Explaining Argument Acceptance in ADFs

Tjitze Rienstra[1], Jesse Heyninck[2], Gabriele Kern-Isberner[3], Kenneth Skiba[4] and Matthias Thimm[4]

[1]*Maastricht University, Maastricht, The Netherlands*
[2]*Open Universiteit, Heerlen, The Netherlands*
[3]*Technische Universität Dortmund, Dortmund, Germany*
[4]*FernUniversität in Hagen, Germany*

Abstract

We present a dialogical proof theory for credulous acceptance in abstract dialectical frameworks under the preferred semantics. Our approach is motivated by the need to *explain* why an argument is accepted. The proof theory defines a set of rules for a dialogue between a proponent and opponent exchanging propositional formulas. The proponent takes on the role of trying to prove that the argument in question is acceptable, and the opponent takes on the role of exhaustively challenges the proponent's moves, with a dialogue where the proponent wins represents a proof that the argument in question is accepted.

Keywords

argumentation, explanation, abstract dialectical frameworks

1. Introduction

Formal argumentation is concerned with models of reasoning that mimic the mechanisms of human argumentation. Since explanation as a human activity relies on the same mechanisms, these models are considered as an ideal basis for explainable AI methods. Most influential within formal argumentation is Dung's model of *abstract argumentation*, which consists of two components [1]. One is an *argumentation framework*, an abstract representation of a debate consisting of a set of arguments and a binary relation of attack between arguments. Another component is the *argumentation semantics*, which represents a criterion to decide which arguments "win the debate", thereby determining the conclusions we can draw. One question left open in the picture sketched so far is how to *explain why an argument is accepted* under a given semantics. Here we can draw on proof theories that have been developed for abstract argumentation, which reflect the dialogical nature of argumentation. For example, if we use the preferred or grounded semantics, we can use the preferred or grounded *discussion games* [2, 3]. These are proof theories where a proof for the acceptance of an argument takes the form of a dialogue (an exchange of arguments) between an imaginary proponent and opponent. The rules defined by these methods determine

ArgXAI
t.rienstra@maastrichtuniversity.nl (T. Rienstra); jesse.heyninck@tu-dortmund.de (J. Heyninck); Gabriele.Kern-Isberner@cs.uni-dortmund.de (G. Kern-Isberner); kenneth.skiba@fernuni-hagen.de (K. Skiba); matthias.thimm@fernuni-hagen.de (M. Thimm)
0000-0002-0877-7063 (T. Rienstra); 0000-0002-3825-4052 (J. Heyninck); 0000-0001-8689-5391 (G. Kern-Isberner); 0000-0003-1250-8920 (K. Skiba); 0000-0002-8157-1053 (M. Thimm)

the permitted dialogue moves and ensure soundness and completeness, meaning that there exists a dialogue in which the proponent wins if and only if the argument of interest is acceptable under the preferred or grounded semantics. The idea to use dialogues of this kind for the purpose of explaining decisions to users, often in an interactive fashion, has appeared in a number of recent works [4, 5].

For some applications, Dung's model of abstract argumentation is not sufficiently expressive. This has led to a number of extensions of Dung's model. Examples are *bipolar argumentation* which allow both support and attack relations between arguments [6] and SETAFs, which allow the expression of collective attacks (i.e., sets of arguments jointly attacking another argument) [7]. The variety of extensions of Dung's model led to the development of a unifying approach based on the notion of *abstract dialectical frameworks* (ADFs) [8]. These consist of a set of arguments associated with *acceptance conditions*, which are propositional formulas with which we can express arbitrary relationships between arguments, including attack, support, and collective attack as in the earlier mentioned approaches. In this paper we take a step towards explaining why an argument is accepted in an ADF. We do this by developing a dialogical proof theory for credulous acceptance in ADFs under the preferred semantics. This proof theory can be used to explain argument acceptance in the presence of arbitrary relationships between arguments.

Our proposal is related to the dialogical proof theory for ADFs under the preferred semantics due to Zafarghandi et al. [9]. However, in their approach, the proponent and opponent exchange *interpretations*, a mathematical notion originating from the definition of the semantics of ADFs. These are functions that assign to each argument a truth value (true, false or undetermined). It is not clear how, in a setting where dialogues are used for explanatory purposes, such mathematical objects should be understood by a user. Our notion of dialogical proof is based on an exchange between a proponent and opponent of statements in the form of propositional formulas. Roughly speaking, the proponent takes on the role of trying to prove that the argument in question is acceptable, and the opponent takes on the role of exhaustively challenging the proponent's moves.

A dialogue that is won by the proponent is a dialogue in which the proponent has satisfied all of the challenges put forward by the opponent. Such a dialogue can be read as an explanation of the form ϕ_1 *because* ϕ_2 *because* ... *because* ϕ_n, where ϕ_1 is the initial claim that the argument of interest is accepted, and ϕ_n is a tautology. We believe that this approach is simpler, easier to interpret and therefore better suited to explain argument acceptance in ADFs.

