Antonis C. Kakas
Paolo Mancarella
Francesca Toni

On Argumentation Logic and Propositional Logic

Abstract. This paper studies the relationship between Argumentation Logic (AL), a recently defined logic based on the study of argumentation in AI, and classical Propositional Logic (PL). In particular, it shows that AL and PL are logically equivalent in that they have the same entailment relation from any given classically consistent theory. This equivalence follows from a correspondence between the non-acceptability of (arguments for) sentences in AL and Natural Deduction (ND) proofs of the complement of these sentences. The proof of this equivalence uses a restricted form of ND proofs, where hypotheses in the application of the Reductio of Absurdum inference rule are required to be "relevant" to the absurdity derived in the rule. The paper also discusses how the argumentative re-interpretation of PL could help control the application of ex-falso quodlibet in the presence of inconsistencies.

Keywords: Argumentation, Propositional logic, Natural deduction, Reductio ad Absurdum.

1. Introduction

Argumentation and logic have traditionally been considered as closely related, but nonetheless formally different. As early as in the work of Aristotle[1], who separates "dialectic argument" from "syllogism", argumentative reasoning has been distinguished from the demonstrative reasoning of deduction in logic. This paper studies the link between the proof theoretic view of Natural Deduction (ND) [11,12,20] for classical Propositional Logic (PL) and the recently proposed Argumentation Logic (AL) [19], and proves that the two are equivalent in the case of classically consistent PL theories.

Artificial Intelligence falls squarely within the traditional view of argumentation in AI (e.g. see [2,23] for overviews), where conflicts are captured via attacks, and conflicts are handled by defending against all attacking arguments. Our equivalence between PL and AL reconciles the aforementioned differences between classical logic and argumentation, by showing that ND proofs, including RA, can be given an argumentative reading and

[1] *Topics I* (100^a25–30), *Prior Analytics I* (24^a22–24^b12).

Presented by **David Makinson**; *Received* September 27, 2016

https://doi.org/10.1007/s11225-017-9736-x

amount to a dialectical process. We will discuss how our equivalence between PL and AL could then pave the way towards a form of paraconsistent reasoning in PL, controlling the application of the ex-falso quodlibet principle from inconsistent (and thus conflicting) information.

In AL, (sets of) propositional formulae are treated as *arguments* and entailment is defined through a notion of *acceptability* of arguments adapted from argumentation in Artificial Intelligence [8,14,15,18]. Informally, conclusions of AL are sentences supported by acceptable arguments and for which no acceptable argument exists supporting the contrary position, i.e. the negation of the sentences. The acceptability of sentences is defined in terms of notions of *attack* and *defence* between arguments and follows the natural dialectic prescription of argumentation, that for an argument to be deemed acceptable, for any attacking argument against it, there must exist an acceptable defending argument against the attacking argument. In AL, attack between arguments is defined in terms of a notion of *direct derivation* (resulting from applying ND without the Reductio ad Absurdum (RA)/Negation Introduction inference rule). Sets of propositional formulae characterise direct derivations from them, and the two views of arguments as sets of formulae and as direct derivations coincide. Informally, an argument attacks another if the two, together, directly derive an inconsistency. Then, defence is defined as a restricted form of attack, originating from the defending argument undermining, by taking the complementary position, a sentence in the argument being defended against.

To prove the desired result of equivalence between AL and PL, we consider a restricted form of ND, motivated by the argumentative view. This restricted form of ND limits the application of RA so that, informally, in any application of this inference rule, the direct derivation of the contradiction necessarily requires the hypothesis posited at the start of the application of RA. We call ND derivations where the RA rule is so restricted as fulfilling the *Genuine Absurdity Property* (GAP). GAP can be seen as a *relevance* property needed for arguments to form "proper" dialectical counter-arguments.

We show that AL and PL are logically equivalent by showing a correspondence between (non)-acceptability of formulae in AL and derivations in the restricted GAP-fulfilling ND. The equivalence result between AL and PL then follows from a central technical lemma showing that in fact the restricted form of ND is not restrictive in the conclusions derived when the given theory of premises is classically consistent, in that any ND proof can be mapped onto (at least) one corresponding GAP-fulfilling ND proof.

AL rests on separating RA from the other ND inference rules. Indeed, whereas the latter rules provide the building blocks for "direct arguments", RA does not, by operating in two steps: firstly it recognises that the posited hypothesis is inconsistent and secondly it derives the complement of the hypothesis. In AL, rather than using RA with a given hypothesis to obtain an "indirect argument" for the complement of the hypothesis, the hypothesis is recognised as (dialectically) *non-acceptable*: in the simplest case, when the hypothesis leads to inconsistency by a direct derivation (without any nested uses of RA), then it is self-attacking and hence a non-acceptable argument; in more complex cases (with nested uses of RA) the posited hypotheses can only be eventually defended by arguments which are self-attacking and again, since the defending arguments are non-acceptable, so is the posited hypothesis. Non-acceptability of the hypothesis in turn implies that it is not *AL-entailed* but *not* that its complement is AL-entailed. Indeed, the complement of the hypothesis could also be non-acceptable if the given theory is classically inconsistent.

The main technical result of the paper gives, in the case of classically consistent theories, equivalence between non-acceptability of arguments $\{\phi\}$ and full ND derivations (possibly using RA) of $\neg\phi$, thus showing that AL can recover all logical consequences of PL, including those requiring indirect proofs using RA. However, in the case of classically inconsistent theories, we show that AL does not behave like PL, by avoiding trivialisation.

The remainder of the paper is organised as follows. In Section 2 we give some preliminary definitions. In Section 3 we provide a motivating illustration of AL (formally given in Section 4) and GAP and the restricted form of ND we introduce (formally given in Section 5). In Section 6 we give the main technical result of equivalence between AL and PL. In Section 7 we discuss some implications of our results, notably in the case of inconsistent theories. Section 8 discusses related work and Section 9 concludes with a summary and a brief discussion of future work.

2. Preliminaries

Let $\mathcal{L}$ be a Propositional Logic (PL) language obtained from a given set of atoms but using only the $\neg$ and $\wedge$ connectives, without loss of generality (as every theory in PL can be equivalently formulated using only these connectives). Throughout the paper, theories and sentences will always be theories and sentences in PL, with respect to $\mathcal{L}$. We assume that $\mathcal{L}$ contains the special sentence $\bot$, informally amounting to inconsistency. Given a sentence ϕ from $\mathcal{L}$, the complement $\overline{\phi}$ of ϕ is ψ if $\phi = \neg\psi$ and $\neg\phi$ otherwise.

We use the following Natural Deduction (ND) rules, for any $\phi, \psi \in \mathcal{L}$:

$$\wedge I : \frac{\phi, \psi}{\phi \wedge \psi} \qquad \wedge E : \frac{\phi \wedge \psi}{\phi} \qquad \wedge E : \frac{\phi \wedge \psi}{\psi} \qquad \text{DN: } \frac{\neg\neg\phi}{\phi}$$

$$\neg E\text{: } \frac{\phi, \neg\phi}{\bot} \qquad RA_1 : \frac{\lceil \phi \ldots \bot \rfloor}{\neg\phi} \qquad RA_2 : \frac{\lceil \neg\phi \ldots \bot \rfloor}{\phi}$$

RA_1 is normally referred to as intuitionistic Reductio ad Absurdum (RA) and RA_2 as classical RA. We will often refer to the use of any of RA_1 and RA_2 simply as RA.[2]

Derivations result from applying these ND rules in sequence.[3] We refer to the premise $\lceil \phi \ldots \bot \rfloor$ or $\lceil \neg\phi \ldots \bot \rfloor$ of an application of RA as a *sub-derivation*, and to ϕ or $\neg\phi$, respectively, as the *hypothesis* of the sub-derivation. Note that, in the fragment of PL we consider, all sub-derivations of a derivation result from applications of RA. Also, sub-derivations may have sub-derivations in turn. We refer to each occurrence of the hypothesis of $\mathtt{d} = \lceil \phi \ldots \bot \rfloor$ or $\mathtt{d} = \lceil \neg\phi \ldots \bot \rfloor$ in a sub-derivation of $\mathtt{d}$ as a *copy* of the hypothesis.

The following example illustrates the use of the ND rules above to give derivations and sub-derivations.

Example 1. Let $T = \{\neg(\neg\alpha \wedge \beta), \neg(\gamma \wedge \neg\beta), \gamma \wedge \delta\}$. The following is a derivation of α from T:[4]

1		$\lceil \neg\alpha$			hypothesis	
2			$\lceil \beta$			hypothesis
3			$c(\neg\alpha)$			copy of $\neg\alpha$
4			$\neg\alpha \wedge \beta$			$\wedge I$
5			$\neg(\neg\alpha \wedge \beta)$			from T
6			$\bot \rfloor$			$\neg E$
7		$\neg\beta$			RA_1	
8		$\gamma \wedge \delta$			from T	
9		γ			$\wedge E$	
10		$\gamma \wedge \neg\beta$			$\wedge I$	
11		$\neg(\gamma \wedge \neg\beta)$			from T	
12		$\bot \rfloor$			$\neg E$	
13	α			RA_2		

[2] Note that using both RA_1 and RA_2 is redundant, due to the inclusion amongst the inference rules of the Double Negation (DN) rule. Moreover, using DN is redundant given the RA_2 and ¬Elimination ($\neg E$) rules. We use this redundant set of rules nonetheless to simplify the presentation of our approach.

[3] Note that we have opted for a Fitch-style notation for derivations, as this makes it easier to link with the argumentation semantics of AL.

[4] We adopt the convention to write a copy of a hypothesis ϕ in a sub-derivation as $c(\phi)$.

Steps 1–12 form a sub-derivation of this derivation, and steps 2–6 are a sub-derivation of this sub-derivation.

The following example gives a further illustration showing, in particular, that ex-falso quodlibet holds using the ND rules given above.

Example 2. Let $T = \{\neg(\alpha \wedge \beta), \neg(\alpha \wedge \neg\beta), \alpha\}$. The following is a derivation of γ from T:

1		$\lceil\neg\gamma$			hypothesis	
2			$\lceil\beta$			hypothesis
3			α			from T
4			$\alpha \wedge \beta$			$\wedge I$
5			$\neg(\alpha \wedge \beta)$			from T
6			$\bot\rfloor$			$\neg E$
7		$\neg\beta$			RA_1	
8		α			from T	
9		$\alpha \wedge \neg\beta$			$\wedge I$	
10		$\neg(\alpha \wedge \neg\beta)$			from T	
11		$\bot\rfloor$			$\neg E$	
12	γ			RA_2		

Steps 1–11 form a sub-derivation of this derivation, and steps 2–6 are a sub-derivation of this sub-derivation. Note that here the hypothesis of the sub-derivation consisting of steps 1–11 is "vacuously discharged", resulting in the derivation of (any) γ from the (classically inconsistent) theory T.

We use $T \vdash \phi$ to indicate that there is a derivation of a sentence ϕ from a theory T, using the ND rules given above. Thus, in Example 1, $T \vdash \alpha$. Note that ND as given above is a sound and complete proof system for PL when theories are formulated using only the $\neg$ and $\wedge$ connectives. Thus, $T \vdash \phi$ iff $T \models \phi$ (namely ϕ is logically entailed by T in PL).

We distinguish between (direct) derivations not using the RA rule and (indirect) derivations using the RA rule, and use the notion of direct derivation to define a notion of direct consistency:

Definition 1. *[Direct Derivation and Direct Consistency]* Let T be a theory and ϕ a sentence. A *direct derivation* of ϕ (from T) is a derivation of ϕ (from T) that does not contain any application of the RA rule. If there is a direct derivation of ϕ (from T) we say that ϕ is *directly derived* (or *derived modulo RA*) from T, denoted as $T \vdash_{MRA} \phi$. T is *directly inconsistent* iff $T \vdash_{MRA} \bot$, and *directly consistent* otherwise.

Thus, in Example 1, $T \vdash_{MRA} \gamma$ by virtue of the following simple derivation:

$$\begin{array}{lll} \mathit{1} & \gamma \wedge \delta & \quad \textit{from } T \\ \mathit{2} & \gamma & \quad \wedge E \end{array}$$

However, for the same example, it is easy to see that $T \nvdash_{MRA} \alpha$.

Trivially, if a theory is classically consistent then it is directly consistent, e.g. as in the case of T in Example 1. However, a directly consistent theory may be classically inconsistent, e.g. as in the case of $T = \{\neg(\alpha \wedge \beta), \neg(\alpha \wedge \neg\beta), \alpha\}$ in Example 2. Throughout the paper, unless specified otherwise, we assume as given a *directly consistent* theory T.

3. Motivation and Illustration

Let α, β, γ and δ be propositional formulae standing, respectively, for 'avoid steroids' 'get vaccine against hepatitis B', 'plan to travel to Africa' and 'plan to visit friends'. Then, T in Example 1 represents the information that you should not travel to Africa without getting vaccinated against hepatitis B, and you should avoid steroids when getting a hepatitis B vaccine; in addition, you are planning to travel to Africa to visit some friends. Given this information:

> Should you avoid steroids and if so why?

The ND derivation in Example 1 can be deemed to provide a positive answer to the main question. We will see that the AL reading of this derivation, afforded by our correspondence results, gives an argumentative interpretation of the classical ND reasoning, allowing also to extract from the derivation a dialectical explanation as to why this is so.

AL interprets the question as to whether you should avoid steroids as:

> Is α supported by an *argument* that can be deemed (dialectically) *acceptable*? Orthogonally, is every argument supporting $\neg\alpha$ guaranteed to be (dialectically) *non-acceptable*?

Moreover, we deem α to be *AL-entailed* if and only if both these subsidiary questions are answered positively, although, as we will prove later in Section 6, in the case of classically consistent theories a positive answer to the second question implies a positive answer to the first. Thus in this special case AL-entailment of a formula simply amounts to non-acceptability of its negation and the question as to whether and why you should avoid steroids boils down to the single subsidiary question:

Is every argument supporting $\neg\alpha$ (dialectically) *non-acceptable*?

In general, note that the definition of AL-entailment in terms of both acceptability of a formula and non-acceptability of its complement amounts to a "sceptical" view (not taking any chances when determining what is entailed). A "credulous" view, corresponding to satisfiability rather than entailment in PL, amounts to allowing for a formula and its complement to be both acceptable, but for neither to be AL-entailed. This view may be useful, in particular, when the given theory is classically inconsistent.

In order to answer these questions we need to decide what we want to consider as supporting arguments and how to determine the non-acceptability of arguments. We could take the view that any ND derivation of α from a given theory constitutes an argument supporting α, but this would imply that, in the case of a classically inconsistent theory, we could find a supporting argument for every formula, causing an "explosion" of the argumentative reasoning (resulting from the application of the ex-falso quodlibet principle). To prevent this explosion, we could alternatively take the view that any direct derivation of α constitutes an argument supporting α. We take the view that *arguments* are *sets of propositional formulae, added to the given theory* T and consider direct derivations from arguments during the dialectical process to ascertain their (non-)acceptability. So, in Example 1, $T \cup \{\alpha\}$ is an argument supporting α.

