Abstraction in Argumentation: Necessary but Dangerous

Henry PRAKKEN [a] Michiel DE WINTER [b]

[a] *Department of Information and Computing Sciences, Utrecht University and Faculty of Law, University of Groningen, The Netherlands*

[b] *ABN AMRO Bank, The Netherlands*

Abstract. While work on abstract argumentation frameworks has greatly advanced the study of argumentation in AI, its use is not without danger. One danger is that the direct modelling of examples in abstract frameworks instead of through a theory of the structure of arguments and the nature of attacks leads to ad-hoc modellings. Another danger is that it may be overlooked that abstract accounts of argumentation can implicitly make assumptions that are not shared by many of their instantiations. A variant of this is where assumptions valid for specific argumentation contexts are incorrectly generalised by abstracting away from the context. This paper gives examples of both dangers. A lesson drawn from this is that abstraction in AI research, although necessary for understanding the essentials of the object of study, can oversimplify in ways that are not easily noticed without an explicit account of the structure of arguments and the nature of attack.

Keywords. Abstract argumentation frameworks, Structure of arguments, Nature of attack

1. Introduction

Since Dung's seminal paper [8], work on abstract argumentation frameworks has greatly advanced the study of argumentation in AI. Among other things, commonalities and differences between existing nonmonotonic logics and argumentation systems can be studied in terms of variations on just a small number of notions, and results can be proven for large classes of systems instead of just for particular systems. However, abstract argumentation frameworks (henceforth AFs for short) should be used with care. It is worth noting (as earlier done in [16]) that the word 'abstract' as used by Dung in [8] does not qualify 'argumentation' but 'frameworks'. In Dung's terminology, it is the framework that is abstract, not the argumentation. Strictly speaking there is no such thing as abstract argumentation, just as there is no such thing as structured argumentation. All there is, is argumentation, which can be studied at various levels of abstraction. There is nothing wrong in principle with abstract studies of argumentation, since abstraction is an indispensable tool in any kind of research. However, it should be used with care.

In particular, one should resist the temptation to think that for any given argumentation phenomenon the most principled analysis is at the level of abstract argumentation frameworks. In fact, it often is the other way around, since at the abstract level crucial notions like claims, reasons and grounds are abstracted away. For example, work on rationality postulates since [3] has shown that the theory of AFs is not all there is to say about argument evaluation. Notions of consistency and deductive closure of conclusion sets are also important and these notions can only be studied by making the structure of arguments and the nature of attacks explicit (cf. [2]). Moreover, several proposals for extending AFs with new elements, such as preferences, values or probabilities, or for studying the dynamics of AFs, implicitly make assumptions about the arguments and attacks in an AF that are not in general satisfied. The resulting formalisms are thus abstract but not general in that they model special cases, such as the special case in which all arguments, or all attacks, are independent of each other, or the special case in which all arguments are attackable.

These observations are not new. See e.g. [4,15,12,11,16] for earlier discussions. The purpose of this paper[1] is to illustrate two further dangers of naive uses of AFs. The first danger is that natural-language examples of argumentation are directly modelled in AFs instead of through a theory of the nature of arguments and attacks, leading to ad-hoc modellings so that the observations made about the resulting AFs lack general validity. In Section 3 we will give two examples of this danger from the literature on, respectively, bipolar and probabilistic abstract argumentation frameworks. The second danger is that an approach implicitly makes assumptions about the context that gives rise to an AF but then incorrectly generalises its conclusions by abstracting away from this context. In Section 4 we will illustrate this danger with a conceptual discussion of recent research on gradual notions of argument acceptability. This discussion will also provide a further example of the danger of making assumptions on the nature of arguments and attack. Before these discussions, first Dung's theory of AFs (used throughout this paper) and the *ASPIC*$^+$ framework (used in Section 3 and referred to in Section 4) will be summarised.