Our proposal is based on a novel result relating the semantics of ADFs with the notion of *prime implicant*. Prime implicants are an important concept in applications such as model-based diagnosis [10], knowledge compilation [11], explaining decisions made by classifiers [12, 13, 14]. Briefly, a prime implicant of a formula ϕ is a minimal conjunction of literals that entails ϕ. We show that the admissible interpretation of an ADF are exactly those where an argument is true only if a prime implicant of its acceptance condition is true, and false only if a prime implicant of the negation of its acceptance condition is true. Prime implicants also play an important role in our dialogical proof theory, where the proponent replies to formulas put forward by the opponent by putting forward suitable prime implicants of these formulas.

The overview of this paper is as follows. In Section 2 we recall the necessary basics concerning ADFs. We introduce the concept of prime implicant, and establish its connection with the semantics of ADFs in Section 3. In Section 4 we present our proof theory for credulous acceptance under the preferred semantics of ADFs. In the remaining sections we discuss our findings, make

a comparison with related work and discuss future work.

2. Preliminaries

In this section we recall the necessary basics concerning ADFs and their semantics [8]. Given a set At of atoms we denote by $\mathscr{L}(At)$ the propositional language constructed from At and the logical connectives $\wedge$, $\vee$ and $\neg$ in the usual way. An ADF (abstract dialectical framework) is defined as follows.

Definition 1. *An* abstract dialectical framework *(in short, ADF) is a tuple $D = (At, L, C)$ where*

- *At is a finite set of* atoms *(also referred to as* arguments*)*
- *$L \subseteq At \times At$ is a set of* links
- *$C = \{\phi_x\}_{x \in At}$ is a set of* acceptance conditions *(elements of $\mathscr{L}(At)$), such that an atom $y \in At$ appears in ϕ_x only if $(y,x) \in L$.*

Given an ADF $D = (At, L, C)$ and argument $x \in At$ we denote by $par_D(x) = \{y \in At | (y,x) \in L\}$ the set of *parents* of x. As is common, if we omit the set L from the definition of an ADF then we assume L to be determined by the condition that $(x,y) \in L$ if and only if the atom x appears in ϕ_y.

An *interpretation* for $\mathscr{L}(At)$ is a function $v : At \to \{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$. We denote by $\mathscr{V}^3(At)$ the set of interpretations for $\mathscr{L}(At)$. Given an interpretation $v \in \mathscr{V}^3(At)$ and a set $B \subseteq At$, we use $v|_B$ to denote the restriction of v to B. Given a set $V \subseteq \mathscr{V}^3(At)$ we denote by $V|_B$ the set $\{v|_B | v \in V\}$.

An interpretation v is *two-valued* if $v(x) \in \{\mathbf{t}, \mathbf{f}\}$ for all $x \in At$. We denote by $\mathscr{V}(At)$ the set of two-valued interpretations for $\mathscr{L}(At)$. We extend an interpretation v to assign truth values to elements of $\mathscr{L}(At)$ using Kleene semantics [15]: $v(\neg\phi) = \mathbf{f}$ iff $v(\phi) = \mathbf{t}$, $v(\neg\phi) = \mathbf{t}$ iff $v(\phi) = \mathbf{f}$, and $v(\neg\phi) = \mathbf{u}$ iff $v(\phi) = \mathbf{u}$; $v(\phi \wedge \psi) = \mathbf{t}$ iff $v(\phi) = v(\psi) = \mathbf{t}$, $v(\phi \wedge \psi) = \mathbf{f}$ iff $v(\phi) = \mathbf{f}$ or $v(\psi) = \mathbf{f}$, and $v(\phi \wedge \psi) = \mathbf{u}$, otherwise; $v(\phi \vee \psi) = \mathbf{f}$ iff $v(\phi) = v(\psi) = \mathbf{f}$, $v(\phi \vee \psi) = \mathbf{t}$ iff $v(\phi) = \mathbf{t}$ or $v(\psi) = \mathbf{t}$, and $v(\phi \vee \psi) = \mathbf{u}$, otherwise. We say that v *satisfies* ϕ (written $v \models \phi$) if $v(\phi) = \mathbf{t}$.

The *information order* $\leq_i$ is the reflexive closure of the strict partial order $<_i$ over $\{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$ defined by $\mathbf{u} <_i \mathbf{t}$ and $\mathbf{u} <_i \mathbf{f}$. Intuitively, $\leq_i$ orders truth values according to their information content. We extend $\leq_i$ to an order over interpretations by setting $v \leq_i u$ if and only if $v(x) \leq_i u(x)$ for all $x \in At$.

We denote by $\sqcap$ the meet operation of the complete meet-semilattice $(\{\mathbf{t}, \mathbf{f}, \mathbf{u}\}, \leq_i)$. Intuitively, $\sqcap$ represents a *consensus* that assigns $\mathbf{t} \sqcap \mathbf{t} = \mathbf{t}$, $\mathbf{f} \sqcap \mathbf{f} = \mathbf{f}$ and assigns $\mathbf{u}$ in all other cases. The meet $\sqcap$ of two interpretations v, u of At is defined by $(v \sqcap u)(x) = v(x) \sqcap u(x)$ for all $x \in At$. The set of interpretations of a set At of atoms forms a complete meet-semilattice with respect to $\leq_i$.