Does $T \cup \{\neg\alpha\}$ form a non-acceptable argument for its (direct) consequence $\neg\alpha$? This argument may be deemed to be *attacked*, for example, by $T \cup \{\neg\beta\}$, since $T \cup \{\neg\alpha\} \cup \{\neg\beta\}$ is directly inconsistent, and thus $T \cup \{\neg\beta\}$ provides an objection against $\neg\alpha$. Inconsistency is indeed a basic requirement for attack in argumentation in Artificial Intelligence when arguments are considered in the setting of classical logic [2]. Alternative attacks against $T \cup \{\neg\alpha\}$ exist, e.g. $T \cup \{\beta\}$, as again $T \cup \{\neg\alpha\} \cup \{\beta\}$ is directly inconsistent. In order to deem $T \cup \{\neg\alpha\}$ (dialectically) non-acceptable, all possible *defences* against at least one attack against it (e.g. the attack by $T \cup \{\neg\beta\}$ or the attack by $T \cup \{\beta\}$) need to be (dialectically) non-acceptable in turn, rendering the particular attacking argument an insurmountable hurdle for the acceptability of $T \cup \{\neg\alpha\}$. Overall, a (recursive) dialectical process is required to ascertain this non-acceptability.

How do we define defence between arguments? Clearly, inconsistency needs to be at the heart of this dialectical relation too. If we choose defence to coincide with attack, however, since (direct) inconsistency is symmetric, each argument can defend against any attack trivially by simply attacking back. To avoid this trivialisation we can either tune the dialectical process carefully, e.g. as done in [2], or choose for defence not to coincide with attack. In AL we follow the latter option and define defence as a restricted form of attack:

- defending against a directly consistent argument amounts to "undermining" the attack by providing an argument extending T with the negation of some sentence in the attacking argument being defended against, and
- defending against a directly inconsistent, and hence "self-attacking" and harmless, argument can be simply achieved by the trivial "empty" argument, consisting solely of the original theory.

Thus, in our example, both $T \cup \{\beta\}$ and $T \cup \{\}$ defend against the (directly inconsistent) attacking argument $T \cup \{\neg\beta\}$, and $T \cup \{\neg\beta\}$ defends against the (directly consistent) attacking argument $T \cup \{\beta\}$.

Given notions of arguments, attack and defence as above, non-acceptability can be defined, informally, as the "inability of arguments to defend against their attacking arguments", with empty attacks being trivially defensible against and empty defences being trivially unattackable, and hence trivially acceptable.

Overall, in our example, can $T \cup \{\neg\alpha\}$ be deemed non-acceptable? As we have seen, the argument $T \cup \{\beta\}$ attacks it and the only possible defence against it is $T \cup \{\neg\beta\}$. This is directly inconsistent with the given theory, and in particular with γ in it. Thus, the defence $T \cup \{\neg\beta\}$ is attacked by the empty argument and hence it is non-acceptable, giving that $T \cup \{\neg\alpha\}$ is non-acceptable too. Intuitively, this dialectical process sanctioning the non-acceptability of $\neg\alpha$ (and thus, since T is classically consistent, the AL-entailment of α) can be read as follows:

- a legitimate objection against taking steroids/not avoiding them ($\neg\alpha$) is supported by needing to get the hepatitis B vaccine (β);
- this objection can be defended against by simply deciding not to take this vaccine ($\neg\beta$);
- but this is at odds with the information in the given theory, and in particular the decision to travel to Africa (γ)!

As we will see later in this paper, the choices of argument, non-acceptability and AL-entailment informally illustrated here will allow us to provide an

$$\begin{array}{llll}
1 & & \lceil\neg\alpha & \\
2 & & & \lceil\neg\beta \\
3 & & & \gamma\wedge\delta \\
4 & & & \gamma \\
5 & & & \gamma\wedge\neg\beta \\
6 & & & \neg(\gamma\wedge\neg\beta) \\
7 & & & \bot\rfloor \\
8 & & \beta & \\
9 & & \neg\alpha\wedge\beta & \\
10 & & \neg(\neg\alpha\wedge\beta) & \\
11 & & \bot\rfloor & \\
12 & \alpha & &
\end{array}$$

Figure 1. An alternative ND derivation of α for Example 1

equivalent dialectical counter-part of ND in the case of classically consistent theories. Specifically, the dialectical process sanctioning the non-acceptability of $\neg\alpha$ informally illustrated before is the dialectical counterpart of the sub-derivation of the ND derivation in Figure 1.[5] Here, steps 1, 8–11 prove that $T\cup\{\beta\}$ attacks $T\cup\{\neg\alpha\}$ and the sub-derivation consisting of steps 2–7 prove that all defences against this attack (namely $T\cup\{\neg\beta\}$) are attacked by the empty argument and are thus non-acceptable.

The dialectical reading of the sub-derivation in Figure 1 is only possible because this sub-derivation and its own sub-derivation each need their posited hypothesis to directly derive the inconsistency, namely, they satisfy what we will call the *genuine absurdity property* (GAP). Instead, the original derivation given in Example 1 is not GAP-fulfilling as the direct derivation of inconsistency at step 12 does not need the hypothesis $\neg\alpha$ at step 1. Argumentatively, the original derivation identifies, at steps 7–11, the attacking argument $T\cup\{\neg\beta\}$ against $T\cup\{\neg\alpha\}$, and, at steps 2–6, that the only possible defence $T\cup\{\beta\}$ against this attack is attacked by the empty argument and thus is non-acceptable, but it fails to recognise that the attacking argument $T\cup\{\neg\beta\}$ is directly inconsistent and hence trivially defended against by the (acceptable) empty argument. Thus, since non-acceptability requires that *all* defences against *some* attack are non-acceptable, the argumentative reading of this derivation does not allow us to conclude that $\neg\alpha$ is non-acceptable. Indeed, the attacking argument $T\cup\{\neg\beta\}$ this derivation identifies is an "irrelevant" attack against the argument $T\cup\{\neg\alpha\}$, intuitively sanctioning

[5] From now on, for simplicity, we omit to indicate the ND rules used in derivations.

that an objection against taking steroids/not avoiding them is supported by not getting vaccinated against hepatitis B. Being (directly) inconsistent on its own with the information in the given theory that you are travelling to Africa, this argument could actually attack any argument, but, similarly, be defended against by appealing to the decision to travel to Africa.

Thus, in general, only GAP-fulfilling ND derivations can be given a dialectical, argumentative reading in AL. We will prove that, in the case of classically consistent theories, we can use restricted, GAP-fulfilling derivations without loss of generality, and thus always guarantee a dialectical reading of entailment in PL as illustrated above.

4. Argumentation Logic

In this section we define Argumentation Logic (AL) as in [19] but using $\vdash_{MRA}$ as given in Section 2 for syntactically restricted propositional languages with $\wedge$ and $\neg$ as the sole connectives rather than the notion of $\vdash_{MRA}$ in [19] for propositional languages with all connectives. Note that the syntactic restriction to $\wedge$ and $\neg$ forces the classical interpretation of the other connectives and does not cause a loss of generality.[6]

Definition 2. *[Argumentation Logic Framework]* The *Argumentation Logic (AL) framework corresponding to* T is the triple $\langle Args^T, Att^T, Def^T \rangle$ with:

- $Args^T = \{T \cup \Sigma | \Sigma$ is a set of sentences $\}$ is the set of all expansions of T by sets of sentences Σ ;
- given $a, b \in Args^T$, with $a = T \cup \Delta$, $b = T \cup \Gamma$, such that $\Delta \neq \{\}$, $(b, a) \in Att^T$ iff $a \cup b \vdash_{MRA} \bot$;
- given $a, d \in Args^T$, with $a = T \cup \Delta$, $(d, a) \in Def^T$ iff
  1. $d = T \cup \{\overline{\phi}\}$ for some sentence $\phi \in \Delta$ such that $\overline{\phi}$ is the complement of ϕ,[7] or
  2. $d = T \cup \{\}$ and $a \vdash_{MRA} \bot$.

[6] AL can be defined, as in [19], for unrestricted propositional languages simply by using a suitably extended notion of direct derivation ($\vdash_{MRA}$). Whether the correspondence between AL and PL can still be proven when these connectives are explicit in the underlying language is an open question, that we briefly discuss in Section 9.

[7] This notion of defence is a simplification of the notion in [19]. There, an attacking argument with a negative sentence $\neg\psi$ could be defended against by using ψ as well as $\neg\neg\psi$. Since $T \cup \{\psi\}$ and $T \cup \{\neg\neg\psi\}$ are attacked by the same arguments, the simpler notion here results in the same definition of AL.

In the remainder, *b attacks a (with respect to T)* stands for $(b, a) \in Att^T$ and *d defends*, or *is a defence against*, *a* (with respect to T) stands for $(d, a) \in Def^T$. Also, as in [19], we often use the set of sentences Σ to stand for argument $T \cup \Sigma$.

As an illustration, consider again the theory in Example 1: as we have informally seen in Section 3, $\{\neg\alpha\}$ is *attacked* by $\{\beta\}$, as $T \cup \{\neg\alpha\} \cup \{\beta\} \vdash_{MRA} \bot$, i.e. the two arguments together directly derive an inconsistency; this attack can be *defended* against by $\{\neg\beta\}$, attacked in turn by $\{\}$, as $T \cup \{\neg\beta\} \vdash_{MRA} \bot$. This attack by the empty set (i.e. simply by the given theory T) cannot possibly be defended against, since T is (classically and thus directly) consistent.

Note that the attack relation is symmetric except for the case of the empty argument. Indeed, for a, b both non-empty, it is always the case that a attacks b iff b attacks a. However, the empty argument cannot be attacked by any argument (as the attacked argument is required to be non-empty), but the empty argument can attack an argument. Finally, note that our notion of attack includes the special case of attack between a sentence and its complement, since, for any theory T, for any sentence ϕ, $\{\phi\}$ attacks $\{\neg\phi\}$ (and vice-versa).

Note also that the defence relation is a subset of the attack relation. In the first case of the definition we defend against an argument by adopting the complement of some sentence in the argument, whereas in the second case we defend against any directly inconsistent set using the empty argument. Also, trivially, if T is directly consistent as we are assuming, the empty argument cannot be possibly defended against.

AL is defined in terms of notions of acceptability and non-acceptability of arguments to determine which conclusions can be dialectically justified (or not) from the given theory. Given a generic AL framework, the intuition behind acceptability is that "an argument is acceptable iff all arguments attacking it can be successfully defended against". This intuition can be (and has been) formalised in many alternative ways, e.g. via a notion of *admissibility* as in [8], or via a *judge function* as in [2]. Here, in the tradition of [9,14] and more recently [16], we define acceptability as a "relative" notion, whereby an acceptable argument can render each of its attacking arguments non-acceptable by rendering a defence against it acceptable. As in [19], we formalise acceptability and non-acceptability as least fixed points of (monotonic) operators on the binary Cartesian product of the set of arguments, allowing to express (non-)acceptability as a relative notion, as follows.

DEFINITION 3. *[Acceptability and Non-Acceptability Operators]* Let $\langle Args^T, Att^T, Def^T \rangle$ be the AL framework corresponding to a directly consistent theory T, and $\mathcal{R}$ the set of binary relations over $Args^T$.

- The *acceptability operator* $\mathcal{A}_T : \mathcal{R} \rightarrow \mathcal{R}$ is defined as follows: for any $acc \in \mathcal{R}$ and $a, a_0 \in Args^T$: $(a, a_0) \in \mathcal{A}_T(acc)$ iff
  - $a \subseteq a_0$, or
  - for any $b \in Args^T$ such that b attacks a with respect to T,
    - $b \not\subseteq a_0 \cup a$, and
    - there is $d \in Args^T$ that defends against b with respect to T such that $(d, a_0 \cup a) \in acc$.
- The *non-acceptability operator* $\mathcal{N}_T : \mathcal{R} \rightarrow \mathcal{R}$ is defined as follows: for any $nacc \in \mathcal{R}$ and $a, a_0 \in Args^T$: $(a, a_0) \in \mathcal{N}_T(nacc)$ iff
  - $a \not\subseteq a_0$, and
  - there is $b \in Args^T$ such that b attacks a with respect to T and
    - $b \subseteq a_0 \cup a$, or
    - for any $d \in Args^T$ that defends against b with respect to T, $(d, a_0 \cup a) \in nacc$.

  $NACC^T$ and ACC^T are defined as the least fixed points of the $\mathcal{N}_T$ and $\mathcal{A}_T$ operators, respectively.

It is easy to see that the given operators are monotonic, and thus acceptability and non-acceptability are well defined. Note that non-acceptability, $NACC^T(a, a_0)$, is the same as the classical negation of $ACC^T(a, a_0)$, i.e. $NACC^T(a, a_0) = \neg ACC^T(a, a_0)$. Also, note that $NACC^T$ and ACC^T are such that

- $ACC^T(a, a_0)$, read a *is acceptable with respect to* a_0 *in* T, iff
  - $a \subseteq a_0$, or
  - for all $b \in Args^T$ such that b attacks a:
    - $b \not\subseteq a_0 \cup a$, and
    - there exists $d \in Args^T$ such that d defends against b and $ACC^T(d, a_0 \cup a)$;
- $NACC^T(a, a_0)$, read a *is not acceptable with respect to* a_0 *in* T, iff
  - $a \not\subseteq a_0$ and
  - there exists $b \in Args^T$ such that b attacks a and
    - $b \subseteq a_0 \cup a$, or
    - for all $d \in Args^T$ such that d defends against b it holds that $NACC^T(d, a_0 \cup a)$.

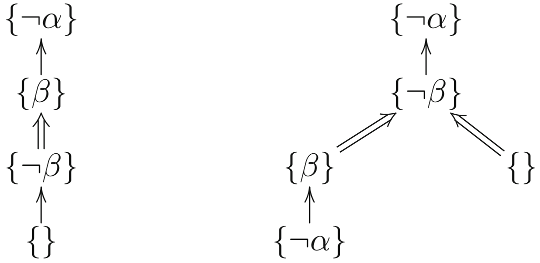

Figure 2. Illustration of $NACC^T(\{\neg\alpha\}, \{\})$ (*left*) and of failure to prove $NACC^T(\{\neg\alpha\}, \{\})$ (*right*), for the theory in Example 1 (where ↑ indicates 'attack' and ⇑ indicates 'defence')

In the remainder, we say that a is *(non-)acceptable (in T)* when a is (non-)acceptable with respect to the empty set (in T).

Figure 2 (left) illustrates, for the theory T in Example 1, the non-acceptability of $\neg\alpha$ (i.e. that $NACC^T(\{\neg\alpha\}, \{\})$ holds). Here, the (empty) leaf attack cannot be defended against and hence $\neg\beta$ is non-acceptable, giving in turn that $\neg\alpha$ is non-acceptable. This proof of non-acceptability corresponds to the dialectical process informally sanctioning the non-acceptability of $\neg\alpha$ in Section 3. Figure 2 (right) illustrates a failed attempt at determining the non-acceptability of $\neg\alpha$. Here, whereas the defence $\{\beta\}$ in the left-most branch is non-acceptable, relative to $\{\neg\alpha\}$ (as $NACC^T(\{\beta\}, \{\neg\alpha\})$ holds), the empty defence in the right-most branch is acceptable, relative to $\{\neg\alpha\}$ or any other argument. Thus, the right-most branch fails to meet the requirement for $NACC^T(\{\neg\alpha\}, \{\})$ to hold.