2. Formal Preliminaries

An *abstract argumentation framework* (AF) is a pair $\langle \mathcal{A}, attack \rangle$, where $\mathcal{A}$ is a set of arguments and $attack \subseteq \mathcal{A} \times \mathcal{A}$. The theory of AFs, initiated by [8], identifies sets of arguments (called *extensions*) which are internally coherent and defend themselves against attack. A key notion here is that of an argument $A \in \mathcal{A}$ being *defended* by a set by $S \subseteq \mathcal{A}$ if for all $B \in \mathcal{A}$: if B attacks A, then some $C \in S$ attacks B. Then relative to a given AF, $E \subseteq \mathcal{A}$ is *admissible* if E is conflict-free and defends all its members; E is a *complete extension* if E is admissible and $A \in E$ iff A is defended by E; E is a *preferred extension* if E is a $\subseteq$-maximal admissible set; E is a *stable extension* if E is admissible and attacks all arguments outside it; and $E \subseteq \mathcal{A}$ is the *grounded extension* if E is the least fixpoint of operator F, where $F(S)$ returns all arguments defended by S. It holds that any preferred, stable or grounded extension is a complete extension. For $T \in \{$complete, preferred, grounded, stable$\}$, X is *sceptically* or *credulously* justified under the T semantics if X belongs to all, respectively at least one, T extension.

We next summarise *ASPIC*$^+$ as presented in [13]. It defines the notion of an abstract *argumentation system* as a structure consisting of a logical language $\mathcal{L}$ with negation, two sets $\mathcal{R}_s$ and $\mathcal{R}_d$ of strict and defeasible inference rules, and a naming convention n in $\mathcal{L}$ for defeasible rules in order to talk about the applicability of defeasible rules in $\mathcal{L}$.

[1] The title of this paper is inspired by the title of [18].

Definition 1 [Argumentation System] An *argumentation system* (AS) is a tuple $AS = (\mathcal{L}, \mathcal{R}, ^{-}, n)$ where:

- $\mathcal{L}$ is a logical language consisting of propositional or ground predicate-logic literals
- $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict ($\mathcal{R}_s$) and defeasible ($\mathcal{R}_d$) inference rules of the form $\varphi_1, \ldots, \varphi_n \to \varphi$ and $\varphi_1, \ldots, \varphi_n \Rightarrow \varphi$ respectively (where φ_i, φ are meta-variables ranging over wff in $\mathcal{L}$), such that $\mathcal{R}_s \cap \mathcal{R}_d = \emptyset$. $\varphi_1, \ldots, \varphi_n$ are called the *antecedents* and φ the *consequent* of the rule.
- $^{-}$ is a function from $\mathcal{L}$ to $2^{\mathcal{L}}$, such that:

  φ is a *contrary* of ψ if $\varphi \in \overline{\psi}$, $\psi \notin \overline{\varphi}$;

  φ is a *contradictory* of ψ (denoted by '$\varphi = -\psi$'), if $\varphi \in \overline{\psi}$, $\psi \in \overline{\varphi}$.
- n is a partial function such that $n : \mathcal{R}_d \longrightarrow \mathcal{L}$.

Below we will for all φ assume that $\varphi = -\neg\varphi$ and $\neg\varphi = -\varphi$. Further contrariness relations will be defined when needed.

Definition 2 A *knowledge base* in an $AS = (\mathcal{L}, \mathcal{R}, n)$ is a set $\mathcal{K} \subseteq \mathcal{L}$ consisting of two disjoint subsets $\mathcal{K}_n$ (the *axioms*) and $\mathcal{K}_p$ (the *ordinary premises*).

Definition 3 [**Arguments**] An *argument* A on the basis of a knowledge base $\mathcal{K}$ in an argumentation system AS is a structure obtainable by applying one or more of the following steps finitely many times:

1. φ if $\varphi \in \mathcal{K}$ with: $\texttt{Prem}(A) = \{\varphi\}$; $\texttt{Conc}(A) = \varphi$; $\texttt{Sub}(A) = \{\varphi\}$; $\texttt{TopRule}(A)$ = undefined.
2. $[A_1], \ldots, [A_n] \to \psi$[2] if $A_1, \ldots, A_n$ are arguments such that $\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \to \psi \in \mathcal{R}$ with:
   $\texttt{Prem}(A) = \texttt{Prem}(A_1) \cup \ldots \cup \texttt{Prem}(A_n)$, $\texttt{Conc}(A) = \psi$, $\texttt{Sub}(A) = \texttt{Sub}(A_1) \cup \ldots \cup \texttt{Sub}(A_n) \cup \{A\}$, $\texttt{TopRule}(A) = \texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \to \psi$.
3. $[A_1], \ldots, [A_n] \Rightarrow \psi$ if $A_1, \ldots, A_n$ are arguments such that $\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \Rightarrow \psi \in \mathcal{R}$ with:
   $\texttt{Prem}(A) = \texttt{Prem}(A_1) \cup \ldots \cup \texttt{Prem}(A_n)$, $\texttt{Conc}(A) = \psi$, $\texttt{Sub}(A) = \texttt{Sub}(A_1) \cup \ldots \cup \texttt{Sub}(A_n) \cup \{A\}$, $\texttt{TopRule}(A) = \texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \Rightarrow \psi$.