A two-valued interpretation u *extends* a three-valued interpretation v iff $v \leq_i u$. We denote by $[v]_2$ the set of two-valued interpretations that extend v. Given an ADF D and three-valued-interpretation v, $\Gamma_D(v)$ denotes the three-valued interpretation defined by

$$\Gamma_D(v)(x) = \sqcap\{u(\phi_x) | u \in [v]_2\}.$$

The admissible, complete and preferred semantics are defined as follows.

Definition 2. *A three-valued interpretation v of an ADF D is:*

- *admissible iff* $v \leq_i \Gamma_D(v)$
- *complete iff* $v = \Gamma_D(v)$
- *preferred iff it is* $\leq_i$*-maximal admissible.*

We denote by $ad(D)$*,* $co(D)$ *and* $pr(D)$ *the set of admissible, complete and preferred interpretations of* D*.*

Given an ADF $D = (At, C)$ and argument $x \in At$ we say that x is *skeptically* accepted under the preferred semantics if $v(x) = \mathbf{t}$ for *every* preferred interpretation v of D, and that x is *credulously* accepted under the preferred semantics if $v(x) = \mathbf{t}$ for *some* preferred interpretation v of D. Note that, in what follows, we focus on explaining credulous acceptance under the preferred semantics, making use of the following fact: Since preferred interpretations are $\leq_i$-maximal admissible, proving credulous acceptance of an argument x under the preferred semantics amounts to checking whether there exists an admissible interpretation in which x is true.

3. Prime Implicants

In this section we introduce the notion of *prime implicant* [16] and establish its role in the context of the admissible semantics for ADFs. Prime implicants are an important concept in applications such as model-based diagnosis [10], knowledge compilation [11], and to explain decisions made by classifiers [12, 13, 14]. First some definitions. A *literal* is an atom or its negation. A *term* is a consistent conjunction of literals with $\top$ denoting the empty conjunction. We sometimes equate a term τ with the set of literals it contains. For instance, $\tau \setminus \tau'$ denotes the term τ with every literal that appears in τ' removed. An *implicant* of a formula ϕ is a term τ such that $\tau \models \phi$. A *prime implicant* of ϕ is an implicant τ of ϕ such that there exists no implicant τ' of ϕ such that $\tau' \subset \tau$. Thus a prime implicant can be thought of as a minimal assignment of truth values to atoms ensuring the truth of ϕ.

The following lemma establishes a characterisation of admissible interpretations in terms of prime implicants. This lemma forms the basis for the method presented in the next section. It states that the admissible interpretations are exactly those interpretations that satisfy the condition that an argument x is true only if a prime implicant of ϕ_x is true, and false only if a prime implicant of $\neg\phi_x$ is true.

Lemma 1. *An interpretation* v *of an ADF* $D = (At, C)$ *is admissible if and only if, for all* $x \in At$*:*

(1) *If* $v(x) = \mathbf{t}$ *then* $v(\tau) = \mathbf{t}$ *for some prime implicant* τ *of* ϕ_x*.*
(2) *If* $v(x) = \mathbf{f}$ *then* $v(\tau) = \mathbf{t}$ *for some prime implicant* τ *of* $\neg\phi_x$*.*

Proof. Let $D = (At, L, C)$ be an ADF.

For the *only-if* direction, suppose v is an admissible interpretation of D and let $x \in At$. For the case $v(x) = \mathbf{t}$, let τ be the conjunction of literals contained in the sets $\{y \mid y \in par_D(x), v(y) = \mathbf{t}\}$ and $\{\neg y \mid y \in par_D(x), v(y) = \mathbf{f}\}$. We will show that τ is an implicant of ϕ_x by showing that $w(\tau) = \mathbf{t}$ implies $w(\phi_x) = \mathbf{t}$ for every two-valued interpretation w of $\mathscr{L}(par_D(x))$. Let w be a two-valued interpretation of $\mathscr{L}(par_D(x))$ such that $w(\tau) = \mathbf{t}$. Then $w \in [v]_2|_{par_D(x)}$. Since v is admissible we have $v(x) \leq_i \sqcap\{u(\phi_x) | u \in [v]_2\}$, which implies that $u(\phi_x) = \mathbf{t}$ for every $u \in$

$[v]_2|_{par_D(x)}$. It thus follows that $w(\phi_x) = \mathbf{t}$. Hence, τ is an implicant of ϕ_x which in turn implies that $v(\tau') = \mathbf{t}$ for some prime implicant $\tau' \subseteq \tau$ of φ_x. Hence, condition (1) in Lemma 1 is satisfied. For the case $v(x) = \mathbf{f}$ it similarly follows that condition (2) in Lemma 1 is satisfied.

