Contamination in Formal Argumentation Systems

Martin Caminada [a]

[a]*Utrecht University, P.O.Box 80089, 3508TB Utrecht*

Abstract

Over the last decennia, many systems for formal argumentation have been defined. The problem, however, is that these systems do not always satisfy reasonable properties. In this paper, we focus on the particular property that a conflict between two arguments cannot keep other unrelated arguments from becoming justified. Although this property appears obvious, it is in fact violated by several existing argumentation formalisms. In this paper we examine what exactly goes wrong and how things can be improved.

1 Introduction

Argumentation has become an Artificial Intelligence keyword for the last fifteen years, especially in subfields such as nonmonotonic reasoning, inconsistency-tolerant reasoning, multiple-source information systems, natural language processing and human-machine interface also in connection with multi-agents systems [1, 11, 12, 6].

One of the most abstract argumentation systems is Dung's one. It has been shown that several formalisms for nonmonotonic reasoning can be expressed in terms of this argumentation system [4]. Since its original formulation, Dung's system has become very popular and different instantiations of it have been defined. This may have caused some to believe that defining an argumentation formalism is simply a matter of defining how arguments and their defeat relation can be constructed from a knowledge base. Unfortunately, things are not that simple. Several systems that apply Dung's standard semantics, such as [11, 7] (and even systems that do not, like [5]) can lead to very unintuitive results, as described in [3].

In order to avoid such anomalies, the aim of this paper is twofold: on the one hand, like in the field of belief revision where the well-known AGM-postulates serve as general properties a system for belief revision should fulfill, we are interested in defining some *principles* (also called *quality postulates* or *axioms*) that any argumentation system should fulfill. These postulates will govern the well-definition of an argumentation system and will ensure the correctness of its results. In a previous paper [3] the postulates *consistency* and *closeness* have been treated, and it was shown that these are violated by several existing argumentation formalisms [11, 7, 5]. In the current paper, we focus on the postulate of *non-contamination*. Non-contamination basically means that two arguments that rebut each-other cannot be combined into an argument that keeps any arbitrary argument from becoming justified. This postulate is especially problematic when defeasible reasoning is combined with some kind of classical logic. The fact that the postulate of non-contamination is currently not sufficiently understood can be illustrated using Pollock's New System [9]. This particular formalism violates non-contamination, and the effects of this can be far-reaching, as is explained in section 3.

This paper is structured as follows. First, in section 2, a reference argumentation formalism is described, in which some of the main problems of formal argumentation will be illustrated. Then, in section 3, the problem of non-contamination is specified, and it is shown how Pollock's formalism violates this postulate and what the effects of this violation are. In section 4, we provide a solution and prove that this solution complies with the quality aspects stated earlier. The discussion is then rounded off with some concluding remarks in section 5.

2 The reference formalism

In this section we lay out a reference formalism (inspired by [14]) in which arguments are composed by repeatedly applying strict and defeasible rules.

Definition 1 (argumentation theory). *Let $\mathcal{L}$ be a logical language that is closed under classical negation and $\phi_1, \ldots, \phi_n, \psi \in \mathcal{L}$. An* argumentation theory *is a pair $(\mathcal{S}, \mathcal{D})$ where $\mathcal{S}$ is a set of strict rules of the form $\phi_1, \ldots, \phi_n \rightarrow \psi$ and $\mathcal{D}$ is a set of defeasible rules of the form $\phi_1, \ldots, \phi_n \Rightarrow \psi$. A strict rule with an empty antecedent is called a* premisse*. A defeasible rule with an empty antecedent is called an* assumption.