Each of these functions Func are also defined on sets of arguments $S = \{A_1, \ldots, A_n\}$ as follows: $\texttt{Func}(S) = \texttt{Func}(A_1) \cup \ldots \cup \texttt{Func}(A_n)$.

Arguments can be attacked in three ways: on an application of a defeasible rule, on the conclusion of such an application or on an ordinary premise.

Definition 4 [**Attack**] An argument A *attacks* an argument B iff A *undercuts* or *rebuts* or undermines B, where:

- A *undercuts* B (on B') iff $\texttt{Conc}(A) = -n(r)$ and $B' \in \texttt{Sub}(B)$ such that B''s top rule r is defeasible.
- A *rebuts* B (on B') iff $\texttt{Conc}(A) = -\varphi$ for some $B' \in \texttt{Sub}(B)$ of the form $B''_1, \ldots, B''_n \Rightarrow \varphi$.

[2] The square brackets make the presentation of examples more concise. They will be omitted if there is no danger for confusion.

- A *undermines* B (on φ) iff $\texttt{Conc}(A) = -\varphi$ for some $\varphi \in \texttt{Prem}(B) \cap \mathcal{K}_p$.

The *ASPIC*$^+$ counterpart of an abstract argumentation framework is a structured argumentation framework.

Definition 5 [**Structured Argumentation Frameworks**] Let AT be an *argumentation theory*, that is, a pair $(AS, \mathcal{K})$. A *structured argumentation framework (*SAF*)* defined by AT, is a triple $\langle \mathcal{A}, \mathcal{C}, \preceq \rangle$ where $\mathcal{A}$ is the set of all arguments on the basis of $\mathcal{K}$ in AS, $\preceq$ is an ordering on $\mathcal{A}$, and $(X, Y) \in \mathcal{C}$ iff X attacks Y.

The notion of *defeat* is then defined as follows ($A \prec B$ is defined as usual as $A \preceq B$ and $B \npreceq A$ and $A \approx B$ as $A \preceq B$ and $B \preceq A$).

Definition 6 [**Defeat**] A *defeats* B iff either A undercuts B; or A rebuts or undermines B on B' and $A \nprec B'$.

Abstract argumentation frameworks are then generated from $SAFs$ by letting the attacks from an AF be the defeats from a SAF.

Definition 7 [**Argumentation frameworks**] An *abstract argumentation framework* (AF) *corresponding to a* SAF = $\langle \mathcal{A}, \mathcal{C}, \preceq \rangle$ (where $\mathcal{C}$ is *ASPIC*$^+$'s attack relation) is a pair $(\mathcal{A}, attack)$ such that $attack$ is the defeat relation on $\mathcal{A}$ determined by SAF.

A nonmonotonic consequence notion can then be defined as follows. Let $T \in$ {complete, preferred, grounded, stable} and let $\mathcal{L}$ be from the AT defining SAF. A wff $\varphi \in \mathcal{L}$ is *sceptically* T*-justified* in SAF if φ is the conclusion of a sceptically T-justified argument, and *credulously* T*-justified* in SAF if φ is not sceptically T-justified and is the conclusion of a credulously T-justified argument.

3. Directly Encoding Examples as Abstract Argumentation Frameworks

In much recent work on extensions of AFs with new elements, the extended theory is motivated by natural-language examples of argumentation that are directly translated into AFs instead of through a theory of the nature of arguments and attacks. The danger of this is that the resulting AFs are ad-hoc modellings, so that the observations made about the resulting AFs have no general validity. In this section this danger is illustrated with discussions of proposals concerning bipolar argumentation frameworks (Section 3.1) and probabilistic abstract argumentation (Section 3.2). The discussion of the bipolar example is based on [7] while the discussion of the probabilistic example is fully our own.

3.1. Abstract Support Relations

In [5] the following example is used for concluding that Dung-style AFs cannot represent a certain type of support relation between arguments.