For the *if* direction, let v be an interpretation that, for all $x \in At$, satisfies conditions (1) and (2) in Lemma 1. We will prove that v is admissible. Let $x \in At$. We will prove that $v(x) \leq_i \Gamma_D(v)(x)$. There are two cases: $v(x) = \mathbf{t}$ and $v(x) = \mathbf{f}$. For the case $v(x) = \mathbf{t}$, let τ be the prime implicant of ϕ_x such that $v(\tau) = \mathbf{t}$ whose existence follows from condition (1) in Lemma 1. We will prove that for all $u \in [v]_2$, $u(\phi_x) = \mathbf{t}$. Let $u \in [v]_2$. Then $v \leq_i u$ and, since $v(\tau) = \mathbf{t}$, it follows that $u(\tau) = \mathbf{t}$ and hence $u(\phi_x) = \mathbf{t}$. It follows that $\Gamma_D(v)(x) = \mathbf{t}$ and hence $v(x) \leq_i \Gamma_D(v)(x)$. The case $v(x) = \mathbf{f}$ similarly implies, using condition (2) in Lemma 1, that $v(x) \leq_i \Gamma_D(v)(x)$. It thus follows that $v \leq_i \Gamma_D(v)$ and hence that v is admissible. □

Example 1. *Let D be an ADF with argument a. We consider three examples of acceptance conditions for a, along with the conditions that an admissible interpretation of D must satisfy for a to be accepted or rejected.*

1. *$\phi_a = \neg b \wedge \neg c$. Then ϕ_a has one prime implicant, namely ϕ_a itself. Accepting a therefore requires both b and c to be rejected. The prime implicants of $\neg\phi_a$ are b and c. Rejecting a therefore requires b or c to be accepted.*
2. *$\phi_a = b \wedge (c \vee d)$. Then the prime implicants of ϕ_a are $b \wedge c$ and $b \wedge d$. Accepting a thus requires b to be accepted as well as c or d. The prime implicants of $\neg\phi_a$ are $\neg b$ and $\neg c \wedge \neg d$. Rejecting a thus requires b to be rejected or both c and d to be rejected.*
3. *$\phi_a = b \vee \neg b$. Then the only prime implicant of ϕ_a is $\top$. We can therefore accept a regardless of the status of b. Rejecting a is impossible since $\neg\phi_a$ does not have a prime implicant.*

Note that case three in the example above demonstrates that Lemma 1 is not equivalent to the condition that x is true only if ϕ_x is true, and x is false only if $\neg\phi_x$ is true. To see why, note that we have $\phi_a = b \vee \neg b$ with ϕ_a having one prime implicant $\top$. Then, if $v(b) = \mathbf{u}$, we have that v satisfies a prime implicant of ϕ_a but it does not satisfy ϕ_a itself. Furthermore, note that we can adapt Lemma 1 to characterise complete interpretations by turning these only-if conditions into if-and-only-if conditions.

4. Dialogical Proofs for Credulous Acceptance under the Preferred Semantics

We now present a method to prove, given an ADF D and argument x, whether x is credulously accepted under the preferred semantics. We build on ideas behind the so called *preferred game* for abstract argumentation, a dialogical procedure that similarly determines whether an argument of an abstract argumentation framework is credulously accepted under the preferred semantics [17]. Let us sketch the idea behind the preferred game: two imaginary players (the *proponent* and *opponent*) take alternating turns in putting forward arguments according to a set of rules. The initial argument put forward by the proponent is the argument whose acceptance is determined. The opponent challenges the arguments put forward by the proponent by putting forward attacking arguments, and subsequent proponent moves represent defences against the opponent's attacks.

Credulous acceptance is proven if the proponent can win the game by ending the dialogue in its favour according to a "last-word" principle. Our method is a similar dialogical procedure based on a game between a proponent and opponent, formalised as follows.

Definition 3. *Let $D = (At, L, C)$ be an ADF and $x \in At$ be an argument. A* dialogical proof for x *is a sequence $(p_1, o_1, \ldots, p_n, o_n)$ of formulas (with p_i called the i-th* proponent move *and o_i the i-th* opponent move*) such that:*

1. $p_1 = x$
2. *For $i \geq 1$, $o_i = \Theta_D(p_i \setminus p_1 \cup \ldots \cup p_{i-1})$, where $\Theta_D(\phi)$ denotes the result of replacing every atom in ϕ with its acceptance condition.*
3. *For $i > 1$, p_i is a prime implicant of o_{i-1}*
4. *$p_1 \wedge \ldots \wedge p_n$ is satisfiable.*

We say that the dialogical proof is successful *if $o_n \equiv \top$.*

Explanation: the initial formula put forward by the proponent is the argument x whose acceptance is being proven (condition 1). The opponent replies to every proponent move by putting forward a challenge in the form of a formula that must hold for the proponent's claim to be hold (condition 2). This formula is constructed by taking the preceding proponent move (which is a conjunction of literals) and then removing the literals that have already been challenged in previous moves, and then replacing every atom with its acceptance condition. Subsequent proponent moves represent reasons for why these conditions are true, in the form of prime implicants (condition 3). The proponent's moves must furthermore be jointly consistent (condition 4). Credulous acceptance is proven if the dialogue is *successful*, meaning that the opponent has no other challenge left but a vacuous one represented by a tautology. We can think of this as the opponent "conceding" and the proponent "winning the discussion".