Definition 2 (arguments). *Let $(\mathcal{S}, \mathcal{D})$ be an argumentation theory. The following are arguments under this theory:*

strict construction

if $A_1, \ldots, A_n$ $(n \geq 0)$ are arguments and there exists a strict rule $\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \rightarrow \psi$ then $A_1, \ldots A_n \rightarrow \psi$ is an argument (A) with:

- $\texttt{Conc}(A) = \psi$
- $\texttt{StrictRules}(A) = \texttt{StrictRules}(A_1) \cup \ldots \cup \texttt{StrictRules}(A_n) \cup \{\phi_1, \ldots, \phi_n \rightarrow \psi\}$
- $\texttt{DefRules}(A) = \texttt{DefRules}(A_1) \cup \ldots \cup \texttt{DefRules}(A_n)$
- $\texttt{SubArgs}(A) = \texttt{SubArgs}(A_1) \cup \ldots \cup \texttt{SubArgs}(A_n) \cup \{A\}$

defeasible construction

if $A_1, \ldots, A_n$ $(n \geq 0)$ are arguments and there exists a defeasible rule $\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \Rightarrow \psi$ then $A_1, \ldots A_n \Rightarrow \psi$ is an argument (A) with:

- $\texttt{Conc}(A) = \psi$
- $\texttt{StrictRules}(A) = \texttt{StrictRules}(A_1) \cup \ldots \cup \texttt{StrictRules}(A_n)$
- $\texttt{DefRules}(A) = \texttt{DefRules}(A_1) \cup \ldots \cup \texttt{DefRules}(A_n) \cup \{\phi_1, \ldots, \phi_n \Rightarrow \psi\}$
- $\texttt{SubArgs}(A) = \texttt{SubArgs}(A_1) \cup \ldots \cup \texttt{SubArgs}(A_n) \cup \{A\}$

We say that an argument A is strict *iff $\texttt{DefRules}(A) = \emptyset$. An argument A is* consistent *iff $\{\texttt{Conc}(A') \mid A' \in \texttt{SubArg}(A)\}$ does not contain a formula and its negation. We assume $\mathcal{S}$ to be* coherent*, that is, it is not possible to construct two strict arguments with opposite conclusions. Furthermore, if $\mathcal{A}$ is a set of arguments, we define $\texttt{Concs}(\mathcal{A})$ as $\{\texttt{Conc}(A) \mid A \in \mathcal{A}\}$.*

In the definition of defeat, $\lceil \ldots \rceil$ stands for the objectivation operator, as introduced by Pollock [8, 9], which translates a meta-level expression to an object-level expression (in our case: an element of $\mathcal{L}$). We also assume the presence of a syntactic function $- : \mathcal{L} \rightarrow \mathcal{L}$ such that $-\phi = \psi$ (if $\phi \equiv \neg\psi$) and $-\phi = \neg\phi$ (otherwise).

Definition 3 (defeat). *Let A and B be arguments.*

- *A* rebuts *B iff A has a conclusion ψ and B contains a defeasible rule with consequent $-\psi$.*
- *A* undercuts *B iff A has a conclusion $\neg\lceil \phi_1, \ldots, \phi_n \Rightarrow \psi \rceil$ and B contains a defeasible rule $\phi_1, \ldots, \phi_n \Rightarrow \psi$.*

A defeats *B iff A rebuts or undercuts B.*

In the following definition, the notion of defense is the same as the notion of acceptability in [4].

Definition 4 (defense / conflict-free). *Let $S \subseteq \mathcal{A}$.*

- *S* defends *an argument A iff each argument that defeats A is defeated by some argument in S.*
- *S is* conflict-free *iff there exist no $A, B \in S$ such that A defeats B.*

Definition 5 (admissibility semantics). *Let S be a conflict-free set of arguments and let $F : 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$ be a function such that $F(S) = \{A \mid A$ is defended by $S\}$.*

- *S is* admissible *iff $S \subseteq F(S)$.*
- *S is a* complete *extension iff $S = F(S)$.*
- *S is a* grounded *extension iff S is the minimal (w.r.t. set-inclusion) complete extension.*
- *S is a* preferred *extension iff S is a maximal (w.r.t. set-inclusion) complete extension.*
- *S is a* stable *extension iff S is a preferred extension that defeats every argument in $\mathcal{A} \backslash S$.*

In [3], two postulates were given that authors think should be satisfied in each formalism for defeasible argumentation: *consistency* and *closeness*.