Example 1 [5] We want to begin a hike. We prefer a sunny weather, then a sunny and cloudy one, then a cloudy but not rainy one, in this order. We will cancel the hike only if the weather is rainy. But clouds could be a sign of rain. We look at the sky early in the morning. It is cloudy. The following exchange of informal arguments occurs between Tom, Ben and Dan:

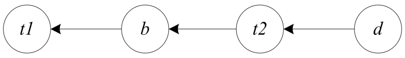

Figure 1. [5]'s AF modelling of Example 1.

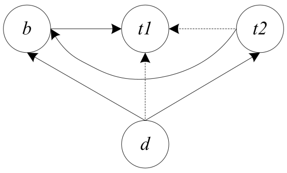

Figure 2. [5]'s BAF modelling of Example 1.

t_1: Today we have time, we begin a hike.
b: The weather is cloudy, clouds are a sign of rain, we had better cancel the hike.
t_2: These clouds are early patches of mist, the day will be sunny, without clouds, so the weather will be not cloudy (and we can begin the hike).
d: These clouds are not early patches of mist, so the weather will be not sunny but cloudy; however these clouds will not grow, so it will not rain (and we can begin the hike).

In [5] this is modelled as the AF of Figure 1 and then discussed in terms of a notion of defence that generalises Dung's notions of defence and attack (by [5] called 'defeat') as follows: argument A *indirectly defends (defeats) argument* B iff the AF contains an odd-length (even-length) chain beginning with A and ending with B. According to this definition, t_2 indirectly defends t_1, while d indirectly defeats it t_1. According to [5] this is counterintuitive since both t_2 and d have the same conclusion. They conclude:

> So, the idea of a chain of arguments and counterarguments in which we just have to count the links and take the even one as defeaters and the odd ones as supporters is an oversimplification. So, the notion of defence proposed by [8] is not sufficient to represent support.

In [5] this analysis is then (with other examples) used to motivate semantics for so-called bipolar argumentation frameworks (BAF), which add abstract support relations between arguments to Dung's AFs. Example 1 is then remodelled by adding an attack relation from d to b and support relations from d to t_1 and from t_2 to t_1, resulting in the BAF of Figure 2 (in which dashed arrows depict support relations). Then [5]'s semantics for $BAFs$ yield $\{d, t_1\}$ as the unique extension, so $hike$ is skeptically acceptable.

As shown by [7], it turns out that the following arguably principled formalisation in *ASPIC*$^+$ yields a different AF, in which the problems noted by [5] do not arise. Let $\mathcal{K} = \mathcal{K}_n = \{cloudy, rain, \neg grow\}$ and $\mathcal{R} = \mathcal{R}_d = \{time \Rightarrow hike; sunny \Rightarrow hike; drycloudy \Rightarrow hike; rain \Rightarrow \neg hike; cloudy \Rightarrow rain; cloudy \Rightarrow mist; mist \Rightarrow sunny; cloudy \Rightarrow \neg mist; cloudy, \neg grow \Rightarrow drycloudy\}$. Furthermore, to keep the formalisation concise, we encode some some conflict relations between elements of $\mathcal{L}$ as contrariness relations. Alternatively, they can be encoded with strict rules and negation.

$rain = -drycloudy, drycloudy = -rain,$
$sunny = -drycloudy, drycloudy = -sunny,$
$rain = -sunny, sunny = -rain$

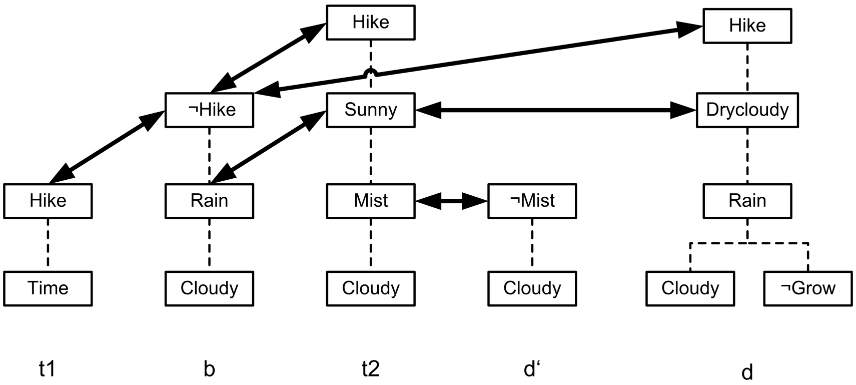

Figure 3. An *ASPIC*$^+$ modelling of Example 1.

