that the connections addressed in this paper can serve to motivate further work in all the fields we addressed. We think there is much to be gained from keeping formal links in mind, also if one feels that the underlying phenomena under consideration require different conceptual frameworks.[21] However, we think striking similarities at the formal level might also suggest deeper theoretical connections, and that this possibility should be explored further.

We are particularly keen on philosophical assessment of the formal link established between truth, argumentation and three-valued belief. It seems likely to us that it can inspire new philosophical ideas concerning the nature of these notions. To what extent are they mutually dependent? How are they related at a high level of abstraction? More concretely: Is it always possible for the truth to prevail in an argument? Should it be? Can false belief be distinguished from true belief on the basis of assessing arguments? Does this hold if "true" is replaced by "rational"? The formal connections mapped out in this paper naturally raise questions such as these, and we think they should be addressed. It seems, moreover, that we have identified a versatile formal framework for doing so.

References

1. Aczel, P.: Non-wellfounded sets. Technical Report 14, CSLI (1988)
2. Arieli, O., Caminada, M.W.A.: A QBF-based formalization of abstract argumentation semantics. Journal of Applied Logic 11(2), 229–252 (2013)
3. Barwise, J., Moss, L.: Vicious Circles: On the Mathematics of Non-Wellfounded Phenomena. CSLI, Stanford (1996)
4. Beall, J.C.: Revenge of the Liar: New Essays on the Paradox. Oxford University Press (2007)
5. Bezem, M., Grabmayer, C., Walicki, M.: Expressive power of digraph solvability. Ann. Pure Appl. Logic 163(3), 200–213 (2012)
6. Béziau, J.-Y.: A sequent calculus for Łukasiewicz's three-valued logic based on Suszko's bivalent semantics. Bulletin of the Section of Logic 28(2), 89–97 (1998)
7. Boros, E., Gurvich, V.: Perfect graphs, kernels and cooperative games. Discrete Mathematics 306, 2336–2354 (2006)
8. Caminada, M.: On the issue of reinstatement in argumentation. In: Fisher, M., van der Hoek, W., Konev, B., Lisitsa, A. (eds.) JELIA 2006. LNCS (LNAI), vol. 4160, pp. 111–123. Springer, Heidelberg (2006)
9. Caminada, M.W.A., Gabbay, D.M.: A logical account of formal argumentation. Studia Logica 93(2-3), 109–145 (2009)

[21] It is worth noting, for instance, that some results from formal argumentation (such as the existence of stable labellings for AFs without odd cycles [16]) are immediate corollaries of much older results from kernel theory ([37]). In the search for new results, it seems prudent to look across disciplinary boundaries to assess whether they have in fact already been established, or follow as trivial corollaries from existing theorems.