Postulate 1 (consistency). *Let $\mathcal{C}$ be a set of formulas and $\mathcal{A}$ be a set of arguments.*

- *We say that $\mathcal{C}$ satisfies consistency iff there are no $\psi, \chi \in \mathcal{C}$ such that $\psi = -\chi$.*
- *We say that $\mathcal{A}$ satisfies direct consistency iff $\texttt{Concs}(\mathcal{A})$ satisfies consistency.*

Postulate 2 (closeness). *Let $(\mathcal{S}, \mathcal{D})$ be a defeasible theory, $\mathcal{A}$ a set of arguments under this defeasible theory and $\mathcal{C}$ a set of formulas.*

- *We say that $\mathcal{C}$ satisfies closeness iff for each strict rule $\phi_1, \ldots, \phi_n \rightarrow \psi$ it holds that if $\phi_1, \ldots, \phi_n \in \mathcal{C}$ then also $\psi \in \mathcal{C}$.*
- *We say that $\mathcal{A}$ satisfies closeness iff $\texttt{Concs}(\mathcal{A})$ satisfies closeness.*

The desirability of direct consistency is obvious, and it is satisfied by every formalism that we know of. The postulate of closeness is somewhat more subtle. The idea is, roughly, that the conclusions of the argumentation formalism should be complete. If the argumentation formalism were to be implemented in the form of an automatic inference engine, then the engine should be able to make all the inferences by its own and not leave part of the reasoning to the user. In [3], it was observed that several argumentation formalisms (such as [7, 11, 5]) violate closeness. Worse yet, when the user starts making the inferences that appear to be missing, he may even entail inconsistencies simply by applying modus ponens on strict (nondefeasible) rules. We refer to [3] for more information.

There exists, however, a relatively simple way in which both quality postulates can be satisfied. The idea is roughly not to have strict rules directly provided by the knowledge-base but instead be generated by classical entailment.

Definition 6. *Let $\mathcal{P}$ be a set of propositional formulas. We define $S(\mathcal{P})$ as*
$\{\rightarrow \psi \mid \psi \in \mathcal{P}\} \cup \{\phi_1, \ldots, \phi_n \rightarrow \psi \mid \phi_1, \ldots, \phi_n \vdash \psi\}$.

When strict rules are generated by classical entailment, like in Definition 6, both consistency and closeness are satisfied, under the original definition of defeat (Definition 3).

Theorem 1. *Let $\mathcal{P}$ be a set of propositional formulas and $\mathcal{D}$ be a set of defeasible rules. Every complete extension of the argumentation theory $(S(\mathcal{P}), \mathcal{D})$ satisfies consistency (Postulate 1) and closeness (Postulate 2).*

Proof. This follows from Theorem 2 and Theorem 3 of [3]. □

3 The problem of contamination

Basing strict rules on classical entailment may satisfy consistency and closeness; it does, however, also introduce a new kind of problem. Consider the following example.

Example 1 (John and Mary).
$\mathcal{P}$ = $\{Says(J, s)$, *"John says sugar has been added."*
$Says(M, \neg s)$, *"Mary says sugar has not been added."*
$Says(WF, r)\}$ *"The weather forecaster predicts rain today."*
$\mathcal{D}$ = $\{Says(X, y) \Rightarrow y\}$ *"People usually tell the truth."*
Notice that the defeasible rule in $\mathcal{D}$ contains two free variables; it is ment to be read as all possible instantiations. Now consider the following arguments:
J1: $Says(J, s) \Rightarrow s$
M1: $Says(M, \neg s) \Rightarrow \neg s$
JM: $J1, M1 \rightarrow \neg r$
W1: $Says(WF, r) \Rightarrow r$
Now, it holds that $J1$ defeats $M1$ and JM, $M1$ defeats $J1$ and JM, and JM defeats $W1$. This situation is graphically depicted in Figure 1. Now, if one for instance takes grounded semantics, argument $W1$ is not justified ($W1$ is not in the grounded extension) and conclusion r will hence not be considered justified. In this way, the conflict between John and Mary over a cup of coffee keeps an intuitively totally unrelated argument from becoming justified, which is clearly undesirable.