This yields the following arguments (visualised in Figure 3):

t_1: $time \Rightarrow hike$
b: $[cloudy \Rightarrow rain] \Rightarrow \neg hike$
t_2: $[[cloudy \Rightarrow mist] \Rightarrow sunny] \rightarrow hike$
d_2': $[[cloudy \Rightarrow \neg mist]$
d: $[[cloudy, \neg grow \Rightarrow rain] \Rightarrow drycloudy] \Rightarrow hike$

Note that the sets $\{b, d\}$, $\{b, t_2\}$ and $\{t_2, d\}$ are not admissible while $\{t_1, t_2\}$ and $\{t_1, d\}$ are admissible. Moreover, the AF has several preferred extensions, some of which contain t_1 and d but not b and t_2, others contain b and d but not t_1 and t_2 and yet others contain t_1 and t_2 but not b and d. So both $hike$ and $\neg hike$ are credulously but not skeptically acceptable. This arguably is a better outcome than in [5]'s BAF, since there is an unresolved conflict between subarguments of d, b and t_2 concerning whether it will be rainy, sunny or dry cloudy. Moreover, in this modelling d does not indirectly defeat t_1. Note that $\{t_1, d\}$ is admissible, unlike in [5]'s AF of the example. In conclusion, in the AF generated by the *ASPIC*$^+$ modelling of the example the problem noted by [5] in their AF directly generared from the natural-language example does not arise.

3.2. Probabilistic Abstract Argumentation

Another field of study where natural-language examples are often directly encoded in abstract AFs is probabilistic abstract argumentation, in which Dung's AFs are extended with probability functions on (sets of) arguments. We will analyse a proposal by Hunter [10] in which the probability of an argument is the degree to which the argument is true, which is equated with the probability that the conjunction of all its premises is true. Among other things, Hunter discusses the following medical diagnosis example.

Example 2 [10]

A_1: From these symptoms, the patient has a cold
A_2: Influenza is an option as a diagnosis for this patient, since it is currently very common.

Where in Hunter's modelling A_2 asymmetrically attacks A_1.

Hunter then assigns probability 0.9 to A_1 and probability 0.1 to A_2. He gives this example as part of an attempt to argue that it makes sense to consider AFs in which one argument asymmetrically attacks another but still has lower probability.

But why is the attack asymmetric? One way to interpret the example is to say that A_2 blocks the inference from A_1's premise to its conclusion. In *ASPIC*$^+$ this can be modelled as an undercutting attack, for instance, as follows (for reasons of space and ease of presentation the representation is left semiformal):

A'_1: The patient has these symptoms, patients that have these symptoms usually have a cold ⇒ the patient has a cold.

A'_2: The patient has these symptoms, influenza is common these days, if influenza is common then for patients with these symptoms influenza is an option as a diagnosis ⇒ influenza is an option as a diagnosis for this patient.

Here A'_1 is taken to use a defeasible inference rule 'φ, usually if φ then $\psi \Rightarrow \psi$' and the conclusion of A'_2 is taken to undercut this rule. Then [10]'s attack graph is indeed obtained (leaving the arguments' proper subarguments implicit). But in the premises approach to argument probability it makes no sense to assign probability 0.9 to A'_1 and probability 0.1 to A'_2, since both have certainly true premises. In particular, it is definitely true that patients that have these symptoms *usually* have a cold.

But perhaps the arguments can be rephrased in classical argumentation as deductively valid arguments with uncertain premises as follows:

A''_1: The patient has these symptoms, patients that have these symptoms have a cold → the patient has a cold, therefore, influenza is not an option as a diagnosis for this patient.

A''_2: The patient has these symptoms, influenza is common these days, if influenza is common then for patients with these symptoms influenza is an option as a diagnosis → influenza is an option as a diagnosis for this patient.

Can the probabilities now be assigned to the last premises of these arguments interpreted as material implications? This seems impossible, since the probabilities are arguably best interpreted as conditional probabilities of having a cold given these symptoms and having influenza given these symptoms and the fact that influenza is currently common. And it is well-known that the conditional probability of Q given P is not equivalent to the unconditional probability of the material implication $P \supset Q$. For instance, the latter contraposes while the former does not.

Let us next try to interpret the last premises directly as statements of conditional probability. Then Pollock's [14]'s *statistical syllogism* can in *ASPIC*$^+$ be employed as a defeasible inference rule: a is an F, the probability of being a G given being an F is x ⇒ the probability of a being a G is x.