10. Caminada, M.W.A.: Semi-stable semantics. In: Proceedings of the 2006 Conference on Computational Models of Argument: Proceedings of COMMA 2006, pp. 121–130. IOS Press, Amsterdam (2006)
11. Cook, R.: Patterns of paradox. The Journal of Symbolic Logic 69(3), 767–774 (2004)
12. Dimopoulos, Y., Torres, A.: Graph theoretical structures in logic programs and default theories. Theoretical Computer Science 170(1-2), 209–244 (1996)
13. Doutre, S.: Autour de la sématique préférée des systèmes d'argumentation. PhD thesis, Université Paul Sabatier, Toulouse (2002)
14. Duchet, P.: Graphes noyau-parfaits, II. Annals of Discrete Mathematics 9, 93–101 (1980)
15. Duchet, P., Meyniel, H.: Une généralisation du théorème de Richardson sur l'existence de noyaux dans les graphes orientés. Discrete Mathematics 43(1), 21–27 (1983)
16. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence 77, 321–357 (1995)
17. Dyrkolbotn, S., Walicki, M.: Kernels in digraphs that are not kernel perfect. Discrete Mathematics 312(16), 2498–2505 (2012)
18. Dyrkolbotn, S., Walicki, M.: Propositional discourse logic. Synthese 191(5), 863–899 (2014)
19. Gabbay, D.: Modal provability foundations for argumentation networks. Studia Logica 93(2-3), 181–198 (2009)
20. Galeana-Sánchez, H., Guevara, M.-K.: Some sufficient conditions for the existence of kernels in infinite digraphs. Discrete Mathematics 309(11), 3680–3693 (2009); 7th International Colloquium on Graph Theory (ICGT) (2005)
21. Galeana-Sánchez, H., Neumann-Lara, V.: On kernels and semikernels of digraphs. Discrete Mathematics 48(1), 67–76 (1984)
22. Gottlob, G.: Complexity results for nonmonotonic logics. Journal of Logic and Computation 2(3), 397–425 (1992)
23. Grossi, D.: Argumentation in the view of modal logic. In: McBurney, P., Rahwan, I., Parsons, S. (eds.) ArgMAS 2010. LNCS (LNAI), vol. 6614, pp. 190–208. Springer, Heidelberg (2011)
24. Grossi, D.: On the logic of argumentation theory. In: van der Hoek, W., Kaminka, G.A., Lespérance, Y., Luck, M., Sen, S. (eds.) AAMAS, pp. 409–416. IFAAMAS (2010)
25. Kripke, S.: Outline of a theory of truth. The Journal of Philosophy 72(19), 690–716 (1975)
26. Łukasiewicz, J.: Selected works, edited by L. Borkowski. Studies in Logic and the Foundations of Mathematics. North Holland, Amsterdam (1970)
27. Mercier, H., Sperber, D.: Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences 34, 57–74 (2011)
28. Milner, E.C., Woodrow, R.E.: On directed graphs with an independent covering set. Graphs and Combinatorics 5, 363–369 (1989)
29. Minari, P.: A note on Łukasiewicz's three-valued logic. Annali del Dipartimento di Filosofia dell'Universitá di Firenze 8(1) (2002)
30. Neumann-Lara, V.: Seminúcleos de una digráfica. Technical report, Anales del Instituto de Matemáticas II, Universidad Nacional Autónoma México (1971)
31. Pollock, J.L.: How to reason defeasibly. Artif. Intell. 57(1), 1–42 (1992)
32. Priest, G.: The logic of paradox. Journal of Philosophical Logic 8, 219–241 (1979)
33. Quine, W.V.: The ways of paradox and other essays. Random House, New York (1966)
34. Rabern, L., Rabern, B., Macauley, M.: Dangerous reference graphs and semantic paradoxes. Journal of Philosophical Logic 42(5), 727–765 (2013)
35. Rahwan, I., Simari, G.R. (eds.): Argumentation in artificial intelligence. Springer (2009)
36. Reiter, R.: A logic for default reasoning. Artif. Intell. 13(1-2), 81–132 (1980)
37. Richardson, M.: On weakly ordered systems. Bulletin of the American Mathematical Society 52, 113–116 (1946)
38. Richardson, M.: Solutions of irreflexive relations. The Annals of Mathematics, Second Series 58(3), 573–590 (1953)

39. Tarski, A.: The concept of truth in formalised languages. In: Corcoran, J. (ed.) Logic, Semantics, Metamathematics, papers from 1923 to 1938, Hackett Publishing Company (1983) (translation of the Polish original from 1933)
40. von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press (1944, 1947)
41. Walicki, M., Dyrkolbotn, S.: Finding kernels or solving SAT. Journal of Discrete Algorithms 10, 146–164 (2012)
42. Wu, Y., Caminada, M.W.A., Gabbay, D.M.: Complete extensions in argumentation coincide with 3-valued stable models in logic programming. Studia Logica 93(2-3), 383–403 (2009)
43. Yablo, S.: Paradox without self-reference. Analysis 53(4), 251–252 (1993)