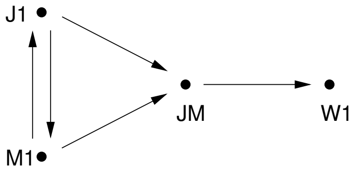

Figure 1: Arguments $J1$ and $M1$ "contaminate" argument $W1$.

Before continuing with attempts to "solve" Example 1 it is useful first to examine what it is that it violates. In essence, the example boils down to the fact that two arguments that rebut each other can keep an arbitrary argument from becoming justified. It is like the two conflicting arguments can "contaminate" an otherwise perfectly healthy argument.

Postulate 3 (non-contamination). *It may not be the case that two arguments that rebut each other can be combined into an argument that can keep any arbitrary other argument from becoming justified.*

It may be interesting to see how other formalisms for defeasible argumentation deal with the issue of non-contamination. Pollock's formalism, for instance, originally tried to deal with the problem basically by ruling out self-defeating arguments, although Pollock admitted that this approach has its difficulties [10, 8].

Another approach would be to change the semantics. An example of this is Reiter's default logic [13], which can also be given an argument-theoretic interpretation [4]. Default logic approaches the problem of non-contamination by applying stable semantics. In Figure 1 this results in two extensions: $\{J1, W1\}$ and $\{M1, W1\}$. As both of these include $W1$, $W1$ is justified under credulous as well as under sceptical stable semantics.

One important and well-known problem of stable semantics, however, is that the occurrence of odd defeat cycles may cause that no extensions exist. A simple example, which includes a self-defeating argument (a defeat cycle of 1) is the following.

Example 2 (unreliable John).
Let $\mathcal{P} = \{Says(J, unrel(J)),\ unrel(X) \supset \neg\lceil Says(X, y) \Rightarrow y\rceil\}$
and $\mathcal{D} = \{Says(X, y) \Rightarrow y\}$.
Now consider the following arguments:

J1: $Says(J, unrel(J)) \Rightarrow unrel(J)$

J2: $J1, (unrel(J) \supset \neg\lceil Says(J, unrel(J)) \Rightarrow unrel(J)\rceil) \rightarrow \neg\lceil Says(J, unrel(J)) \Rightarrow unrel(J)\rceil$

Here, argument $J2$ defeats itself and $J1$. This situation is depicted in Figure 2. Under stable semantics, no extension exists.

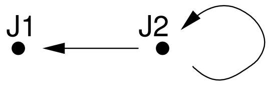

Figure 2: A self-defeating argument can cause no stable extensions to exist.

Stable semantics has often been criticized for the possible non-existence of extensions. A better alternative is sometimes seen in preferred semantics. It is not difficult to see why. In Figure 1, for instance, preferred semantics yields the same two extensions as stable semantics: $\{J1, W1\}$ and $\{M1, W1\}$. In Figure 2, preferred semantics yields exactly one extension: the empty one.

At first, preferred semantics may appear to solve the problem of non-contamination. This at least seems to be the approach of John Pollock, who made preferred semantics one of the key ingredients of his revised formalism [9]. Unfortunately, preferred semantics in itself is still not enough to satisfy non-contamination as a general property. This can be illustrated by means of the following example, in which our two characters John and Mary are still arguing over a cup of tea, but they now also both claim they are unreliable.

Example 3 (unreliable John and unreliable Mary).
Let $\mathcal{P} = \{Says(J, s),\ Says(M, \neg s),\ Says(J, unrel(J)),\ Says(M, unrel(M)),$
$unrel(X) \supset \neg\lceil Says(X, y) \Rightarrow y\rceil,\ Says(WF, r)\}$ *and* $\mathcal{D} = \{Says(X, y) \Rightarrow y\}$.
Now consider the following arguments:

J1: $Says(J, unrel(J)) \Rightarrow unrel(J)$
J2: $J1, (unrel(J) \supset \neg\lceil Says(J, unrel(J)) \Rightarrow unrel(J)\rceil) \rightarrow \neg\lceil Says(J, unrel(J)) \Rightarrow unrel(J)\rceil$
J3: $J1, (unrel(J) \supset \neg\lceil Says(J, s) \Rightarrow s\rceil) \rightarrow \neg\lceil Says(J, s) \Rightarrow s\rceil$
J4: $Says(J, s)$
J5: $J4 \Rightarrow s$
M1: $Says(M, unrel(M)) \Rightarrow unrel(M)$
M2: $M1, (unrel(M) \supset \neg\lceil Says(M, unrel(M)) \Rightarrow unrel(M)\rceil) \rightarrow \neg\lceil Says(M, unrel(M)) \Rightarrow unrel(M)\rceil$
M3: $M1, (unrel(M) \supset \neg\lceil Says(M, \neg s) \Rightarrow \neg s\rceil) \rightarrow \neg\lceil Says(M, \neg s) \Rightarrow \neg s\rceil$
M4: $Says(M, \neg s)$
M5: $M4 \Rightarrow \neg s$
W1: $Says(WF, r) \Rightarrow r$
JM: $J5, M5 \rightarrow \neg r$

The defeat relation is now as follows:
J2 defeats J1, J2, J3
J3 defeats J5
J5 defeats M5, JM
M2 defeats M1, M2, M3
M3 defeats M5
M5 defeats J5, JM
JM defeats W1

This situation is depicted in Figure 3. In this case, the only preferred extension is the empty set. The weather forecast is not justified because John and Mary are having a quarrel.

When the above example is translated to the particular syntax of Pollock's new system, as described in [9], it turns out the Pollock's system has exactly the same problem. This situation is actually quite serious. Imagine a robot that is equipped with a Pollock-style reasoning engine. One or more malicious person could then supply this robot with carefully selected information (like in the case of Example 3) after which all defeasibly inferred information becomes invalid: a total *crash* of the robot's inference capabilities.

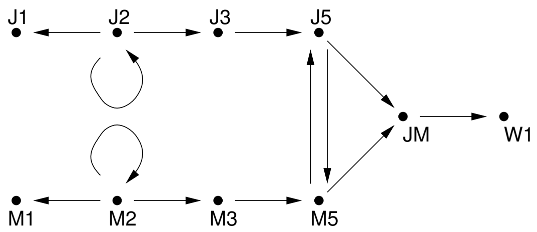

Figure 3: Preferred semantics needs not to solve non-contamination.

4 Ruling out inconsistent arguments

In this section, we provide a general solution to the issue of non-contamination in argumentation systems that combine defeasible rules with classical entailment. In the following definition, $(Args, def)_{(S(\mathcal{P}),\mathcal{D})}$ stands for the Dung-style argumentation framework associated with $\mathcal{P}$ and $\mathcal{D}$, and $(Args_c, def_c)_{(S(\mathcal{P}),\mathcal{D})}$ stands for the same argumentation framework, but without any inconsistent arguments.

Definition 7. *Let $\mathcal{P}$ be a consistent set of propositions and $\mathcal{D}$ be a set of defeasible rules. We define $(Args, def)_{(S(\mathcal{P}),\mathcal{D})}$ as the pair $(\{A \mid A$ is an argument under $(S(\mathcal{P}),\mathcal{D})\}, \{(A,B) \mid A, B$ are arguments under $(S(\mathcal{P}),\mathcal{D})$ such that A defeats B under $(S(\mathcal{P}),\mathcal{D})\})$. We say that an argument A is* consistent *iff $\{\texttt{Conc}(A') \mid A'$ is a subargument of $A\}$ is consistent. We define $(Args_c, def_c)_{(S(\mathcal{P}),\mathcal{D})}$ as $(\{A \mid A$ is a consistent argument under $(S(\mathcal{P}),\mathcal{D})\}, \{(A,B) \mid A, B$ are consistent arguments under $(S(\mathcal{P}),\mathcal{D})$ such that A defeats B under $(S(\mathcal{P}),\mathcal{D})\})$.*