A'''_1: The patient has these symptoms, 90% of the patients that have these symptoms have a cold ⇒ this patient has a cold.

A'''_2: The patient has these symptoms, influenza is common these days, when influenza is common then 10% of the patients with these symptoms has influenza ⇒ this patient has influenza.

Here having a cold and having influenza are contradictories. Note first that this is not fully according to Pollock's treatment of the statistical syllogism, since that does not apply if $x \leq 0.5$. Let us nevertheless allow the rule in this case also. Then two further things should be noted. First, the arguments' conclusions are contradictories so the arguments rebut each other, so the attack relation in Hunter's graph cannot equate *ASPIC*$^+$'s rebutting attack. The only way to make it asymmetric is to interpret it as defeat. And this makes sense, since the second conditional probability is about a more specific class then the first, so Pollock's subproperty defeater yields that A_2''' undercuts A_1'''. However, this is not all, since the second conditional probability implies that when influenza is common then 90% of the patients with these symptoms has no influenza, so the statistical syllogism also gives rise to a third argument:

A_3: The patient has these symptoms, influenza is common these days, when influenza is common then 90% of the patients with these symptoms has no influenza $\Rightarrow$ this patient has no influenza.

And on any reasonable account of argument strength A_3 asymmetrically defeats A_2'''.

The upshot of all this is that it seems impossible to give a principled modelling of Example 2 that yields [10]'s probabilistic AF. Either the AF is retained but [10]'s argument probabilities do not make sense, or a different AF is obtained. So this example fails to show that it makes sense to consider probabilistic AFs in which a weaker argument asymmetrically attacks a stronger one. The underlying reason is arguably the reluctance to consider interpretations of these arguments as defeasible arguments or more generally the reluctance to analyse this example in terms of a theory of argumentation with defeasible generalisations.

4. Abstracting Away from the Context: Degrees of Acceptability of Arguments

We next illustrate the danger of abstracting from the context that gave rise to an AF. We will do so with a discussion of recent attempts to define gradual notions of argument acceptability in terms of the topology of abstract argumentation frameworks. For instance, [9] refine the standard Dung semantics by formalising the following two intuitions:

A1: having fewer attackers is 'better' than having more.
A2: having more defenders is 'better' than having fewer.

Similar intuitions have been expressed by [1] in support of their ranking-based semantics of abstract argumentation. Consider the two AFs displayed in Figure 4. AF_1 has

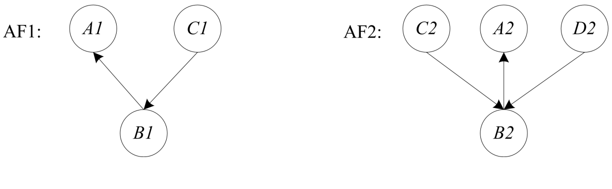

Figure 4. Two example AFs from [9]

the grounded extension $\{A_1, C_1\}$ while AF_2 has the grounded extension $\{A_2, C_2, D_2\}$, which are also the unique complete, stable and preferred extensions. So according to standard Dung's semantics A_1 and A_2 are both skeptically justified. By contrast, in [9]'s semantics A_2 is justified to a higher degree than A_1, since A_2 has two defenders while A_1 has only one defender.

However, this neglects possible differences in the nature of the arguments. For example, if C_1 is unattackable (e.g. strict and firm in *ASPIC*$^+$ or without assumptions in assumption-based argumentation) while C_2 and D_2 are attackable (e.g. defeasible or fallible in *ASPIC*$^+$ or with assumptions in assumption-based argumentation) then A_1 is arguably better justified than A_2, since A_1's defender can never be attacked while A_2's defenders can be attacked. In other words, the gradual semantic semantics proposed by [9] is based on the implicit assumptions that there is no difference in attackability of the arguments. But this assumption is not generally valid, so this is another case of the danger of abstracting from the nature of the arguments in an AF. (Note that this criticism does not apply to Dung's original semantics).