As an illustration of acceptability, consider again the theory in Example 1 and a model of $T \cup \{\alpha\}$. We can then use the chosen model to show that α is acceptable (i.e. that $ACC^T(\{\alpha\}, \{\})$ holds). Indeed each attack against α must contain at least a sentence that is false in the model and thus each attack can be defended against by the complement of this sentence, which is true in the chosen model. For example, in the case of T in Example 1, the attack by $\neg\alpha$ can be defended against by α itself (as α is necessarily true in any model of $T \cup \{\alpha\}$). Similarly, the attack by $\neg\neg\neg\alpha$ can be defended against by $\neg\neg\alpha$, which is true in the model. Also, the attack by $\{\neg(\alpha \wedge \neg\gamma), \neg\gamma\}$ can be defended against by γ or by $\{\}$. Figure 3 illustrates the acceptability of α. Note that there are infinitely many attacks against α but for any such attack there exists a sentence in the attack that is false in the chosen model and whose complement can be used as defence. Since these defences are true in the model we can again choose defences against arguments attacking

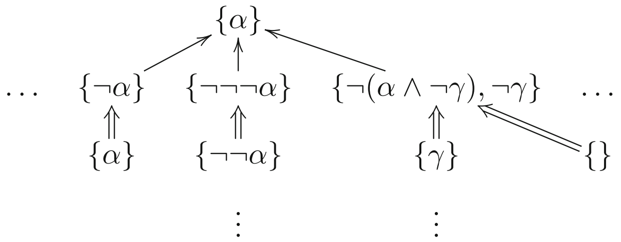

Figure 3. (Partial) illustration of $ACC^T(\{\alpha\}, \{\})$ for the theory in Example 1

these defences from the chosen model, and so recursively determine the acceptability of $\{\alpha\}$ by determining the acceptability of the defences.

Informally, in the spirit of the illustration in Section 3, the dialectical process sanctioning the acceptability of α in Figure 3 can be read as follows:

- objections to (attacks against) avoiding steroids based on recommending to take them instead may be defended against by restating one's position to avoid steroids, and
- objection based on the hypothesis of not planning to go to Africa may be defended against by pointing out that this is at odds with the contrary belief in T that you are actually planning to go to Africa;
- since each of these defending arguments (as well as defending arguments against attacks omitted for simplicity) can be deemed to be acceptable in turn, the original argument for avoiding steroids can be too.

Motivated by the argumentation perspective, where an argument is held if it can be successfully defended and it cannot be successfully objected against, entailment in AL is defined in terms of the notions of acceptability and non-acceptability, as follows:

Definition 4. *[AL-entailment]* Let ϕ be a sentence and $\overline{\phi}$ its complement. Then ϕ is *AL-entailed by* T (denoted $T \models_{AL} \phi$) iff $ACC^T(\{\phi\}, \{\})$ and $NACC^T(\{\overline{\phi}\}, \{\})$.

Hence, for T of Example 1, since we have seen that α is acceptable and $\neg\alpha$ is not, it holds that $T \models_{AL} \alpha$. Note that in general, trivially, a theory cannot AL-entail both a sentence and its complement.

5. Genuine Absurdity Property (GAP)

In this section we formalise GAP for ND derivations of the form considered in Section 2. As we discussed in Section 3, we will use GAP to define the formal correspondence between AL and PL. GAP is motivated by the dialectical argumentative perspective that self-attacking arguments can be trivially defended against and thus should not influence the result of argumentative reasoning. In the setting of AL, these self-attacking arguments amount to directly inconsistent extensions of the given (directly consistent) theory, and GAP amounts to excluding ND derivations that correspond to the use of such self-attacking arguments.

The GAP property is a property on the application of RA. We refer to ND derivations with an outermost application of the RA rule as Reductio ad Absurdum ND (RAND) derivations.

Definition 5. *[RAND-derivation]* Let ϕ be a sentence and $\overline{\phi}$ its complement. A *RAND derivation of ϕ from T* is a ND derivation of ϕ from T of the form (for $n > 1$)

$$\begin{array}{llll}
1 & & \lceil\overline{\phi} & \text{hypothesis} \\
\vdots & & \vdots & \vdots \\
\vdots & & \bot\rfloor & \vdots \\
n & \phi & & RA
\end{array}$$

Trivially, if there is a ND derivation of a sentence, then there is a RAND derivation of that sentence. Thus, in the remainder we focus on RAND derivations.

We use the following terminology: an *immediate sub-derivation* of a (sub-)derivation $\mathtt{d}$ is the sequence of all steps in a sub-derivation $\mathtt{d}'$ of $\mathtt{d}$ which are not part of any sub-derivation of $\mathtt{d}'$. As an illustration, in Example 1, steps 1,7–12 form an immediate sub-derivation $\mathtt{d}'$ of the derivation consisting of steps 1–13 and steps 2–6 form an immediate sub-derivation $\mathtt{d}''$ of $\mathtt{d}'$. In the remainder of the paper, with an abuse of notation we will refer to *immediate* sub-derivations simply as sub-derivations. Moreover, if $\mathtt{d}$ is a sub-derivation of a derivation from T we say that $\mathtt{d}$ is a sub-derivation *from T*. Furthermore, if $\mathtt{d}'$ is a sub-derivation of $\mathtt{d}$, then $\mathtt{d}$ (respectively $\mathtt{d}'$) is called a *direct ancestor* (respectively *child*) of $\mathtt{d}'$ (respectively $\mathtt{d}$) in $\mathtt{d}$; an *ancestor* of a sub-derivation $\mathtt{d}'$ is a direct ancestor of $\mathtt{d}'$ or, recursively, a direct ancestor of an ancestor of $\mathtt{d}'$; similarly, a *descendant* of a (sub-)derivation $\mathtt{d}$ is a child of $\mathtt{d}$ or, recursively, a child of a descendant of $\mathtt{d}$; a *leaf* sub-derivation is a descendant with no children. As an illustration, in

Example 1, the sub-derivation consisting of steps 1, 7–12 is a direct ancestor of the sub-derivation consisting of steps 2–6, and the latter is a leaf and a child of the former. Further, every RAND derivation $\mathtt{d}$ has exactly one child, which we call the *root* sub-derivation of $\mathtt{d}$. Non-root sub-derivations may have one, none or several children/descendants. Finally, we adopt the following notation :

Notation 1. *Given a RAND derivation* $\mathtt{d}$ *and a sub-derivation* $\mathtt{d}'$ *of* $\mathtt{d}$ *,* $\mathtt{d}'$ *is denoted by*

$$\mathtt{d}' = \lceil \phi : c(\phi_1), \ldots, c(\phi_k); \overline{\psi}_1, \ldots, \overline{\psi}_l : \bot \rfloor$$

where $k, l \geq 0$ *and*

- ϕ *is the hypothesis of* $\mathtt{d}'$*;*
- $\{\phi_1, \ldots, \phi_k\}$ *is the set of all hypotheses* χ *of RAND (sub-)derivations* $\mathtt{d}''$ *of* $\mathtt{d}$ *such that* $\mathtt{d}''$ *is an ancestor of* $\mathtt{d}'$ *in* $\mathtt{d}$ *and* χ *is copied in* $\mathtt{d}'$*;*
- $\{\psi_1, \ldots, \psi_l\}$ *is the set of all hypotheses of child RAND sub-derivations of* $\mathtt{d}'$ *in* $\mathtt{d}$*, and, for all* $j = 1, \ldots, l$*,* $\overline{\psi}_i$ *is the complement of* ψ_i*.*

Example 3. Given the theory $T = \{\neg(\alpha \wedge \neg\beta), \neg(\beta \wedge \gamma), \neg(\alpha \wedge \beta \wedge \neg\gamma)\}$, the child sub-derivation $\mathtt{d}_1$ of the RAND derivation $\mathtt{d}$ of $\neg\alpha$ in Figure 4, consisting of steps 1 and 13–16, is denoted as[8]

$$\lceil \alpha : -; \neg\beta : \bot \rfloor$$

Note that $\mathtt{d}_1$ is also the root sub-derivation of $\mathtt{d}$. The sub-derivation $\mathtt{d}_2$ consisting of steps 2, 9–12 in Figure 4 is a descendant of $\mathtt{d}$, a child of $\mathtt{d}_1$, and the (direct) ancestor of the sub-derivation $\mathtt{d}_3$ consisting of steps 3–8 in Figure 4. $\mathtt{d}_3$ is a leaf of $\mathtt{d}$.

Note that, in general, there are no copies (namely $k = 0$) in root sub-derivations, such as the earlier $\mathtt{d}_1 = \lceil \alpha : -; \neg\beta : \bot \rfloor$ in Figure 4. Note also that there are no hypotheses of sub-derivations in leaves of RAND derivations. The leaf $\mathtt{d}_3$ for the same Example 3 will thus be denoted as

$$\lceil \neg\gamma : c(\alpha), c(\beta); - : \bot \rfloor$$

With an abuse of notation, we will often use a combined notation for RAND derivations and their sub-derivations, e.g. the RAND derivation in Example 3 may also be denoted as

$$\lceil \alpha : -; \lceil \beta : -; \lceil \neg\gamma : c(\alpha), c(\beta); - : \bot \rfloor : \bot \rfloor : \bot \rfloor$$

[8] We use – to denote the empty sequence.

$$\begin{array}{lllll}
 & & \mathtt{d}_1: & & \\
1 & & \lceil\alpha & \mathtt{d}_2: & \\
2 & & & \lceil\beta & \mathtt{d}_3: \\
3 & & & & \lceil\neg\gamma \\
4 & & & & c(\alpha) \\
5 & & & & c(\beta) \\
6 & & & & \alpha\wedge\beta\wedge\neg\gamma \\
7 & & & & \neg(\alpha\wedge\beta\wedge\neg\gamma) \\
8 & & & & \bot\rfloor \\
9 & & & \gamma & \\
10 & & & \beta\wedge\gamma & \\
11 & & & \neg(\beta\wedge\gamma) & \\
12 & & & \bot\rfloor & \\
13 & & \neg\beta & & \\
14 & & \alpha\wedge\neg\beta & & \\
15 & & \neg(\alpha\wedge\neg\beta) & & \\
16 & & \bot\rfloor & & \\
17 & \neg\alpha & & &
\end{array}$$

Figure 4. RAND derivation of $\neg\alpha$ from T in Example 3, with sub-derivations $\mathtt{d}_1, \mathtt{d}_2, \mathtt{d}_3$ explicitly indicated

It is important to note that, for any sub-derivation

$$\lceil\phi : c(\phi_1), \dots, c(\phi_k); \overline{\psi}_1, \dots, \overline{\psi}_l : \bot\rfloor$$

from T, it holds that

$$T \cup \{\phi\} \cup \{\phi_1, \dots, \phi_k\} \cup \{\overline{\psi}_1, \dots, \overline{\psi}_l\} \vdash_{MRA} \bot$$

i.e. the inconsistency is directly derivable from the set of hypotheses copied from ancestor sub-derivations and the set of the complements of the hypotheses of all children sub-derivations, together with the hypothesis posited at the start of the RAND derivation. This simple observation provides a basic link between RAND derivations and their argumentative reading, as it indicates that $\{\phi\} \cup \{\phi_1, \dots, \phi_k\} \cup \{\overline{\psi}_1, \dots, \overline{\psi}_l\}$ attacks $\{\phi\}$. Note that the argumentative reading of RAND derivations imply that, for any sub-derivation $\mathtt{d}' = \lceil\phi : c(\phi_1), \dots, c(\phi_k); \overline{\psi}_1, \dots, \overline{\psi}_l : \bot\rfloor$, $\{\phi_1, \dots, \phi_k\}$ can exclude any hypotheses in ancestors of $\mathtt{d}'$ that are not used for the direct derivation of $\bot$ in $\mathtt{d}'$, as any other such hypotheses would be superfluous to the corresponding attack. Moreover, $\{\overline{\psi}_1, \ldots, \overline{\psi}_l\}$ can be chosen to be minimal, in the sense that all $\overline{\psi}_i$ are needed for the derivation of $\bot$ in $\mathtt{d}'$. In other words, we can assume that there are no "redundancies" in any sub-derivation.[9]

The GAP property is defined to disallow sub-derivations that correspond to attacking arguments which are directly inconsistent with the given (directly consistent) theory:

Definition 6. *[Genuine Absurdity Property (GAP)]* Let $\mathtt{d} = \lceil \phi : c(\phi_1), \ldots, c(\phi_k); \overline{\psi}_1, \ldots, \overline{\psi}_l : \bot \rfloor$ be a sub-derivation from T. Then $\mathtt{d}$ *satisfies the genuine absurdity property* (with respect to T) iff

$$T \cup \{\phi_1, \ldots, \phi_k\} \cup \{\overline{\psi}_1, \ldots, \overline{\psi}_l\} \nvdash_{MRA} \bot.$$

Moreover, $\mathtt{d}$ *fully satisfies the genuine absurdity property* (with respect to T) iff it satisfies the genuine absurdity property (with respect to T) and all its descendant sub-derivations satisfy the genuine absurdity property (with respect to T). Finally, a RAND derivation from T *fully satisfies the genuine absurdity property* (with respect to T) iff all its descendant sub-derivations satisfy the genuine absurdity property (with respect to T).

In other words, the genuine absurdity property is satisfied by a sub-derivation when its hypothesis is *necessary* (as opposed to simply *used*) for the direct derivation of $\bot$. Note that a RAND derivation *fully* satisfies the genuine absurdity property iff all its sub-derivations and, recursively, their sub-derivations do so.

The RAND derivation in Figure 4 fully satisfies the genuine absurdity property (with respect to T in Example 3). This is because $T \cup \{\neg\beta\} \nvdash_{MRA} \bot$, $T \cup \{\gamma\} \nvdash_{MRA} \bot$ and $T \cup \{\alpha, \beta\} \nvdash_{MRA} \bot$. Instead, the RAND derivation of α in Example 1 does not satisfy the genuine absurdity property in the root sub-derivation, as $\neg\beta$ is directly inconsistent with T without the need of the hypothesis $\neg\alpha$, whereas the alternative RAND derivation of α given in Figure 1, from the same theory of Example 1, fully satisfies the genuine absurdity property.

RAND derivations that fully satisfy the genuine absurdity property always exist for a classically consistent theory:

Lemma 1. *Let T be a* classically consistent *theory, ϕ be a sentence, and $\mathtt{d}$ be a RAND derivation of the complement $\overline{\phi}$ of ϕ from T. Then there exists*

[9] We are not addressing here the computational issues of constructing or determining RAND derivations with no redundancies. These and other computational issues concern future work.

$$
\begin{array}{lll}
 & \lceil\alpha & \\
 & & \lceil\beta \\
 & & c(\alpha) \\
 & & \alpha\wedge\beta \\
 & & \bot\rfloor \\
 & \neg\beta & \\
 & \gamma & \\
 & \gamma\wedge\neg\beta & \\
 & \bot\rfloor & \\
\neg\alpha & &
\end{array}
\qquad\qquad
\begin{array}{lll}
 & \lceil\alpha & \\
 & & \lceil\neg\beta \\
 & & \gamma \\
 & & \gamma\wedge\neg\beta \\
 & & \bot\rfloor \\
 & \beta & \\
 & \alpha\wedge\beta & \\
 & \bot\rfloor & \\
\neg\alpha & &
\end{array}
$$

Figure 5. Alternative RAND derivations of $\neg\alpha$ in Example 4: non-GAP fulfilling $\mathtt{d}$ (*left*) and GAP fulfilling $\mathtt{d}'$ (*right*)

a RAND derivation $\mathtt{d}'$ *of* $\overline{\phi}$ *from* T *that fully satisfies the genuine absurdity property.*

The proof of this central technical lemma is given in "Appendix A". The following example illustrates the lemma and the main idea behind its proof.