The basic idea of Definition 7 is to rule out inconsistent arguments before applying one of Dung's standard semantics. The approach of ruling out a specific class of arguments, however, does not necessarily satisfy the earlier mentioned quality postulates. For instance, if one would simply rule out self-defeating arguments, the quality postulate of closeness does not hold anymore. As an illustration, if in Figure 2, one would rule out J2 because it is self-defeating, then J1 becomes justified, so $unrel(J)$ becomes a justified conclusion, even though the strict consequence of this (namely $\neg\ulcorner Says(J, unrel(J)) \Rightarrow unrel(J)\urcorner$) is not justified, therefore violating closeness.

In general, the question of whether the earlier mentioned quality postulates are still warranted once a class of arguments has been ruled out, is a non-trivial one. Fortunately, in case of ruling out inconsistent arguments, it *is* possible to prove that this approach still satisfies closeness and consistency.

Theorem 2. *Let $\mathcal{P}$ be a consistent set of propositions and $\mathcal{D}$ be a set of defeasible rules. Every complete extension of $(Args_c, def_c)_{(S(\mathcal{P}),\mathcal{D})}$ satisfies consistency (Postulate 1) and closeness (Postulate 2).*

Proof. See [2] □

The approach of ruling out incoherent arguments also means that, by definition, the quality postulate of non-contamination is satisfied.

5 Discussion

In this paper, we hope to have convinced the reader that a purely semantical approach to the problem of non-contamination (as taken by [13, 9]) is not sufficient. Our proposed solution of ruling out inconsistent arguments, on the other hand, does not have a semantical bias; it works fine not only for complete semantics (Theorem 2) but also for grounded and preferred semantics (this is because every preferred or grounded extension is also a complete extension).

A more general problem is that much of today's research regarding argumentation and defeasible logic appears to be example driven. Pollock, for instance, proudly claims that his new formalism has successfully solved example 1 [9], but forgets to examine whether any underlying fundamental properties have been satisfied. By stating general properties, like postulate 1, 2 and 3 we aim to provide a more solid basis for the evaluation of formal argumentation systems.

thanks

This work was supported by the Commission of the European Communities under contract IST-2004-002307 (ASPIC)

References

[1] L. Amgoud and C. Cayrol. Inferring from inconsistency in preference-based argumentation frameworks. *International Journal of Automated Reasoning*, Volume 29 (2):125–169, 2002.

[2] M. Caminada. Collapse in formal argumentation systems. Technical Report UU-CS-2005-023, Utrecht University, 2005.

[3] M. Caminada and L. Amgoud. An axiomatic account of formal argumentation. In *Proceedings of the AAAI-2005*, 2005.

[4] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[5] A.J. García and G.R. Simari. Defeasible logic programming: an argumentative approach. *Theory and Practice of Logic Programming*, 4(1):95–138, 2004.

[6] Sergio Alejandro Gómez and Carlos Iván Chesñevar. Integrating defeasible argumentation with fuzzy art neural networks for pattern classification. In *Proc. ECML'03*, Dubrovnik, September 2003.

[7] G. Governatori, M.J. Maher, G. Antoniou, and D. Billington. Argumentation semantics for defeasible logic. *Journal of Logic and Computation*, 14(5):675–702, 2004.

[8] J. L. Pollock. How to reason defeasibly. *Artificial Intelligence*, 57:1–42, 1992.

[9] J. L. Pollock. *Cognitive Carpentry. A Blueprint for How to Build a Person*. MIT Press, Cambridge, MA, 1995.

[10] J.L. Pollock. Self-defeating arguments. *Minds and Machines*, 1:367–392, 1991.

[11] H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7:25–75, 1997.

[12] I. Rahwan, S. D. Ramchurn, N. R. Jennings, P. McBurney, S. Parsons, and L. Sonenberg. Argumentation-based negotiation. *Knowledge engineering review*, 2004. In press.

[13] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[14] G. A. W. Vreeswijk. Studies in defeasible argumentation. *PhD thesis at Free University of Amsterdam*, 1993.