Note that the opponent's moves are fully determined by the preceding proponent moves, whereas the proponent may need to make a choice among possible prime implicants, where some choices may lead to a successful dialogical proof and others may not. Checking whether a successful dialogical proof exists thus amounts to finding a sequence of prime implicants leading to a dialogical proof that is successful. Furthermore, whereas the moves of the opponent are arbitrary formulas, the moves of the proponent, being either the initial argument or a prime implicant of the preceding opponent move, are conjunctions of literals. This means that the consistency requirement (condition 4) amounts to checking for the presence of contradictory literals.

The following theorem establishes soundness and completeness of our method.

Theorem 1. *Let $D = (At, L, C)$ be an ADF and $x \in A$ be an argument. There exists a successful dialogical proof for x if and only if there exists a preferred interpretation v of D such that $v(x) = \mathbf{t}$.*

Proof. Let $D = (At, L, C)$ be an ADF and $x \in At$ an argument.

ONLY IF: Let $(p_1, o_1, \ldots, p_n, o_n)$ be a successful dialogical proof for x. Let v be the interpretation of D defined by

$$v(x) = \begin{cases} \mathbf{t} & \text{if } x \in p_1 \cup \ldots \cup p_n \\ \mathbf{f} & \text{if } \neg x \in p_1 \cup \ldots \cup p_n \\ \mathbf{u} & \text{otherwise} \end{cases}$$

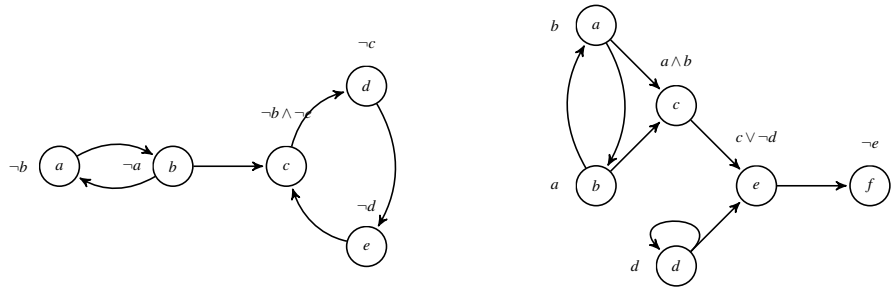

Figure 1: Two example ADFs

We will prove that v is admissible. Let $y \in At$. If $v(y) = \mathbf{t}$ then the construction of v implies that there is an i such that $y \in p_i$. It then follows that $o_i \models \phi_y$. Because p_{i+1} is a prime implicant of o_i, it follows that $p_{i+1} \models \phi_y$. Then there must be a term $\tau \subseteq p_{i+1}$ that is a prime implicant of ϕ_y. We furthermore have that $v(p_{i+1}) = \mathbf{t}$ and hence $v(\tau) = \mathbf{t}$. Thus, condition (1) in Lemma 1 is satisfied. If $v(y) = \mathbf{f}$ it follows similarly that condition (2) in Lemma 1 is satisfied. Using Lemma 1 it follows that v is an admissible interpretation of D such that $v(x) = \mathbf{t}$. This implies that there is a preferred interpretation w of D such that $v \leq_i w$ and hence $w(x) = \mathbf{t}$.

IF: Let v be a preferred interpretation of D such that $v(x) = \mathbf{t}$. We say that a sequence $(p_1, o_1, \ldots, p_n, o_n)$ is *v-valid* if for all i in $[1, \ldots, n]$, $v(p_i) = v(o_i) = \mathbf{t}$. We inductively and non-deterministically define a v-valid sequence $A(n)$ for any positive integer n as follows:

- $A(1) = (x, \Theta_D(x))$.
- If $A(n) = (p_1, o_1, \ldots, p_n, o_n)$ then $A(n+1) = (p_1, o_1, \ldots, p_n, o_n, p_{n+1}, o_{n+1})$ where
  1. p_{n+1} is a prime implicant of o_n such that $v(p_{i+1}) = \mathbf{t}$.
  2. $o_{n+1} = \Theta_D(p_{n+1} \setminus p_1 \cup \ldots \cup p_n)$.