Example 4. Consider the theory $T = \{\neg(\alpha\wedge\beta), \neg(\gamma\wedge\neg\beta), \gamma\}$. A possible RAND derivation of $\neg\alpha$ from T is $\mathtt{d}$ in Figure 5 (left), with root sub-derivation $\mathtt{d}_1$ denoted:

$$\mathtt{d}_1 = \lceil\alpha : -; \lceil\beta : c(\alpha); - : \bot\rfloor : \bot\rfloor$$

This does not satisfy the genuine absurdity property as in the root sub-derivation $\lceil\alpha : -; \neg\beta : \bot\rfloor$ the hypothesis α is not needed to derive $\bot$, since $\neg\beta$ is directly inconsistent with T. But we can "flip" the sub-derivation $\mathtt{d}_2 = \lceil\beta : c(\alpha); - : \bot\rfloor$ "on α" and "reverse" the root sub-derivation $\mathtt{d}_1$ to construct a RAND derivation $\mathtt{d}'$ of $\neg\alpha$ that satisfies the genuine absurdity property. This is $\mathtt{d}'$ given in Figure 5 (right), with root sub-derivation:

$$\mathtt{d}'_1 = \lceil\alpha : -; \lceil\neg\beta : -; - : \bot\rfloor : \bot\rfloor$$

We are thus using the copy of the sentence α in the sub-derivation $\mathtt{d}_2$ of $\mathtt{d}$ to guide the construction of a new RAND derivation $\mathtt{d}'$ that satisfies the genuine absurdity property.

Note that the lemma does not hold in the case of classically inconsistent theories, as illustrated by the following example.

Example 5. Let $T = \{\alpha, \neg\alpha\}$, and let $\mathcal{L}$ include atoms α and β. Then no RAND derivation of β from T fully satisfies the genuine absurdity property.

6. Correspondence Between AL and PL

In order to show the correspondence of logical entailment in PL and AL-entailment, we first show that, for any directly consistent theory, if a sentence is non-acceptable in AL then its complement is provable in PL:

Theorem 2. *Let ϕ be a sentence. If $NACC^T(\{\phi\}, \{\})$ holds then there exists a RAND derivation of the complement $\overline{\phi}$ of ϕ from T.*

The proof of this theorem, already given in [19], can be found in "Appendix B" for completeness of presentation.

Conversely, for any directly consistent theory, if the complement of a sentence is provable in PL by a GAP-fulfilling RAND derivation, then the sentence is non-acceptable in AL:

Theorem 3. *Let ϕ be a sentence, and $\overline{\phi}$ its complement. If there exists a RAND derivation of $\overline{\phi}$ from T that fully satisfies the genuine absurdity property (with respect to T) then $NACC^T(\{\phi\}, \{\})$ holds.*

The proof of this theorem can be found in "Appendix C".

As an illustration of this result, consider again the RAND derivation in Figure 1, a derivation that fully satisfies the genuine absurdity property. Its root sub-derivation can be interpreted as not sanctioning $\neg\alpha$ or rendering $\neg\alpha$ non-acceptable on argumentative grounds, as we have already illustrated in Section 3.

Finally, we give the equivalence result, for classically consistent theories:

Theorem 4. *Let T be a* classically consistent *theory and ϕ be a sentence. Then $T \vdash \phi$ iff $T \models_{AL} \phi$*

Proof. Let $T \models_{AL} \phi$. By Definition 4, $NACC^T(\{\overline{\phi}\}, \{\})$ holds. By Theorem 2, there exists a RAND derivation of ϕ from T and trivially $T \vdash \phi$.

Let $T \vdash \phi$. Then there exists a RAND derivation of ϕ from T. By Lemma 1, since T is classically consistent, there is a RAND derivation of ϕ that fully satisfies the genuine absurdity property. By Theorem 3, $NACC^T(\{\overline{\phi}\}, \{\})$ holds. To prove that $T \models_{AL} \phi$ we are left to prove that $ACC^T(\{\phi\}, \{\})$ holds. Suppose, by contradiction, that $ACC^T(\{\phi\}, \{\})$ does not hold. Then $NACC^T(\{\phi\}, \{\})$ holds (since $NACC^T(\{\phi\}, \{\}) = \neg ACC^T(\{\phi\}, \{\})$) and by Theorem 2 there is a RAND derivation of $\overline{\phi}$ from

T. Thus $T \vdash \overline{\phi}$, which implies that T is classically inconsistent: contradiction. ■

This theorem shows that, for classically consistent theories expressed using only the connectives $\neg, \wedge$, a restricted form of the RA inference rules does not compromise the completeness of the ND proof system. Since all PL theories can be equivalently expressed using the connectives $\neg, \wedge$, our restricted form of ND is a complete proof system for (consistent) PL.

Theorem 2 also implies that, in the case of classically consistent theories, AL-entailment of propositional formulae, amounting to sceptically determining both the acceptability of the formulae and the non-acceptability of their complements, can be solely determined by determining the latter.

Proposition 1. *Let T be* classically consistent *and ϕ a sentence with complement $\overline{\phi}$. If $NACC^T(\{\overline{\phi}\}, \{\})$ holds then $ACC^T(\{\phi\}, \{\})$ also holds.*

Proof. By Theorem 2, since $NACC^T(\{\overline{\phi}\}, \{\})$, then $T \vdash \phi$. Suppose, by contradiction, that $ACC^T(\{\phi\}, \{\})$ does not hold. Then $NACC^T(\{\phi\}, \{\})$ holds (since $NACC^T(\{\phi\}, \{\}) = \neg ACC^T(\{\phi\}, \{\})$) and by Theorem 2 there is a RAND derivation of $\overline{\phi}$ from T and thus $T \vdash \overline{\phi}$. This implies that T is classically inconsistent: contradiction. ■

Corollary 5. *Let T be a* classically consistent *theory and ϕ a sentence with complement $\overline{\phi}$. Then $T \models_{AL} \phi$ iff $NACC^T(\{\overline{\phi}\}, \{\})$.*

7. Discussion: AL Beyond Consistent PL

We have shown in Section 6 that AL-entailment and classical entailment/ND in PL coincide in the special case of classically consistent theories expressed in the PL fragment restricted to the $\wedge$ and $\neg$ connectives. This result is based upon an argumentative reading of GAP-fulfilling RAND derivations, and it holds for any directly (rather than just classically) consistent theories. In this section we discuss the use of GAP-fulfilling RAND derivations and their equivalent argumentative reading in the case of non-classically (but directly) consistent theories, exploring in particular whether these derivations give a paraconsistent form of reasoning and their corresponding AL reading fulfils in full the vocation of argumentation to handle conflicts.

As a first example, consider $T = \{\neg(\alpha \wedge \beta), \neg(\gamma \wedge \neg\beta), \gamma, \alpha\}$ (this is the theory in Example 4 extended with α). T is directly, but not classically, consistent. It is easy to see that there are GAP-fulfilling RAND derivations for both β and $\neg\beta$. Hence, by Theorem 3, $NACC^T(\{\neg\beta\}, \{\})$ and

$NACC^T(\{\beta\},\{\})$ both hold and thus neither β nor $\neg\beta$ are (sceptically) AL-entailed by T (but both may be credulously adopted). However, both γ and α are (sceptically) AL-entailed by T while their negation is not.[10] Thus, AL does not trivialise in this example. In general, in the case of directly consistent theories, we can still draw the following (weaker) connection between direct derivations and AL-entailment, giving that AL may be deemed to give a paraconsistent extension of classical PL where (at least) the direct consequences of the given theory are AL-entailed :

PROPOSITION 2. *Let T be a directly consistent theory and ϕ a sentence such that $T \vdash_{MRA} \phi$. Then $T \models_{AL} \phi$.*

PROOF.[11] Let $a = T \cup \Delta$ be any attack against $\{\phi\}$, i.e. $T \cup \{\phi\} \cup \Delta \vdash_{MRA} \bot$. Since $T \vdash_{MRA} \phi$ then $T \cup \Delta \vdash_{MRA} \bot$. Since T is directly consistent, $\Delta \neq \{\}$. Hence any such a can be defended against by the empty argument. Since $ACC^T(\{\},\Sigma)$, for any set of sentences Σ, then $ACC^T(\{\phi\},\{\})$ holds. Moreover, since $T \vdash_{MRA} \phi$, necessarily $T \cup \{\neg\phi\} \vdash_{MRA} \bot$. Hence the empty argument attacks $\{\neg\phi\}$ and thus $NACC^T(\{\neg\phi\},\{\})$ holds. ■

The proof of this result indicates that AL can contain the explosion of PL, resulting from applying the ex-falso quodlibet principle, when directly inconsistent attacking (defending) arguments can be identified while proving the acceptability (non-acceptability, respectively) of arguments, since directly inconsistent arguments are trivially defended against (attacked, respectively) by the (acceptable) empty argument. This feature of AL, of essentially factoring out directly inconsistent arguments, is linked to the GAP restriction in RAND derivations. We illustrate this feature and its link to the containment of ex-falso quodlibet beyond Proposition 2 (namely for sentences that are not direct consequences of the theory) with the following example.[12]

EXAMPLE 6. Let $T_1 = \{\neg(\neg\alpha \wedge \gamma), \neg(\neg\alpha \wedge \neg\gamma), \neg\alpha\}$ and $T_2 = \{\neg(\neg\alpha \wedge \gamma), \neg(\neg\alpha \wedge \neg\gamma), \neg(\alpha \wedge \delta), \neg(\alpha \wedge \neg\delta)\}$. Both are directly consistent but classically inconsistent. Consider an arbitrary sentence β which is not a direct consequence of any of T_i for $i = 1, 2$, i.e. such that $T_i \not\vdash_{MRA} \beta$. It is easy to see that there are RAND derivations of β and $\neg\beta$ from both theories: using the combined notation for RAND derivations and their sub-derivations from

[10] As we have mentioned at the end of Section 4, in general, for any directly consistent theory T and sentence ϕ, if $T \models_{AL} \phi$ then $T \not\models_{AL} \overline{\phi}$, trivially by definition of $\models_{AL}$.

[11] This proof, already given in [19], is included here for completeness of presentation.

[12] We thank an anonymous reviewer for suggesting a variant of this example.

Section 5, the root sub-derivations of two possible RAND derivations of $\neg\beta$ (from T_1 and T_2) may be represented (respectively) as

$$\begin{aligned}\mathtt{d}_1 : \lceil \beta : -; \lceil \neg\alpha : -; \lceil \neg\gamma : c(\neg\alpha); - : \bot\rfloor : \bot\rfloor,\\ \lceil \alpha\wedge\beta : -; - : \bot\rfloor : \bot\rfloor\\ \mathtt{d}_2 : \lceil \beta : -; \lceil \neg\alpha : -; \lceil \neg\gamma : c(\neg\alpha); - : \bot\rfloor : \bot\rfloor,\\ \lceil \alpha\wedge\beta : -; \lceil \delta : c(\alpha\wedge\beta); - : \bot\rfloor : \bot\rfloor : \bot\rfloor\end{aligned}$$

Note that both $\mathtt{d}_1$ and $\mathtt{d}_2$ have two children sub-derivations, with hypotheses $\neg\alpha$ and $\alpha\wedge\beta$. More importantly, the derivation with sub-derivation $\mathtt{d}_2$ is GAP-fulfilling, whereas the one with sub-derivation $\mathtt{d}_1$ is not. Indeed, $T_2 \cup \{\alpha, \neg(\alpha\wedge\beta)\} \nvdash_{MRA} \bot$ whereas $T_1 \cup \{\alpha\} \vdash_{MRA} \bot$ and thus $T_1 \cup \{\alpha, \neg(\alpha\wedge\beta)\} \vdash_{MRA} \bot$, namely the hypothesis of $\mathtt{d}_1$ is not necessary for the direct derivation of inconsistency.[13] Similarly, we can construct a GAP-fulfilling RAND derivation of β from T_2.

Thus, in the case of the first, but not the second, theory in this example, GAP-fulfilment acts as a barrier against explosion of non-direct consequences. In order to prevent this explosion in the case of the second theory and in general, the notion of GAP-fulfilment needs to be revised. Although this revision is outside the scope of this paper, we discuss here a possible direction for it, driven by the dialectical perspective afforded by AL. Let us revisit Example 6 dialectically (with $A = \{\alpha, \neg(\alpha\wedge\beta)\}$ from now on):

- $\mathtt{d}_1$ identifies the attacking argument A against $\{\beta\}$, but, as the GAP-violation shows, this is directly inconsistent with the theory (T_1) alone, and thus trivially defended against by the (acceptable) empty argument; hence A is not a suitable attacking argument to ascertain the non-acceptability of $\{\beta\}$;
- $\mathtt{d}_2$ also identifies the attacking argument A against $\{\beta\}$, but this is directly consistent with the theory (T_2) and hence can no longer be defended against by the empty argument (or any other acceptable argument; indeed, by Theorem 3, $NACC^{T_2}(\{\beta\},\{\})$ holds).

Nevertheless, in both cases the attacking argument A is dialectically "problematic": in the first case it is "self-attacking", and thus non-acceptable (with respect to any argument that does not contain it, formally, for any $a_0 \not\supseteq A$, $NACC^{T_2}(A, a_0)$), and, in the second case, it is non-self-attacking

[13] Note that, since $T_1 \cup \{\alpha\} \vdash_{MRA} \bot$, an alternative root sub-derivation of a RAND-derivation of $\neg\beta$ from T_1 is $\lceil \beta : -; \lceil \neg\alpha : -; \lceil \neg\gamma : c(\neg\alpha); - : \bot\rfloor : \bot\rfloor : \bot\rfloor$. This is also non-GAP-fulfilling, exactly because $\{\alpha\}$ is directly inconsistent with T_1.

but still non-acceptable with respect to the empty argument and, importantly, with respect to $\{\beta\}$ whose acceptability status we are examining (formally $NACC^{T_2}(A, \{\})$ and $NACC^{T_2}(A, \{\beta\})$). In AL as defined in [19] and in this paper, self-attacking arguments play a role in the base case of non-acceptability (as the empty argument is contained in any argument) but non-self-attacking but still non-acceptable arguments do not contribute to the base case. To strengthen the control of AL over explosion and disregard, in the case of T_2 in Example 6, the problematic attacking argument, we could thus bring the definition of AL closer to the full abstract acceptability semantics of [9,14–16], where the fixed point definitions of acceptability and non-acceptability are interleaved, as expressed informally by "an argument is acceptable if and only if all its attacking arguments are rendered non-acceptable". This possibility and the question of how GAP for ND could correspondingly be adapted are important matters of future work.

8. Related Work

We have drawn a formal link between ND from classically consistent theories in PL and argumentation in AL. The link between logic and argumentation has been widely studied in the field of Artificial Intelligence (e.g. see [1,2,4]). Existing studies are mainly concerned with how to use classical logic or fragments thereof on which to build argumentation frameworks. In contrast, our work shows that argumentation can provide the foundation to build logical frameworks and in particular to reconstruct, from an argumentation basis, classical PL. In particular, the main technical difference between AL and other logic-based argumentation formalisms is that, rather than including RA in the underlying logic layer to build argumentation frameworks, AL recovers the effects of RA semantically, via non-acceptability. This could simplify/facilitate the development of the argumentation frameworks, for example when adding explicit preferences on the arguments or their components, by allowing to control the explosion of (irrelevant) arguments, attacks and defences between arguments. Moreover, the separation between a logic layer and an argumentation layer, widespread in much argumentation in AI, is shown by our results not to be necessary, and a uniform treatment of logic and argumentation, as afforded by AL, may play a significant role in Cognitive Computing, as discussed in [13,17].

We have seen that, in the case of directly consistent but classically inconsistent theories, RAND derivations satisfying GAP and their dialectical re-interpretation in AL provide a form of paraconsistent reasoning. Various paraconsistent propositional logics have been proposed (see e.g. [22]). Some of these logics can also be seen as restricting the application of RA in ND in order to control the ex-falso quodlibet principle. For example, D'Ottaviano and De Castro [5,7] have pointed out that the task of blocking explosion in Da Costa's paraconsistent C-systems [3] can be studied in terms of ND systems where the application of RA is restricted to contradictions derived from a sub-class of well behaved formulae in Da Costa's hierarchy.

AL falls within the purely proof-theoretic logic tradition, because of the link between AL and (GAP-fulfilling) ND. In particular, AL may be seen as following in spirit some of Tennant's works [24–26], sharing their aim to develop proof theories that on the one hand preserve fully classical entailment from consistent theories but on the other hand do not trivialise when the theories are inconsistent.[14] This is given by discarding the ex-falso quodlibet principle as well as some forms of reasoning, e.g., in [24], the excluded middle law, while keeping others, e.g., again in [24], the disjunctive syllogism (or resolution). Moreover, these works consider ways of normalising ND proofs, in the tradition of [20], by forcing *deductive relevance*, namely that conclusions of valid deductions need to follow from premises that are somehow "linked" to them. The GAP-fulfilment restriction in ND derivability shares the same aims as the restrictions to normal proofs.

We leave the study of the formal link between GAP-fulfilling ND and normal proofs as future work. The study of this link may give insights as to how the GAP property could be generalised to further control and contain explosion. Furthermore, the study of this link may help examine the transitivity of AL entailment, from the point of view of AL seen as a system of dialectical argumentation rather than a counterpart of a restricted form of ND. In addition, the notion of normal proofs could be exploited in AL to develop effective decision procedures for determining acceptability by focusing attention on attacking arguments that are linked to normal proofs of contradiction.

The study of the formal relationship between (forms of) AL and paraconsistent variants of PL is beyond the scope of this paper and is left as future work.

[14] Non-proof theoretical approaches to support paraconsistent reasoning with the aim of preserving classical reasoning from consistent theories exist, see e.g. [21] and references therein.

9. Conclusions

We have shown that the recently proposed Argumentation Logic (AL) of [19] and classical Propositional Logic (PL) are equivalent in the case of classically consistent theories. This equivalence implies that classical reasoning (in PL) can be given an alternative argumentative reading. The equivalence is proven using a restricted form of Natural Deduction (ND), where Reductio ad Absurdum (RA) is required to satisfy the *Genuine Absurdity Property* (GAP), forcing relevance of the hypothesis in the direct (namely without using RA) derivation of inconsistency. The equivalence holds since this restriction does not compromise the completeness of ND for classically consistent theories expressed using conjunction and negation only. We have also shown that, in the case of directly consistent theories (namely theories from which inconsistency cannot be directly derived) which are not classically consistent, AL controls, to a certain extent, the explosion of PL by the ex-falso quodlibet principle.

This paper paves the way to much future work, summarised as follows.

We have indicated a possible direction to further control the explosion of PL in the case of classically inconsistent but directly consistent theories, by extending the (non-)acceptability semantics in line with its use in abstract dialectical settings [9,14–16], where acceptability and non-acceptability recursively influence one another. This will also require a modification of the GAP property and the introduction of new restricted forms of ND derivations.

In this paper we did not consider directly inconsistent theories. In [19] we have indicated a possible direction for reasoning with these general theories, by focusing on directly consistent sub-theories. This or alternative directions require additional future work.

We proved that AL and PL are equivalent for restricted theories expressed using the $\neg$ and $\wedge$ connectives only, without loss of generality. When other connectives are explicitly included in the language and ND suitable extended to accommodate them, the interpretation of this connectives in AL may not be classical. A proper investigation of this aspect requires further study.

In our study of the formal connection between AL and PL we have ignored computational issues and in particular issues of computational complexity. Interesting questions in this landscape include: What is the complexity of checking that a derivation is GAP-fulfilling? What is the best algorithm for checking GAP-fulfilment? What are the complexity of/a procedural mechanism for testing whether an argument is (non-)acceptable? Is this complexity higher in the case of directly consistent but classically inconsistent theories? Can complexity results from argumentation theory [6,10] help us with the study of the computational complexity of AL? Identifying answers for these and related questions will require further work.

Several related works define non-classical, paraconsistent, or argumentation-based logics. We have briefly discussed the connection between AL and GAP-fulfilling ND with a subset of these related works that share some of our motivations. The study of the formal relationship between (forms of) AL and these and other extensions/variants of PL is beyond the scope of this paper and is left as future work.

Acknowledgements. We thank Marcello D'Agostino, Krysia Broda, Ian Hodkinson and Yiannis Moschovakis for helpful discussions about the topics of this paper. We are grateful to anonymous reviewers of this paper for helpful suggestions, and in particular to the anonymous reviewer who pointed towards Example 6 in Section 7.

A. Proof of Lemma 1

Throughout this proof, for simplicity, we will use $\neg$ to stand for complement. Thus, for any propositional formula χ, $\neg\chi$ will stand for the complement of χ and $\neg\neg\chi$ will stand for χ, i.e. the complement of the complement of χ. Similarly, for a set of formulae Ψ, $\neg\Psi$ will stand for the set of the complements of formulae in Ψ. Also, we will say that a sub-derivation $\lceil\phi : c(\phi_1), \ldots, c(\phi_k); \neg\psi_1, \ldots, \neg\psi_l : \bot\rfloor$ is a sub-derivation *of* $\neg\phi$.

To prove the lemma, we use an ordering on RAND derivations, as follows.

Definition 7. *[Derivation Ordering]* Let T be a classically consistent propositional theory, χ a sentence in $\mathcal{L}$ and $\mathtt{d}'$ and $\mathtt{d}''$ be two RAND derivations from T of $\neg\chi$ such that $\mathtt{d}''$ can be obtained from $\mathtt{d}'$ by replacing a sub-derivation $\mathtt{sd}'$ of some $\neg\chi'$ with some other sub-derivation, $\mathtt{sd}''$ of $\neg\chi'$, and vice versa. Then $\mathtt{d}'' < \mathtt{d}'$ iff the number of sub-derivations of $\mathtt{d}''$ that violate GAP is strictly smaller than the number of sub-derivations of $\mathtt{d}'$ that violate GAP.

When $\mathtt{d}'' < \mathtt{d}'$ we say that $\mathtt{d}''$ *is simpler than* $\mathtt{d}'$.

We prove the lemma using the following further lemma.

LEMMA 6. *Let T be a classically consistent propositional theory. Let $\mathtt{d}_1$ be a RAND derivation of a sentence $\neg\phi_1 \in \mathcal{L}$ from T, such that $\mathtt{d}_1$ does not satisfy GAP. Then there exists a RAND derivation $\mathtt{d}_1'$ of $\neg\phi_1$ from T that is simpler than $\mathtt{d}_1$.*

Lemma 1 then follows directly by repeatedly applying this Lemma 6 to $\mathtt{d}$ until no GAP violation exists, since there can only be finitely many GAP violations in the given RAND derivation $\mathtt{d}$.

Before we prove Lemma 6 we give some further notation.

A.1. Notation

We will often denote a RAND derivation or a sub-derivation $\mathtt{d}$ of $\neg\phi$ as $\mathtt{d}(\phi)$. Also, we will denote

$$\mathtt{d} = \lceil \phi : c(\phi_1), \ldots, c(\phi_h); \neg\psi_1, \ldots, \neg\psi_n : \bot \rfloor$$

as

$$\mathtt{d}(\phi) = \lceil \phi : c(\Phi(\mathtt{d})); \neg\Psi(\mathtt{d}) : \bot \rfloor \qquad \text{or} \qquad \mathtt{d} = \lceil \phi : c(\Phi); \neg\Psi : \bot \rfloor$$

where $c(\Phi(\mathtt{d}))$ and $c(\Phi)$ are used as a shorthand for $c(\phi_1), \ldots, c(\phi_h)$ and $\neg\Psi(\mathtt{d})$ and $\neg\Psi$ are used as a shorthand for $\neg\psi_1, \ldots, \neg\psi_n$.

With an abuse of notation, we will sometimes treat sequences as sets, and write, for example, $\Phi \cup \Psi$ to indicate $\{\chi | \chi$ is an element in the sequence Φ or $\Psi\}$.

A child sub-derivation, $\mathtt{d}'(\psi)$, of $\mathtt{d}$ will also be denoted by **child(d,ψ)** (where ψ is the hypothesis of the child sub-derivation $\mathtt{d}'$). The set of all child sub-derivations of $\mathtt{d}$ will be denoted by **children(d)** or **children(Ψ(d))**, where $\Psi(\mathtt{d})$ is the sequence/set of the hypotheses of all the child sub-derivations of $\mathtt{d}$. Then we will also denote a sub-derivation $\mathtt{d}(\phi)$ by

$$\lceil \phi : c(\Phi(\mathtt{d})); \mathbf{children}(\Psi(\mathtt{d})) : \bot \rfloor \qquad \text{or} \qquad \lceil \phi : c(\Phi); \mathbf{children}(\Psi) : \bot \rfloor.$$

Given a sub-derivation $\mathtt{d} = \lceil \phi : c(\Phi); \mathbf{children}(\Psi) : \bot \rfloor$ we will denote by the ordered list

$$H(\mathtt{d}) = [\phi_1, \ldots, \phi_m] \ (m \geq 1)$$

the list of hypotheses of the ancestor sub-derivations of $\mathtt{d}$ starting from the hypothesis ϕ_1 of the root sub-derivation to which $\mathtt{d}$ belongs down to ϕ_m, the hypothesis of the sub-derivation of which $\mathtt{d}$ is a child. Note that when $\mathtt{d}$ is a root sub-derivation then $H(\mathtt{d}) = []$.

A.2. Proof of Lemma 6

We will prove the lemma for $\mathtt{d}_1$ "fully non-redundant", in the following sense.

Definition 8. A sub-derivation $\mathtt{d} = \lceil \phi : c(\Phi); \neg\Psi : \bot \rfloor$ from T is *non-redundant* iff there exists no $S \subset \Phi \cup \neg\Psi$ such that $T \cup \{\phi\} \cup S \vdash_{MRA} \bot$ holds. A sub-derivation is *fully non-redundant* iff it is non-redundant and all its descendants are non-redundant. A RAND derivation is fully non-redundant iff all its descendants are non-redundant

A fully non-redundant (sub-)derivation contains no redundant copies of ancestor hypotheses or hypotheses of child sub-derivations. Without loss of generality we can assume that RAND derivations and their sub-derivations are fully non-redundant[15] since we can drop from (sub-)derivations any copies of ancestor hypotheses or child sub-derivations that are not necessary for the direct derivation of inconsistency, without affecting the conclusion of these (sub-)derivations.

Let us assume that the GAP violation occurs in a sub-derivation $\mathtt{d}_j$ of $\mathtt{d}_1$ given by

$$\mathtt{d}_j = \lceil \phi_j : c(\Phi(\mathtt{d}_j)); \textbf{children}(\Psi(\mathtt{d}_j)) : \bot \rfloor \quad (j \geq 1)$$

and that all sub-derivations of $\mathtt{d}_j$ satisfy GAP, i.e. we consider a deepest violation of GAP in $\mathtt{d}_1$. Note that possibly $\mathtt{d}_j = \mathtt{d}_1$. Since $\mathtt{d}_1$ is fully non-redundant, the GAP violation in $\mathtt{d}_j$ means that $T \cup \Phi(\mathtt{d}_j) \cup \neg\Psi(\mathtt{d}_j) \vdash_{MRA} \bot$, namely $T \cup \Phi(\mathtt{d}_j) \cup \neg\Psi(\mathtt{d}_j)$ is directly inconsistent, and no strict subset of $\Phi(\mathtt{d}_j) \cup \neg\Psi(\mathtt{d}_j)$ can be directly inconsistent with T. Let

$$\Phi(\mathtt{d}_j) = \phi^j_1, \ldots, \phi^j_h, \quad \neg\Psi(\mathtt{d}_j) = \neg\psi^j_1, \ldots, \neg\psi^j_n, \text{ for } 0 \leq h < j,\ 0 \leq n$$

where $\phi^j_1, \ldots, \phi^j_h$ is a sub-sequence of $H(\mathtt{d}_j) = \phi_1, \ldots, \phi_{j-1}$. Note that it cannot be that $n = h = 0$, since T is classically (and thus directly) consistent. We thus need to consider the following three cases:

Case 1: $n = 0, h > 0$,

Case 2: $n > 0, h > 0$, and

Case 3: $n > 0, h = 0$.

We will see below that the third case boils down to a sub-case of the second case. In the treatment of the first two cases, we will refer to the sub-derivation $\mathtt{d}_k$ in $\mathtt{d}_1$ that is the deepest ancestor sub-derivation of $\mathtt{d}_j$ such

[15] We are not addressing here the issue of *constructing* fully non-redundant derivations, or what the *computational complexity* of determining fully non-redundant derivations may be. We leave these and other computational issues for future work.

that its hypothesis ϕ_k is in $\Phi(\mathtt{d}_j)$. (Note that $1 \leq k < j$ in these two cases). Thus, $\Phi(\mathtt{d}_j) = \{\phi_k\} \cup \Phi'(\mathtt{d}_j)$, with all hypotheses in $\Phi'(\mathtt{d}_j)$ ancestors of ϕ_k (namely $\Phi'(\mathtt{d}_j) \subseteq H(\mathtt{d}_k)$). In particular, this means that $\phi^j_h = \phi_k$. We will refer to $\Phi'(\mathtt{d}_j)$ also as Φ'^j.

Below, given a sub-derivation $\mathtt{d}_x = \lceil \phi_x : c(\Phi(\mathtt{d}_x)); \textbf{children}(\Psi(\mathtt{d}_x)) : \bot \rfloor$ and a child sub-derivation of $\mathtt{d}_x$ with hypothesis ϕ_{x+1}, $\Psi'(\mathtt{d}_x)$ stands for $\Psi(\mathtt{d}_x)$ without ϕ_{x+1}, (namely, in the set-theoretic presentation, $\Psi(\mathtt{d}_x) = \Psi'(\mathtt{d}_x) \cup \{\phi_{x+1}\}$). Moreover, $\Phi(\mathtt{d}_x)$ and $\Psi(\mathtt{d}_x)$ are sometimes indicated as Φ^x and Ψ^x respectively.

A.2.1. Case 1. In this case, $\mathtt{d}_j$ is a leaf sub-derivation and $\{\phi^j_1, \ldots, \phi^j_h\}$ is (minimally) directly inconsistent with T. Moreover, since ϕ_k is the deepest hypothesis amongst $\phi^j_1, \ldots, \phi^j_h$ and $\phi^j_h = \phi_k$, all hypotheses in $\phi^j_1, \ldots, \phi^j_{h-1}$ are ancestors of $\mathtt{d}_k$ and thus we can construct the sub-derivation $\mathtt{d}'_1$ of $\neg\phi_1$ by replacing in $\mathtt{d}_1$ the entire sub-derivation $\mathtt{d}_k$ of $\neg\phi_k$ (including its sub-derivations) by a new sub-derivation $\mathtt{d}'_k$ of $\neg\phi_k$:[16]

$$\mathtt{d}'_k = \lceil \phi_k : c(\phi^j_1), \ldots, c(\phi^j_{h-1}); - : \bot \rfloor$$

This $\mathtt{d}'_k$ satisfies GAP as the GAP violating set $\{\phi^j_1, \ldots, \phi^j_h\}$ in $\mathtt{d}_j$ is a minimally inconsistent set (namely $\phi_k = \phi^j_h$ in this set is needed to derive the inconsistency). Then $\mathtt{d}'_1$ is simpler than $\mathtt{d}_1$ as $\mathtt{d}'_k$ satisfies GAP whereas the sub-derivation $\mathtt{d}_k$ violates GAP as it contains the sub-derivation $\mathtt{d}_j$.