Note that $A(1)$ is clearly v-valid. Furthermore, if $A(n) = (p_1, o_1, \ldots, p_n, o_n)$ is v-valid then $v(o_n) = \mathbf{t}$ which implies, using Lemma 1 and the fact that v is admissible, that a prime implicant p_{n+1} such that $v(p_{i+1}) = \mathbf{t}$ as mentioned in line 1 above exists, and hence that $A(n+1)$ exists. Furthermore, if $v(p_{i+1}) = \mathbf{t}$ it follows that $v(o_{i+1}) = \mathbf{t}$ and therefore $A(n+1)$ is also v-valid. Finiteness of At implies that there is a k such that $A(k) = A(k+1)$. It then follows that $A(k)$ is a successful dialogical proof for x. □

Example 2. *Consider the ADF shown in Figure 1 on the left. Note that the acceptance conditions of this ADF correspond to the acceptance conditions of an abstract argumentation framework. The ADF therefore represents the abstract argumentation framework with edges interpreted as attacks. The ADF has two preferred interpretations:* $v_1 = \{a = \boldsymbol{t}, b = \boldsymbol{f}, c = \boldsymbol{u}, d = \boldsymbol{u}, e = \boldsymbol{u}\}$ *and* $v_2 = \{a = \boldsymbol{f}, b = \boldsymbol{t}, c = \boldsymbol{f}, d = \boldsymbol{t}, e = \boldsymbol{f}\}$. *Thus, two arguments that are credulously accepted are a (interpretation* v_1*) and d (interpretation* v_2*). The successful dialogical proof for a is*

p_1	o_1	p_2	o_2	p_3	o_3
a	$\neg b$	$\neg b$	a	a	$\top$

(1)

The successful dialogical proof for d is

p_1	o_1	p_2	o_2	p_3	o_3	p_4	o_4	p_5	o_5
d	$\neg c$	$\neg c$	$b \vee e$	b	$\neg a$	$\neg a$	b	b	$\top$

(2)

Example 3. *The ADF shown in Figure 1 on the right has four preferred interpretations:*

$$v_1 = \{a = \boldsymbol{f}, b = \boldsymbol{f}, c = \boldsymbol{f}, d = \boldsymbol{f}, e = \boldsymbol{t}, f = \boldsymbol{f}\}$$
$$v_2 = \{a = \boldsymbol{t}, b = \boldsymbol{t}, c = \boldsymbol{t}, d = \boldsymbol{f}, e = \boldsymbol{t}, f = \boldsymbol{f}\}$$
$$v_3 = \{a = \boldsymbol{t}, b = \boldsymbol{t}, c = \boldsymbol{t}, d = \boldsymbol{t}, e = \boldsymbol{t}, f = \boldsymbol{f}\}$$
$$v_4 = \{a = \boldsymbol{f}, b = \boldsymbol{f}, c = \boldsymbol{f}, d = \boldsymbol{t}, e = \boldsymbol{f}, f = \boldsymbol{t}\}$$

Thus, two arguments that are credulously accepted are e and f. There are two successful dialogical proofs for e. They correspond to the two ways to satisfy the acceptance condition for e namely by making d false (preferred interpretations v_1 and v_2) or making c true (preferred interpretations v_2 and v_3):

p_1	o_1	p_2	o_2	p_3	o_4
e	$c \vee \neg d$	$\neg d$	$\neg d$	$\neg d$	$\top$

(3)

p_1	o_1	p_2	o_2	p_3	o_3	p_4	o_4
e	$c \vee \neg d$	c	$a \wedge b$	$a \wedge b$	$b \wedge a$	$b \wedge a$	$\top$

(4)

There are two distinct successful dialogical proofs for f, both corresponding to preferred interpretation v_4:

p_1	o_1	p_2	o_2	p_3	o_3	p_4	o_4	p_5	o_5	p_6	o_6
f	$\neg e$	$\neg e$	$\neg(c \vee \neg d)$	$\neg c \wedge d$	$\neg(a \wedge b) \wedge d$	$\neg a \wedge d$	$\neg b$	$\neg b$	$\neg a$	$\neg a$	$\top$

(5)

p_1	o_1	p_2	o_2	p_3	o_3	p_4	o_4	p_5	o_5	p_6	o_6
f	$\neg e$	$\neg e$	$\neg(c \vee \neg d)$	$\neg c \wedge d$	$\neg(a \wedge b) \wedge d$	$\neg b \wedge d$	$\neg a$	$\neg a$	$\neg b$	$\neg b$	$\top$

(6)

5. Discussion

We now have a sound and complete dialogical proof method for credulous acceptance in ADFs under the preferred semantics. The approach we took is motivated by the need to explain argument acceptance in ADFs. An important question is therefore whether dialogical proofs can indeed be used for this purpose and, if not, whether they can be transformed into adequate explanations. We plan to address this question more thoroughly in future work and will suffice here with some brief remarks. Firstly, we would like to interpret a dialogical proof as a sequence of statements connected by a *because* relationship. One issue is that of circular justifications. Take dialogical proof (5) as an example. This proof ends with the moves $\neg b, \neg b, \neg a, \neg a$ and $\top$. What happens here is that $\neg b$ is justified by $\neg a$ which is in turn, in a circular fashion, justified by $\neg b$. This circular justification is not made explicit in the proof, however, and finding a way to do so may be necessary to obtain adequate explanations.