A.2.2 Case 2. We consider two sub-cases:

Sub-case 2.1: there are no copies of any of the hypotheses $\phi_{k+1}, \ldots, \phi_j$ in any of the sub-derivations $\textbf{children}(\mathtt{d}_j) = \textbf{children}(\psi^j_1, \ldots, \psi^j_n)$;

Sub-case 2.2: such copies exist.

Sub-case 2.1: $\mathtt{d}_1$ is outlined in Figure 6. Similarly to Case 1, we can construct the sub-derivation $\mathtt{d}'_1$ of $\neg\phi_1$ by replacing in $\mathtt{d}_1$ the entire sub-derivation $\mathtt{d}_k$ of $\neg\phi_k$ by a new sub-derivation $\mathtt{d}'_k$ of $\neg\phi_k$:

$$\mathtt{d}'_k = \lceil \phi_k : c(\phi^j_1), \ldots, c(\phi^j_{h-1}); \textbf{children}(\psi^j_1, \ldots, \psi^j_n) : \bot \rfloor$$

This satisfies GAP as the GAP violating set $\{\phi^j_1, \ldots, \phi^j_h\} \cup \{\psi^j_1, \ldots, \psi^j_n\}$ in $\mathtt{d}_j$ is a minimally inconsistent set (namely $\phi_k = \phi^j_h$ in this set is needed to derive

[16] Here and in the remainder of the proof, we omit to give sub-derivations explicitly and use their denotation instead. Trivially, since $T \cup \{\phi^j_0, \ldots, \phi^j_{h-1}, \phi_k\} \vdash_{MRA} \bot$, this $\mathtt{d}'_k$ can be constructed.

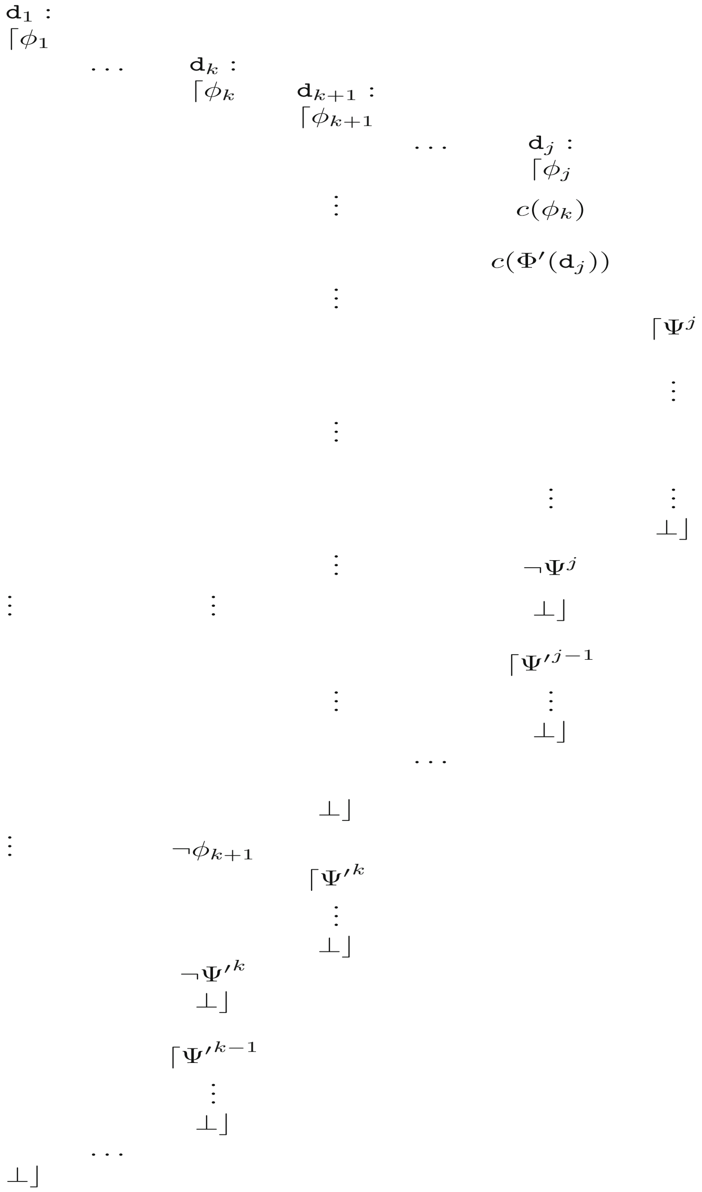

Figure 6. A sub-derivation $\mathtt{d}_1$ with a GAP violation in the sub-derivation $\mathtt{d}_j$ where (i) no sub-derivation of $\mathtt{d}_j$ contains a GAP violation and (ii) the violation necessarily involves copies of ancestor hypotheses of $\mathtt{d}_j$ but no copies of $\phi_{k+1}, \ldots, \phi_j$ are used in the children of $\mathtt{d}_j$ (Sub-case 2.1 in the Proof of Lemma 6). The hypothesis ϕ_k is the deepest ancestor hypothesis of $\mathtt{d}_j$ contributing to the GAP violation in $\mathtt{d}_j$

the inconsistency). Then $\mathtt{d}_1'$ is simpler than $\mathtt{d}_1$ as $\mathtt{d}_k'$ satisfies GAP whereas the sub-derivation $\mathtt{d}_k$ violates GAP as it contains the sub-derivation $\mathtt{d}_j$.

Sub-case 2.2: copies of some of the hypotheses $\phi_{k+1}, \ldots, \phi_j$ exist in the **children**$(\psi_1^j, \ldots, \psi_n^j)$ sub-derivations of $\mathtt{d}_j$. For this sub-case, $\mathtt{d}_1$ is outlined in Figure 7. Given this situation we consider the hypothesis ϕ_i which is the deepest ancestor hypothesis of $\mathtt{d}_j$ amongst $\phi_{k+1}, \ldots, \phi_j$ that is copied in any of the sub-derivations of $\mathtt{d}_j$. We then consider the sub-derivation $\mathtt{d}_m$ which is the first such sub-derivation of $\mathtt{d}_j$ where a copy of ϕ_i appears (note that a copy of ϕ_i may also appear in the sibling sub-derivations of $\mathtt{d}_m$, $\lceil {\Psi'}^{m-1} \ldots \bot \rfloor$, at their top level). (Note that $m > j$ and $k + 1 \leq i \leq j$.) The sub-derivation $\mathtt{d}_m$ (up to the GAP violating sub-derivation $\mathtt{d}_j$) is of the form (see also Figure 7)

$$\mathtt{d}_m = \lceil \phi_m : c(\Phi(\mathtt{d}_m)); \neg\Psi^m : \bot \rfloor = \lceil \phi_m : c(\{\phi_i\} \cup {\Phi'}^m); \neg\Psi^m : \bot \rfloor$$

where $\Phi'_m = \Phi(\mathtt{d}_m) \backslash \{\phi_i\}$. The ancestor sub-derivations of $\mathtt{d}_m$ (up to the GAP violating sub-derivation $\mathtt{d}_j$) and their hypotheses are $\mathtt{d}_{m-1}, \ldots, \mathtt{d}_{j+1}, \mathtt{d}_j$ and $\phi_{m-1}, \ldots, \phi_{j+1}, \phi_j$, respectively, where $\neg\phi_{j+1}$ is part of the GAP violating set of $\mathtt{d}_j$.

We will construct $\mathtt{d}_1'$ simpler than $\mathtt{d}_1$ by replacing $\mathtt{d}_i$ with a new sub-derivation $\mathtt{d}_i'$ of $\neg\phi_i$. This $\mathtt{d}_i'$ will be obtained from $\mathtt{d}_m$ by (I) moving its copy of ϕ_i as the hypothesis of $\mathtt{d}_i'$, (II) introducing a new child sub-derivation (of $\mathtt{d}_i'$) of $\neg\neg\phi_m$ to obtain ϕ_m in $\mathtt{d}_i'$, (III) introducing a new child sub-derivation (of $\mathtt{d}_i'$) of $\neg\neg\phi_*$, for each $\phi_* \in \{\phi_{j+1}, \ldots, \phi_{m-1}\} \cap {\Phi'}^m$, namely for each ϕ_* amongst $\phi_{j+1}, \ldots, \phi_{m-1}$ that is copied in $\mathtt{d}_m$, and (IV) similarly replacing each copy of a hypothesis ϕ_* amongst $\phi_{j+1}, \ldots, \phi_m$ in **children**$(\mathtt{d}_m)$ with a new sub-derivation of $\neg\neg\phi_*$. Overall, $\mathtt{d}_i'$ can be characterised as:

$$\begin{aligned} \mathtt{d}_i' = \lceil \phi_i : c({\Phi'}^m \backslash \{\phi_{j+1}, \ldots, \phi_{m-1}\}); \\ \{\neg\neg\phi_m\} \cup (\neg\neg{\Phi'}^m \cap \{\neg\neg\phi_{j+1}, \ldots, \neg\neg\phi_{m-1}\}) \cup \neg\Psi^m : \bot \rfloor \end{aligned}$$

We will refer to step (I) as *flipping* of $\mathtt{d}_m$ on $c(\phi_i)$, denoted $\mathbf{fl}(\mathtt{d}_m, c(\phi_i))$; to the new sub-derivation at step (II) as $\mathtt{vd}(\neg\phi_m)$; to the sub-derivations at step (III) overall as $\mathtt{vd}(\neg{\Phi''}^m)$, with ${\Phi''}^m = {\Phi'}^m \cap \{\phi_{j+1}, \ldots, \phi_{m-1}\}$; and to the sub-derivations obtained from step (IV) as $\mathbf{r} - \mathbf{children}(\mathtt{d}_m)$: these stand for **children**$(\mathtt{d}_m)$ where each copy of a hypothesis ϕ^* amongst $\phi_{j+1}, \ldots, \phi_m$ is replaced by a new sub-derivation $\mathtt{vd}(\neg\phi_*)$ to obtain ϕ_*. Note that, because of the way ϕ_i has been selected, ${\Phi'}^m$ does not contain any hypotheses from $\{\phi_{i+1}, \ldots, \phi_j\}$. Given these notations, the new sub-derivation $\mathtt{d}_i'$ that we want to construct can be represented as:

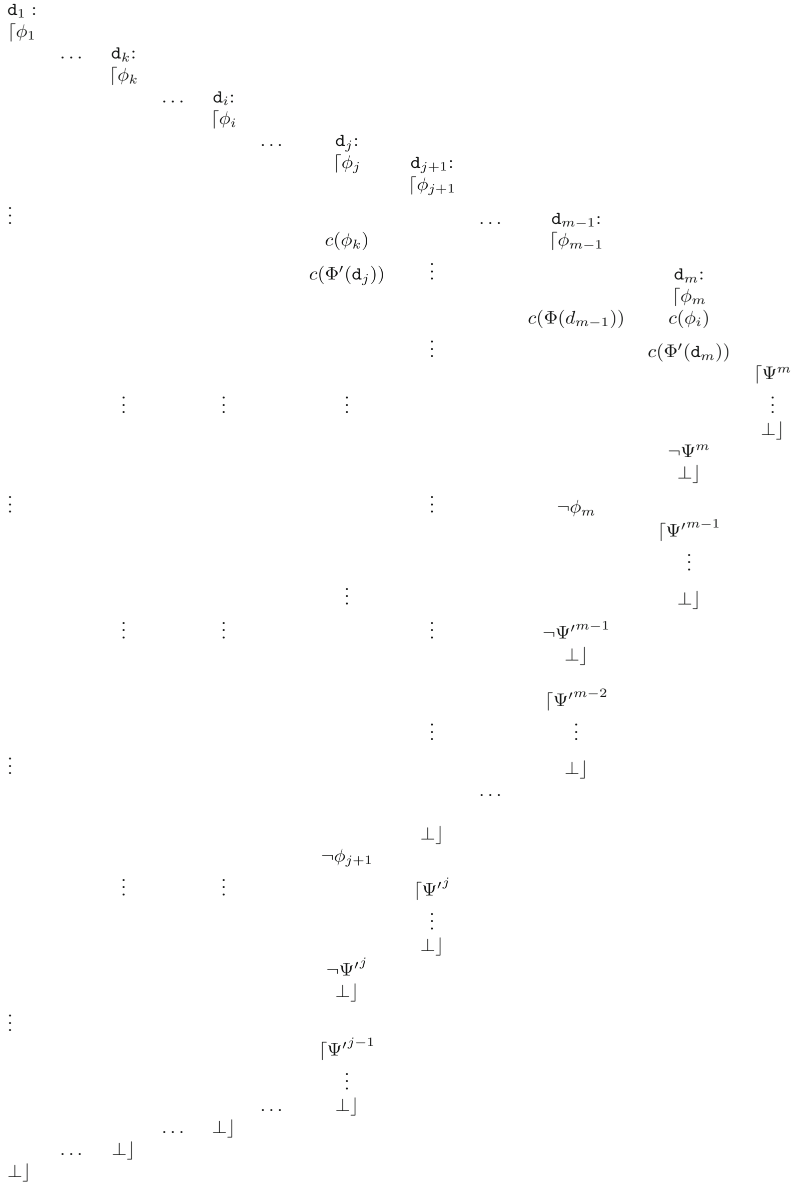

Figure 7. A sub-derivation $\mathtt{d}_1$ with a GAP violation in $\mathtt{d}_j$ and where no sub-derivation of $\mathtt{d}_j$ violates GAP. The hypothesis ϕ_k is the deepest ancestor hypothesis of $\mathtt{d}_j$ contributing to the GAP violation in $\mathtt{d}_j$. The hypothesis ϕ_i is the deepest ancestor hypothesis of $\mathtt{d}_j$ copied in the sub-derivations of $\mathtt{d}_j$, with $\mathtt{d}_m$ the first such sub-derivation (Sub-case 2.2 in the Proof of Lemma 6)

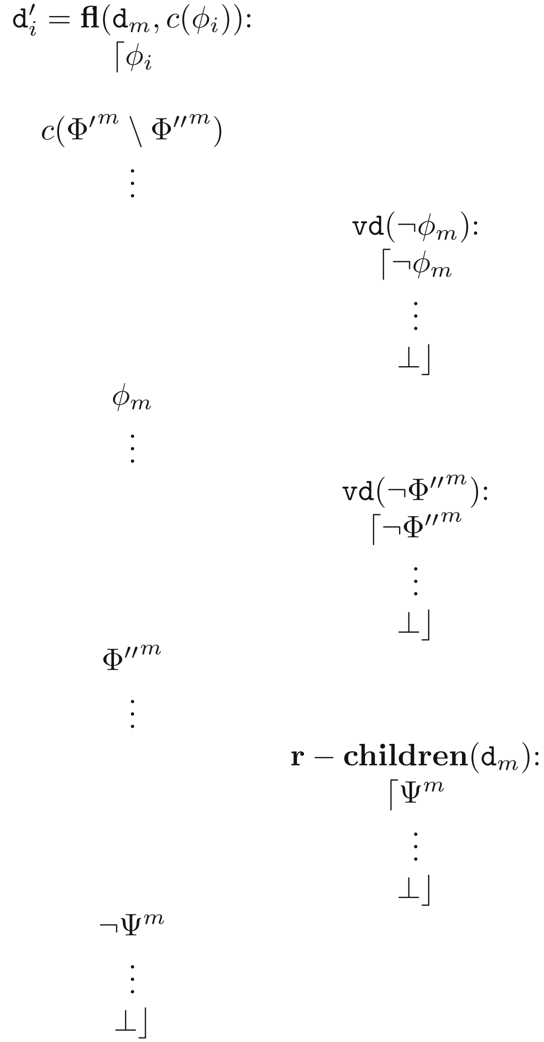

Figure 8. Derivation $\mathtt{d}'_i = \mathbf{fl}(\mathtt{d}_m, c(\phi_i))$ constructed by flipping $\mathtt{d}_m$ on its copy of ϕ_i

$$\begin{aligned} \mathtt{d}'_i = \lceil \phi_i : c(\Phi'^m \backslash \{\phi_{j+1}, \ldots, \phi_{m-1}\}); \\ \mathtt{vd}(\neg\phi_m), \mathtt{vd}(\neg\Phi''^m), \mathbf{r} - \mathbf{children}(\mathtt{d}_m) : \bot \rfloor \end{aligned}$$

This sub-derivation $\mathtt{d}'_i$ can be outlined as in Figure 8.