Let us now consider another issue, namely that of redundancy due to repetition of formulas. The problem is that we do not want to explain a formula ϕ by stating that "ϕ because ϕ". Take again proof (5) as an example. Here we see that every proponent move apart from p_4 is logically equivalent to the preceding opponent move. The reason for this is that these opponent moves are formulas with only one prime implicant, namely the formula itself. Let us refer to such formulas as *deterministic*. Let us furthermore refer to a dialogical proof in which all deterministic proponent moves are removed, as *type 1* explanations. The type 1 explanation corresponding to the dialogical proof (5) is the following sequence (we use $\Longrightarrow$ to denote the "because" relationship).

$$f \Longrightarrow \neg e \Longrightarrow \neg(c \vee \neg d) \Longrightarrow \neg(a \wedge b) \wedge d \Longrightarrow \neg a \wedge d \Longrightarrow \neg b \Longrightarrow \neg a \Longrightarrow \top$$

Another possibility is to simply remove all opponent moves. Let us refer to this as a *type 2* explanation. For the dialogical proof (5) we get the following type 2 explanation:

$$f \Longrightarrow \neg e \Longrightarrow \neg c \wedge d \Longrightarrow \neg a \wedge d \Longrightarrow \neg b \Longrightarrow \neg a$$

We invite the reader to reflect on whether these two explanations adequately explain acceptance of f in the ADF shown in Figure 1 on the right.

6. Related Work

Zafarghandi et al. [9] were the first to propose a sound and complete discussion-based proof method for credulous acceptance under the preferred semantics of ADFs. Our approach is related but different in important respects. The main difference is that in their approach, the dialogue moves of the two players are interpretations of the ADF, rather than formulas. An insight linking the two approaches is that their notion of *minimal interpretation around an argument x* appears to correspond to a prime implicant of the acceptance condition of x. We believe that our approach, where moves are propositional formulas rather than interpretations, is simpler, easier to interpret and therefore better suited to explain argument acceptance in ADFs. Zafarghandi et al. also developed a discussion-based proof method for the grounded semantics of ADFs [18]. We expect to see a similar correspondence with that approach when we investigate dialogical proofs for grounded semantics in future work.

Our proof method can be seen as a generalisation of *Socratic discussion games* for credulous acceptance under the preferred semantics of abstract argumentation frameworks due to Caminada et al. [3]. They refer to their dialogues as Socratic because the role of the opponent in their discussion game is likened to that of Socrates in a Socratic discussion, while the proponent tries to avoid being led to a contradiction by the opponent. The roles of the proponent and opponent in our setting are not the same, however. To illustrate, consider again the ADF shown in Figure 1 on the left. As explained earlier, the acceptance conditions of this ADF correspond to those of an abstract argumentation framework. The ADF therefore represents the argumentation framework with edges interpreted as attacks. Below we show an example that we copied from [3] of a Socratic discussion about the argument d in this argumentation framework. This game is won by the proponent and the dialogue therefore proves credulous acceptance of d.

- PRO: $\texttt{IN}(d)$
  "I have an admissible labelling in which d is labelled IN."

- OPP: $\texttt{OUT}(c)$
  "But then in your labelling it must also be the case that d's attacker c is labelled OUT. Based on which grounds?"
- PRO: $\texttt{IN}(b)$
  "c is labelled OUT because b is labelled IN."
- OPP: $\texttt{OUT}(a)$
  "But then in your labelling it must also be the case that b's attacker a is labelled OUT. Based on which grounds?"
- PRO: $\texttt{IN}(b)$
  "a is labelled OUT because b is labelled IN."

If we translate $\texttt{IN}(x)$ to x and $\texttt{OUT}(x)$ to $\neg x$, then this dialogue consists of the following moves.

PRO	OPP	PRO	OPP	PRO
d	$\neg c$	b	$\neg a$	b

Three observations: Firstly, this sequence is equivalent to the type 2 explanation (i.e., the sequence resulting from removing all opponent moves) corresponding to the dialogical proof (2) for d, and we see the same equivalence in proofs for other ADFs that represent argumentation frameworks. This means that the roles of the proponent and opponent in a Socratic dialogue are merged into one role, which is in our setting played by the proponent. Secondly, in a Socratic dialogue, there is no analogue of the role played by the opponent in our setting. Consider again dialogical proof (2) as an example. When the proponent moves $\neg c$, the opponent replies with the exact condition that must be satisfied for $\neg c$ to be justified, which is $b \vee c$ (i.e., either b or c must be IN). This condition is not made explicit in the Socratic discussion shown above. Third, in a Socratic discussion game, the proponent may only claim that arguments are accepted, and the opponent may only claim that arguments are rejected. Such a restriction does not exist in our setting, where the proponent may claim that arguments are accepted as well as rejected.

7. Future Work

In future work we plan to further investigate the use of dialogical proofs for the purpose of explanation. We will also consider ways to present explanations to users in an interactive way. An interesting question in this context is how to incorporate user feedback in such an interactive setting in case of disagreement with an outcome or its explanation, which would lead to an update of the ADF. We also plan to define variants of our method for other semantics, such as the grounded semantics, which amounts to skeptical acceptance under the complete semantics. Finally, we plan to analyse the complexity of our method and to compare its runtime with existing ADF solvers.