In order to fully define $\mathtt{d}'_i$, we need to define $\mathtt{vd}(\neg\phi_m)$, $\mathtt{vd}(\neg\Phi''^m)$ and all $\mathtt{vd}(\neg\phi_*)$ for $\mathbf{r} - \mathbf{children}(\mathtt{d}_m)$. In order to do this, it suffices to define the sub-derivations $\mathtt{vd}(\neg\phi_x)$ for each $x = j+1, \ldots, m$. Also we need to ensure that these new sub-derivations do not contain any copies of $\neg\phi_x$ for each $x = j+1, \ldots, m$ so that they will be legitimate sub-derivations of $\mathtt{d}'_i$ when this replaces $\mathtt{d}_i$ in $\mathtt{d}_1$.

The construction of these new sub-derivations $\mathtt{vd}(\neg\phi_x)$ will be based on the GAP violation in $\mathtt{d}_j$ from which we will obtain the first such new sub-derivation $\mathtt{vd}(\neg\phi_{j+1})$, then building the rest recursively from this. Hence, for $x = j+1$, the sub-derivation $\mathtt{vd}(\neg\phi_{j+1})$ is obtained from (the GAP violating) $\mathtt{d}_j$ as $\mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1})$ defined as follows:

Definition 9. Given $\mathtt{d}_j = \lceil \phi_j : c(\{\phi_k\} \cup \Phi'(\mathtt{d}_j)); \{\neg\phi_{j+1}\} \cup \neg\Psi'^{j} : \bot \rfloor$ (see Figure 7), the sub-derivation $\mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1})$ is $\lceil \neg\phi_{j+1} : c(\{\phi_k\} \cup \Phi'(\mathtt{d}_j)); \neg\Psi'^{j} : \bot \rfloor$.

Since $\mathtt{d}_j$ is GAP violating trivially $T \cup \{\phi_k\} \cup \Phi'(\mathtt{d}_j) \cup \{\neg\phi_{j+1}\} \cup \neg\Psi'^{j} \vdash_{MRA} \bot$. Moreover, the sub-derivations of $\neg\Psi'^{j}$ contain no copies of $\phi_{j+1}, \ldots \phi_m$ since they contain no copies of $\phi_{j+2}, \ldots \phi_m$ (as copies are from ancestors only) and they contain no copies of ϕ_{j+1} (because these sub-derivations are siblings of $\mathtt{d}_{j+1}$). Also, these sub-derivations do not contain any copies of ϕ_l for $l = i+1, \ldots, j$ (by the choice of i). Note also that all hypotheses in $\{\phi_k\} \cup \Phi'(\mathtt{d}_j)$ that are copies in this new sub-derivation are ancestor hypotheses of ϕ_i (since $i \geq k+1$ and all hypotheses in $\Phi'(\mathtt{d}_j)$ are ancestor hypotheses of ϕ_k since by construction ϕ_k is the deepest hypothesis copied in $\mathtt{d}_j$). As a consequence, $\mathtt{vd}(\neg\phi_{j+1}) = \mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1})$ is a legitimate sub-derivation of $\mathtt{d}'_i$ when $\mathtt{d}'_i$ replaces $\mathtt{d}_i$ in $\mathtt{d}_1$ to give $\mathtt{d}'_1$. Moreover, $\mathtt{vd}(\neg\phi_{j+1})$ fully satisfies GAP as the children sub-derivations of $\neg\Psi'^{j}$ fully satisfy GAP (by the way that $\mathtt{d}_j$ is chosen) and its root sub-derivation satisfies GAP due to the non-redundancy of $\mathtt{d}_j$.

For any $x > j+1$ (and $x \leq m$), we will construct recursively, using $\mathtt{vd}(\neg\phi_{j+1})$ as a base case, new sub-derivations $\mathtt{vd}(\neg\phi_x)$ from the sub-derivations $\mathtt{d}_{x-1}$. Let

$$\mathtt{d}_{x-1} = \lceil \phi_{x-1} : c(\Phi(\mathtt{d}_{x-1})); \{\neg\phi_x\} \cup \neg\Psi'^{x-1} : \bot \rfloor$$

be a descendant sub-derivation of $\mathtt{d}_{j+1}$ (see Figure 7). Let

$$\Phi_*^{x-1} = \{\phi_{j+1}, \ldots, \phi_{x-2}\} \cap \Phi(\mathtt{d}_{x-1}),$$

i.e. Φ_*^{x-1} consists of the ancestor hypotheses of $\mathtt{d}_{x-1}$ amongst $\{\phi_{j+1}, \ldots, \phi_m\}$ that are copied in $\mathtt{d}_{x-1}$. Note that the set $\Phi(\mathtt{d}_{x-1}) \backslash \Phi_*^{x-1}$ contains only ancestor hypotheses $\{\phi_1, \ldots, \phi_i\}$ (due to the choice of the hypothesis ϕ_i as the deepest ancestor hypothesis of $\mathtt{d}_j$ that is copied in the children of $\mathtt{d}_j$). Note also that **children**$(\mathtt{d}_{x-1}, \neg\Psi^{x-1})$ may only contain copies of $\phi_{j+1}, \ldots \phi_{x-1}$

amongst $\{\phi_{j+1}, \ldots, \phi_m\}$ (and possibly copies of $\{\phi_1, \ldots, \phi_i\}$). Then $\mathtt{vd}(\neg\phi_x)$ is obtained by *flipping* $\mathtt{d}_{x-1}$ *on* $\neg\phi_x$, giving $\mathbf{fl}(\mathtt{d}_{x-1}, \neg\phi_x)$ as follows:

Definition 10. Given $\mathtt{d}_{x-1} = \lceil \phi_{x-1} : c(\Phi(\mathtt{d}_{x-1})); \{\neg\phi_x\} \cup \neg\Psi'^{x-1} : \bot \rfloor$ (for $j+2 \leq x \leq m$, see Figure 7), the sub-derivation

$$\mathbf{fl}(\mathtt{d}_{x-1}, \neg\phi_x) = \mathtt{d}'_{x-1} = \lceil \neg\phi_x : c(\Phi(\mathtt{d}_{x-1}) \backslash \Phi_*^{x-1}); \\ \{\neg\neg\phi_{x-1}\} \cup \Phi_*^{x-1} \cup \neg\Psi'^{x-1} : \bot \rfloor$$

is such that $\mathbf{child}(\mathtt{d}'_{x-1}, \neg\neg\phi_{x-1})$, $\mathbf{children}(\mathtt{d}'_{x-1}, \Phi_*^{x-1})$, $\mathbf{children}(\mathtt{d}'_{x-1}, \neg\Psi^{x-1})$ are defined recursively as follows:

- $\mathbf{child}(\mathtt{d}'_{x-1}, \neg\neg\phi_{x-1}) = \begin{cases} \mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1}) \text{ (see Definition 9) if } x = j+2 \\ \mathbf{fl}(\mathtt{d}_{x-2}, \neg\phi_{x-1}) \text{ if } x > j+2 \end{cases}$
- $\mathbf{children}(\mathtt{d}'_{x-1}, \Phi_*^{x-1})$ consists of all sub-derivations $\mathbf{fl}(\mathtt{d}_l, \neg\phi_l)$ for $\phi_l \in \Phi_*^{x-1}$ for $j+2 \leq l \leq x-2$ and the sub-derivation $\mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1})$ (see Definition 9) if $\phi_{j+1} \in \Phi_*^{x-1}$;
- $\mathbf{children}(\mathtt{d}'_{x-1}, \neg\Psi'^{x-1})$ is $\mathbf{children}(\mathtt{d}_{x-1}, \neg\Psi^{x-1})$ without $\mathtt{d}_x$ and with each copy of ϕ_{j+1} replaced by a new sub-derivation $\mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1})$ (see Definition 9) and each copy of ϕ_l for $j+2 \leq l \leq x-1$ replaced by a new sub-derivation $\mathbf{fl}(\mathtt{d}_l, \neg\phi_l)$.

From these two Definitions 9 and 10 we get a set of new sub-derivations, $\mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1}), \mathtt{d}'_{j+1}, \ldots, \mathtt{d}'_{m-1}$ of $\neg\neg\phi_{j+1}, \neg\neg\phi_{j+2}, \ldots, \neg\neg\phi_m$ respectively. We then set in the new sub-derivation $\mathtt{d}'_i$ that we are constructing:

- $\mathtt{vd}(\neg\phi_m) = \begin{cases} \mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1}) \text{ if } m = j+1; \\ \mathbf{fl}(\mathtt{d}_{m-1}, \neg\phi_m) \text{ if } m > j+1; \end{cases}$
- each element $\mathtt{vd}(\neg\phi_l)$ in $\mathtt{vd}(\neg\Phi''^m)$ to $\mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1})$ when $l = j+1$ and to $\mathbf{fl}(\mathtt{d}_l, \neg\phi_{l+1})$ for any $l > j+1$
- $\mathbf{r} - \mathbf{children}(\mathtt{d}_m)$ to $\mathbf{children}(\mathtt{d}_m)$ where any copy of ϕ_{j+1} is replaced by a new sub-derivation $\mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1})$ and any copy of ϕ_l where $l = j+2, \ldots, m$ is replaced by a new sub-derivation $\mathbf{fl}(\mathtt{d}_l, \neg\phi_l)$.

We then replace $\mathtt{d}_i$ by $\mathtt{d}'_i$ to give the new derivation $\mathtt{d}'_1$.

By construction, the new sub-derivation $\mathtt{d}'_i$ does not contain any copies of the hypotheses $\phi_{j+1}, \ldots, \phi_m$ of the sub-derivation $\mathtt{d}_1$ as any such copy has been replaced by new sub-derivations $\mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1}), \mathbf{fl}(\mathtt{d}'_{j+1}, \neg\phi_{j+2}), \ldots, \mathbf{fl}(\mathtt{d}'_{m-1}, \neg\phi_m)$, respectively. Also due to the particular choice of ϕ_i and the way that $\mathtt{d}'_i$ is constructed this can contain only copies of hypotheses

$\phi_1, \ldots, \phi_i$ out of the $\phi_1, \ldots, \phi_m$ in $\mathtt{d}_1$. Hence $\mathtt{d}'_i$ thus obtained is a legitimate sub-derivation of $\mathtt{d}'_1$.

Moreover, $\mathtt{d}'_i$ fully satisfies GAP as this is constructed using sub-derivations of $\mathtt{d}_j$ which already satisfy GAP (by the way that $\mathtt{d}_j$ is chosen) and the two operations of *flipping* and *replacement* of copies by new sub-derivations are GAP preserving operations. For the flipping operation this is because of the non-redundancy property of (sub-)derivations. For the replacement operation this holds trivially as the definition of GAP does not distinguish between copies of ancestor hypotheses and hypotheses derived by children sub-derivations. Finally, the only sub-derivation used in the construction of $\mathtt{d}'_i$ that is not obtained from a descendant sub-derivation by using these operations is the sub-derivation $\mathtt{vd}(\mathtt{d}_j, \neg\phi_{j+1})$ given in Definition 9, and we have shown earlier that this also satisfies GAP. Thus, $\mathtt{d}'_1$ is simpler than $\mathtt{d}_1$, as required.

A.2.3. Case 3. In this case, where $h = 0$, i.e. the GAP violation in $\mathtt{d}_j$ does not involve any copies of ancestor hypotheses but comes alone from the hypotheses, $\neg\psi^j_1, \ldots, \neg\psi^j_n$, of the children sub-derivations of $\mathtt{d}_j$, is covered analogously as Sub-case 2.2, with $k = 0$. Indeed, in this special case there must exist at least one copy of a hypothesis from $\phi_1, \ldots, \phi_j$ in the children sub-derivations of $\mathtt{d}_j$ or their descendants, as, otherwise, these children sub-derivations would also constitute root sub-derivations in T and so by the soundness of Natural Deduction $T \vdash \neg\psi^j_x$ for each $x = 1, \ldots, n$ would hold. But then $T \cup \{\neg\psi^j_1, \ldots, \neg\psi^j_n\} \vdash_{MRA} \bot$ would mean that T is classically inconsistent.

B. Proof of Theorem 2

We will use the following lemma:

Lemma 7. *For any theory T and for any set of sentences Δ such that $T \cup \Delta$ is directly consistent , if $NACC^T(\{\phi\}, \Delta)$ holds then there exists a RAND derivation of the complement $\overline{\phi}$ of ϕ from $T \cup \Delta$.*

Proof of lemma 7. We use induction on the number of iterations of the $\mathcal{N}_T$ operator whose least fixed point defines $NACC^T$ (see Definition 3).

Base Case: $NACC^T(\{\phi\}, \Delta)$ holds at the first iteration of $\mathcal{N}_T$. Then, there exists A such that A attacks $\{\phi\}$ (namely $T \cup A \cup \{\phi\} \vdash_{MRA} \bot$) and $A \subseteq \Delta \cup \{\phi\}$. Thus, $T \cup \Delta \cup \{\phi\} \vdash_{MRA} \bot$ and, trivially, there exists a RAND derivation (with no RAND sub-derivations) of $\overline{\phi}$ from $T \cup \Delta$.

Induction Hypothesis: For any ψ, for any $\mathcal{E}$ such that $T \cup \mathcal{E}$ is directly consistent, if $NACC^T(\{\psi\}, \mathcal{E})$ holds after k iterations of $\mathcal{N}_T$, then there exists a RAND derivation of $\overline{\psi}$ (the complement of ψ) from $T \cup \mathcal{E}$.

Inductive Step: Assume $NACC^T(\{\phi\}, \Delta)$ holds after $k+1$ iterations of $\mathcal{N}_T$, for some Δ such that $T \cup \Delta$ is directly consistent. Then there exists A such that

(i) A attacks $\{\phi\}$ (namely $T \cup A \cup \{\phi\} \vdash_{MRA} \bot$), but $A \not\subseteq \Delta \cup \{\phi\}$; and

(ii) for each defence D against A, $NACC^T(D, \Delta \cup \{\phi\})$ holds after k iterations of $\mathcal{N}_T$.

Since $A \not\subseteq \Delta \cup \{\phi\}$, $A \neq \{\}$. Also, by compactness of $\vdash_{MRA}$ (holding by compactness of $\vdash$), we can assume that A is finite. Let $A = \{\psi_1, \ldots, \psi_n\}$. Then, $D_i = \{\overline{\psi_i}\}$, for any $i = 1, \ldots, n$, is a defence against A and hence satisfies property (ii) above, i.e. $NACC^T(D_i, \Delta \cup \{\phi\})$ holds after k iterations. Note that $T \cup \Delta \cup \{\phi\}$ is directly consistent, as otherwise Δ attacks $\{\phi\}$ with respect to T and $NACC^T(\{\phi\}, \Delta)$ would hold at the first iteration.

By the induction hypothesis, there exists a RAND derivation of ψ_i, for any $i = 1, \ldots, n$. from $T \cup \Delta \cup \{\phi\}$. We can construct a RAND derivation, `d`, of $\overline{\phi}$ from $T \cup \Delta$, with root sub-derivation `d` : $\lceil \phi \ldots \bot \rfloor$ using the RAND derivations of ψ_i from $T \cup \Delta \cup \{\phi\}$ as child sub-derivations. Thus, in the root sub-derivation we can use the formulae ψ_i, for $i = 1, \ldots, n$, from each child sub-derivation, and hence, by definition of the attack A, the derivation `d` indeed leads directly to inconsistency from $T \cup \Delta$. The resulting `d` is a RAND of $\overline{\phi}$ from $T \cup \Delta$ as any use of ϕ in the sub-derivations of ψ_i from $T \cup \Delta \cup \{\phi\}$ can now be replicated using the copy operation of ϕ from `d`. ■

To prove the theorem, assume now that $NACC^T(\{\phi\}, \{\})$ holds. Directly from Lemma 7 with $\Delta = \{\}$, if T is directly consistent then there is a RAND derivation `d` of $\overline{\phi}$ from T.

C. Proof of Theorem 3

We will use the following notation, definition and lemma:

Notation 2. *Given a RAND derivation* `d` *from $T \cup \Delta$, for some $\Delta \subseteq \mathcal{L}$, and a RAND (sub-)derivation* `d'` *of $\overline{\phi}$ (the complement of ϕ) in* `d` *(possibly* `d'=d`*),* `d'` *is denoted by*

$$\lceil \phi : c(\phi_1), \ldots, c(\phi_k); \delta_1, \ldots, \delta_m; \overline{\psi}_1, \ldots, \overline{\psi}_l : \bot \rfloor$$

where $k, m, l \geq 0$ and

- *ϕ is the hypothesis of* $\mathtt{d}'$*;*
- *$\{\phi_1, \ldots, \phi_k\}$ is the set of all hypotheses χ of RAND (sub-)derivations* $\mathtt{d}''$ *of* $\mathtt{d}$ *such that* $\mathtt{d}''$ *is an ancestor of* $\mathtt{d}'$ *in* $\mathtt{d}$ *and χ is copied in* $\mathtt{d}'$*;*
- *$\{\delta_1, \ldots, \delta_j\}$ is the set of all sentences in Δ which are copied in* $\mathtt{d}'$*;*
- *$\{\psi_1, \ldots, \psi_l\}$ is the set of all hypotheses of child RAND sub-derivations of* $\mathtt{d}'$ *in* $\mathtt{d}$*, and, $\forall j = 1, \ldots, l$, $\overline{\psi}_i$ is the complement of ψ_i.*

Definition 11. *[Extended Genuine Absurdity Property]* Let $\mathtt{d} = \lceil \phi : c(\phi_1), \ldots, c(\phi_k); \delta_1, \ldots, \delta_m; \overline{\psi}_1, \ldots, \overline{\psi}_l : \bot \rfloor$ be a RAND (sub-)derivation with respect to $T \cup \Delta$ for some $\Delta \subseteq \mathcal{L}$. Then $\mathtt{d}$ *satisfies the extended genuine absurdity property* (with respect to $\langle T, \Delta \rangle$) if and only if

$$T \cup \{\phi_1, \ldots, \phi_k\} \cup \{\delta_1, \ldots, \delta_m\} \cup \{\overline{\psi}_1, \ldots, \overline{\psi}_l\} \nvdash_{MRA} \bot.$$

Moreover, $\mathtt{d}$ *fully satisfies the extended genuine absurdity property*(with respect to $\langle T, \Delta \rangle$) if and only if it satisfies the extended genuine absurdity property (with respect to $\langle T, \Delta \rangle$) and all its sub-derivations fully satisfy the extended genuine absurdity property (with respect to $\langle T, \Delta \rangle$).

Lemma 8. *For any theory T and for any set of sentences Δ such that $T \cup \Delta$ is directly consistent , if there exists a RAND derivation of the complement $\overline{\phi}$ of a sentence ϕ from $T \cup \Delta$ that fully satisfies the extended genuine absurdity property with respect to $\langle T, \Delta \rangle$, then $NACC^T(\{\phi\}, \Delta)$ holds.*

Proof of lemma 7. We prove the lemma by induction on the depth of (the tree corresponding to) the given RAND derivation $\mathtt{d}$ of $\overline{\phi}$ from $T \cup \Delta$. This depth amounts to the maximum number of nested RAND sub-derivations of $\mathtt{d}$.

Base Case: $\mathtt{d}$ has a root sub-derivation of the form $\lceil \phi : -; - : \bot \rfloor$ (namely with no sub-derivations). Then trivially $T \cup \Delta \cup \{\phi\} \vdash_{MRA} \bot$ and so Δ attacks $\{\phi\}$ with respect to T. Thus, by definition of $NACC^T$, since $\Delta \subseteq \Delta \cup \{\phi\}$, trivially $NACC^T(\{\phi\}, \Delta)$ holds.

Induction Hypothesis: For any $\psi \in \mathcal{L}$, for any $\mathcal{E}$ such that $T \cup \mathcal{E}$ is directly consistent, if there exists a RAND derivation of depth k of the complement $\overline{\psi}$ of some sentence ψ from $T \cup \mathcal{E}$ that fully satisfies the extended genuine absurdity property with respect to $\langle T, \mathcal{E} \rangle$, then $NACC^T(\{\psi\}, \mathcal{E})$ holds.

Inductive Step: Let $\mathtt{d}$ be a RAND derivation of depth $k+1$ from $T \cup \Delta$ of $\overline{\phi}$ with a root sub-derivation $\lceil \phi : -; \delta_1, \ldots, \delta_m; \overline{\psi}_1, \ldots, \overline{\psi}_n : \bot \rfloor$ (for $n \geq 1, m \geq$

0) that satisfies the extended genuine absurdity property with respect to $\langle T, \Delta \rangle$. Let $\Delta' = \{\delta_1, \ldots, \delta_m\}$. Then $T \cup \Delta' \cup \{\overline{\psi}_1, \ldots, \overline{\psi}_n\} \cup \{\phi\} \vdash_{MRA} \bot$ and thus $A = \Delta' \cup \{\overline{\psi}_1, \ldots, \overline{\psi}_n\}$ attacks $\{\phi\}$ with respect to T. We show that every possible defence against A with respect to T is not acceptable with respect to $\Delta \cup \{\phi\}$ in T. By definition of defence there are three possibilities:

- $D = \{\}$ is not a possible defence against A with respect to T because $T \cup A \not\vdash_{MRA} \bot$ since $\Delta' \subseteq \Delta$ and the RAND derivation $\mathtt{d}$ (from $T \cup \Delta$) satisfies the extended genuine absurdity property with respect to $\langle T, \Delta \rangle$, i.e. $T \cup \Delta \cup \{\overline{\psi}_1, \ldots, \overline{\psi}_n\} \not\vdash_{MRA} \bot$.
- $D = \{\psi_i\}$ (for any $i = 1, \ldots, n$) is a candidate defence against A with respect to T.

  If $\psi_i = \phi$ then the child sub-derivation of $\mathtt{d}$ is a RAND derivation of $\overline{\phi}$ from $T \cup \Delta$ of depth k that fully satisfies the extended genuine absurdity property with respect to $\langle T, \Delta \rangle$ and the induction hypothesis gives that $NACC^T(\{\phi\}, \Delta)$ holds.

  If $\psi_i \neq \phi$ then we know that there exists a RAND sub-derivation $\mathtt{d}'$ of $\mathtt{d}$ of $\overline{\psi}_i$ from $T \cup \Delta$. This $\mathtt{d}'$ is also a RAND derivation of $\overline{\psi}_i$ from $T \cup \Delta \cup \{\phi\}$ which is of depth at most k.

  Let $\mathtt{d}' = \lceil \psi_i : X; \epsilon_1, \ldots, \epsilon_h; \overline{\gamma}_1, \ldots, \overline{\gamma}_t : \bot \rfloor$ where X can be either the empty sequence $(-)$ or $c(\phi)$. $\mathtt{d}'$ satisfies the extended genuine absurdity property with respect to $\langle T, \Delta \rangle$ (since $\mathtt{d}'$ is a sub-derivation of $\mathtt{d}$ and $\mathtt{d}$ *fully* satisfies the extended genuine absurdity property with respect to $\langle T, \Delta \rangle$).

  Note also that $T \cup \Delta \cup \{\phi\} \not\vdash_{MRA} \bot$ as otherwise we would be in the base case.

  If $X = -$ then $\mathtt{d}'$ is a RAND derivation of $\overline{\psi}_i$ from $T \cup \Delta$ of depth k that fully satisfies the extended genuine absurdity property with respect to $\langle T, \Delta \cup \{\phi\} \rangle$ and the induction hypothesis gives that $NACC^T(\{\psi_i\}, \Delta \cup \{\phi\})$ holds, as needed for $NACC^T(\{\phi\}, \Delta)$ to hold.

  If $X = c(\phi)$, then $T \cup \{\phi\} \cup \{\epsilon_1, \ldots, \epsilon_h\} \cup \{\overline{\gamma}_1, \ldots, \overline{\gamma}_t\} \not\vdash_{MRA} \bot$ since $\mathtt{d}'$ satisfies the extended genuine absurdity property with respect to $\langle T, \Delta \rangle$. We can then construct from $\mathtt{d}'$ a derivation $\mathtt{d}'' = \lceil \psi_i : -; \epsilon_1, \ldots, \epsilon_h, \phi; \overline{\gamma}_1, \ldots, \overline{\gamma}_t : \bot \rfloor$ from $T \cup \Delta \cup \{\phi\}$ by replacing each $c(\phi)$ in $\mathtt{d}'$ with a use of ϕ from the theory $T \cup \Delta \cup \{\phi\}$. $\mathtt{d}''$ is (the root sub-derivation of) a RAND derivation of $\overline{\psi}_i$ with respect to $T \cup \Delta \cup \{\phi\}$ that *fully* satisfies the extended genuine absurdity property with respect to $\langle T, \Delta \cup \{\phi\} \rangle$. Then, by the induction hypothesis, $NACC^T(\{\psi_i\}, \Delta \cup \{\phi\})$ holds as needed for $NACC^T(\{\phi\}, \Delta)$ to hold.

- $D = \{\overline{\delta}\}$, for $\overline{\delta}$ the complement of any $\delta \in \Delta$, is also a candidate defence against A. But $NACC^T(\{\overline{\delta}\}, \Delta \cup \{\phi\})$ holds trivially as $\{\overline{\delta}\}$ is attacked by $\{\delta\}$ which is a subset of Δ.

  We have shown that all candidate defences against this constructed attack A are non-acceptable with respect to $\Delta \cup \{\phi\}$ in T as required for $NACC^T(\{\phi\}, \Delta)$ to hold. ■

To prove the theorem, assume now that there exists a RAND derivation of $\overline{\phi}$ from T that satisfies the extended genuine absurdity property with respect to $\langle T, \{\}\rangle$. Trivially, the derivation fully satisfies the extended genuine absurdity property with respect to $\langle T, \{\}\rangle$ if and only if it fully satisfies the genuine absurdity property with respect to T. Then, taking $\Delta = \{\}$, directly from Lemma 8, if T is directly consistent then $NACC^T(\{\phi\}, \{\})$ holds.

References

[1] Amgoud, L., Postulates for logic-based argumentation systems, *International Journal of Approximate Reasoning* 55(9): 2028–2048, 2014.

[2] Besnard, P., and A. Hunter, Elements of Argumentation, MIT Press, Cambridge, 2008.

[3] Da Costa, N. C. A., *Sistemas Formais Inconsistentes*, Universidade Federal do Parana, Parana, 1963. Thesis.

[4] D'Agostino, M., and S. Modgil, A rational account of classical logic argumentation for real-world agents in G. A. Kaminka, M. Fox, P. Bouquet, E. Hüllermeier, V. Dignum, F. Dignum and F. van Harmelen (eds.), *ECAI 2016: 22nd European Conference on Artificial Intelligence, vol. 285 of Frontiers in Artificial Intelligence and Applications*, IOS Press, 2016, pp. 141–149.

[5] De Castro, M. A., and I. M. L. D'Ottaviano, Natural deduction for paraconsistent logic, *Logica Trianguli* 4: 3–24, 2000.

[6] Dimopoulos, Y., B. Nebel, and F. Toni, On the computational complexity of assumption-based argumentation for default reasoning, *Artificial Intelligence* 141 (1/2): 57–78, 2002.

[7] D'Ottaviano, I. M. L., On the development of paraconsistent logic and Da Costa's work, *The Journal of Non-Classical Logic* 7: 89–152, 1990.

[8] Dung, P. M., On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* 77(2): 321–358, 1995.

[9] Dung, P. M., A. C. Kakas, and P. Mancarella, Negation as failure revisited, in *Technical Report*, University of Pisa, 1992.

[10] Dunne, P. E., The computational complexity of ideal semantics, *Artificial Intelligence* 173(18): 1559–1591, 2009.

[11] Gentzen, G., Untersuchungen über das Logische Schliessen, *Mathematische Zeitschrift* 39(1): 176–210, 1935.

[12] Jaśkowski, S., On the rules of suppositions in formal logic, *Studia Logica* 1: 232–258, 1934.

[13] Kakas, A., and L. Michael, Cognitive systems: argument and cognition, *IEEE Intelligent Informatics Bulletin* 17(1): 4–20, 2016.

[14] Kakas, A. C., P. Mancarella, and P. M. Dung, The acceptability semantics for logic programs, in *ICLP*, 1994, pp. 504–519.

[15] Kakas, A. C., and F. Toni, Computing argumentation in logic programming, *Journal of Logic and Computation* 9: 515–562, 1999.

[16] Kakas, A., and P. Mancarella, On the semantics of abstract argumentation, *Journal of Logic and Computation* 23: 991–1015, 2013.

[17] Kakas, A., L. Michael, and F. Toni, Argumentation: reconciling human and automated reasoning, in *Second Workshop on Bridging the Gap Between Human and Automated Reasoning, IJCAI-16*, 2016.

[18] Kakas, A. C., and P. Mancarella, On the semantics of abstract argumentation, *Journal of Logic and Computation* 23(5): 991–1015, 2013.

[19] Kakas, A. C., F. Toni, and P. Mancarella, Argumentation logic, in S. Parsons, N. Oren, C. Reed and F. Cerutti (eds.), *Computational Models of Argument: Proceedings of COMMA 2014, vol. 266 of Frontiers in Artificial Intelligence and Applications*, IOS Press, 2014, pp. 345–356.

[20] Prawitz, D., *Natural Deduction: A Proof Theoretical Study*, Almqvist and Wiksell, Stockholm, 1965.

[21] Priest, G., Minimally inconsistent LP, *Studia Logica* 50: 321–331, 1991.

[22] Priest, G., R. Routley, and J. Norman, *Paraconsistent Logic: Essays on the Inconsistent*, Philosophia Verlag, 1989.

[23] Rahwan, I., and G. R. Simari, *Argumentation in Artificial Intelligence*, Springer, New York, 2009.

[24] Tennant, N., A proof-theoretic approach to entailment, *Philosophical Logic* 9(2): 185–209, 1980.

[25] Tennant, N., Perfect validity, entailment and paraconsistency, *Studia Logica* 43: 179–198, 1984.

[26] Tennant, N., Natural deduction and sequent calculus for intuitionistic relevant logic, *Symbolic Logic* 52(3): 665–680, 1987.

A. C. Kakas
Cyprus University
Nicosia
Cyprus

P. Mancarella
University of Pisa
Pisa
Italy

F. Toni
Imperial College London
London
UK
f.toni@imperial.ac.uk