References

[1] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artif. Intell. 77 (1995) 321–358.

URL: https://doi.org/10.1016/0004-3702(94)00041-X. doi:10.1016/0004-3702(94)00041-X.

[2] M. Caminada, A discussion game for grounded semantics, in: E. Black, S. Modgil, N. Oren (Eds.), Theory and Applications of Formal Argumentation - Third International Workshop, TAFA 2015, Buenos Aires, Argentina, July 25-26, 2015, Revised Selected Papers, volume 9524 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 59–73. URL: https://doi.org/10.1007/978-3-319-28460-6_4. doi:10.1007/978-3-319-28460-6_4.

[3] M. W. A. Caminada, W. Dvorák, S. Vesic, Preferred semantics as socratic discussion, J. Log. Comput. 26 (2016) 1257–1292. URL: https://doi.org/10.1093/logcom/exu005. doi:10.1093/logcom/exu005.

[4] K. Cyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 4392–4399. URL: https://doi.org/10.24963/ijcai.2021/600. doi:10.24963/ijcai.2021/600.

[5] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021).

[6] L. Amgoud, C. Cayrol, M. Lagasquie-Schiex, P. Livet, On bipolarity in argumentation frameworks, Int. J. Intell. Syst. 23 (2008) 1062–1093. URL: https://doi.org/10.1002/int.20307. doi:10.1002/int.20307.

[7] S. H. Nielsen, S. Parsons, A generalization of dung's abstract framework for argumentation: Arguing with sets of attacking arguments, in: N. Maudet, S. Parsons, I. Rahwan (Eds.), Argumentation in Multi-Agent Systems, Third International Workshop, ArgMAS 2006, Hakodate, Japan, May 8, 2006, Revised Selected and Invited Papers, volume 4766 of *Lecture Notes in Computer Science*, Springer, 2006, pp. 54–73. URL: https://doi.org/10.1007/978-3-540-75526-5_4. doi:10.1007/978-3-540-75526-5_4.

[8] G. Brewka, H. Strass, S. Ellmauthaler, J. P. Wallner, S. Woltran, Abstract dialectical frameworks revisited, in: F. Rossi (Ed.), IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013, IJCAI/AAAI, 2013, pp. 803–809. URL: http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6551.

[9] A. Keshavarzi Zafarghandi, R. Verbrugge, B. Verheij, Discussion games for preferred semantics of abstract dialectical frameworks, in: G. Kern-Isberner, Z. Ognjanovic (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 15th European Conference, ECSQARU 2019, Belgrade, Serbia, September 18-20, 2019, Proceedings, volume 11726 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 62–73. URL: https://doi.org/10.1007/978-3-030-29765-7_6. doi:10.1007/978-3-030-29765-7_6.

[10] J. de Kleer, A. K. Mackworth, R. Reiter, Characterizing diagnoses and systems, Artif. Intell. 56 (1992) 197–222. URL: https://doi.org/10.1016/0004-3702(92)90027-U. doi:10.1016/0004-3702(92)90027-U.

[11] A. Darwiche, P. Marquis, A knowledge compilation map, Journal of Artificial Intelligence Research 17 (2002) 229–264.

[12] A. Darwiche, A. Hirth, On the reasons behind decisions, in: G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, J. Lang (Eds.), ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela,

Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020), volume 325 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, pp. 712–720. URL: https://doi.org/10.3233/FAIA200158. doi:10.3233/FAIA200158.

[13] A. Ignatiev, N. Narodytska, J. Marques-Silva, Abduction-based explanations for machine learning models, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 1511–1519. URL: https://doi.org/10.1609/aaai.v33i01.33011511. doi:10.1609/aaai.v33i01.33011511.

[14] A. Shih, A. Choi, A. Darwiche, A symbolic approach to explaining bayesian network classifiers, in: J. Lang (Ed.), Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, 2018, pp. 5103–5111. URL: https://doi.org/10.24963/ijcai.2018/708. doi:10.24963/ijcai.2018/708.

[15] S. C. Kleene, N. De Bruijn, J. de Groot, A. C. Zaanen, Introduction to metamathematics, volume 483, van Nostrand New York, 1952.

[16] W. V. Quine, The problem of simplifying truth functions, The American mathematical monthly 59 (1952) 521–531.

[17] S. Modgil, M. Caminada, Proof theories and algorithms for abstract argumentation frameworks, in: G. R. Simari, I. Rahwan (Eds.), Argumentation in Artificial Intelligence, Springer, 2009, pp. 105–129. URL: https://doi.org/10.1007/978-0-387-98197-0_6. doi:10.1007/978-0-387-98197-0_6.

[18] A. Keshavarzi Zafarghandi, R. Verbrugge, B. Verheij, A discussion game for the grounded semantics of abstract dialectical frameworks, in: H. Prakken, S. Bistarelli, F. Santini, C. Taticchi (Eds.), Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020, volume 326 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, pp. 431–442. URL: https://doi.org/10.3233/FAIA200527. doi:10.3233/FAIA200527.