[9] argue for their proposal on further grounds, namely, that their semantics is supported by experimental findings of [17] that humans have higher confidence in the claims of arguments that are unattacked, than when these arguments are subsequently attacked and then defended. Here they refer to the two AFs as displayed in Figure 5. In [17]'s experiments the subjects were first confronted with a single argument, for instance:

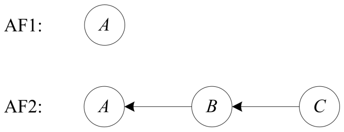

Figure 5. The reinstatement pattern

> A: *The battery of Alex's car is not working. Therefore, Alex's car will halt.*

They were then asked to rate their confidence in its conclusion. Only then were they subsequently confronted with an attacker and defender, for instance:

> B: *The battery of Alex's car has just been changed today. Therefore, the battery of Alex's car is working.*
> C: *The garage was closed today. Therefore, the battery of Alex's car has not been changed today.*

The subjects were then again asked to rate their confidence in the conclusion of the initial argument and it turned out that their average confidence was significantly lower than after being presented with A only (although significantly higher than after being presented with B but not yet with C).

For several reasons these findings cannot be used in support of the general claim that argument A in Figure 5 is more justified in AF_1 than in AF_2. To start with, the subjects were asked to give their degree of confidence in the conclusion of an argument, which is not obviously the same as the degree of justification of an argument. It may be that what the subjects were doing is better modelled as Bayesian updating in probabilistic networks than as considering degrees of justification in AFs.

Second, even if granted that the subjects were considering the latter problem, it is not obvious that the AFs they considered correspond to the ones of Figure 5. For example, if the example is reconstructed in *ASPIC*$^+$ by regarding all premises as ordinary ones and by assuming that all arguments employ defeasible inference rules, then the full set of arguments corresponds to AF_2' as shown in Figure 6, where Ap, Bp and Cp are the subarguments of, respectively, A, B and C consistening of their premise. Note that B and Ap attack each other since B undermines Ap (and A) while Ap rebuts B. Likewise for the other attacks. Note that unlike in AF_2, in AF_2' argument A is not skeptically justified. So there is ambiguity about how the test subjects may have interpreted the example. This is in fact another illustration of the problem discussed in Section 3 that directly formalising natural-language examples as AFs may result in ad-hoc modellings (or in this case in a modelling that is not the only possible one).

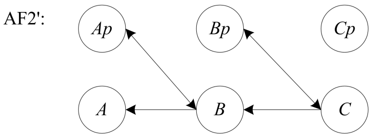

Figure 6. An alternative interpretation of [17]'s examples

Even granted that the subjects interpreted the examples as in Figure 5, the findings cannot be used as support for the general claim that fewer attackers make an argument more justified. The point is that this claim is more general than just about structures as in Figure 5, where A on the left and right refer to the same argument. By contrast, [9]'s claim also covers situations where A on the left and right refer to different arguments, but in [17]'s experiments no examples of this kind were shown to the test subjects.

Even for the restricted case of Figure 5 [17]'s findings do not support the claims of [9], since it is not obvious that the subject's degrees of confidence will remain the same if the arguments are presented to them in a different way. As suggested by [17], one possible explanation for the finding is that the second rating was on average lower than the first is that being confronted with the attacker increased the subject's degree of belief in other possible attackers (for instance, that the battery of Alex's car is old or dirty). If this explanation is true, then different results may be obtained if the examples are presented to the subjects in a different way, for instance, if first the entire theory from which the example arguments are drawn (in the present case a theory on the functioning of car batteries) is presented and only then the arguments are presented. In particular, it is conceivable that compared to the original experiments, the average rating before presenting the attacker and defender goes down while the average rating after presenting the attacker and defender will remain the same or will even go up. Note in the car battery example that before being presented with the attacker and defender the subject has received no evidence at all about whether the car battery was changed, while after being presented with the attacker and defender she has received evidence that the car battery was not changed. So if she was made aware from the start of all possible reasons why the battery could not be working, then her degree of confidence in the belief that the battery was not changed could have increased while her degree of confidence in all other reasons why the battery might not work remained the same. In that case her degree of belief in the conclusion of argument A would also have increased. Therefore, before additional experiments are conducted in which the arguments are presented to the subjects in different ways, [17]'s findings cannot be regarded as supporting the abstract semantics of [9] or similar semantics such as the one of [1]. This also illustrates a danger of abstracting from the context of the argumentation, namely, that observations valid for particular contexts are without supporting evidence presented as valid in general.

These observations can also be explained in Bayesian terms. Consider the following arguments:

A: *Tweety is a bird. Therefore, Tweety can fly.*
B: *Tweety is a penguin, since John says so. Therefore, Tweety cannot fly*
C: *This camera footage shows that Tweety is not a penguin. Therefore, Tweety is not a penguin*

If the subject is first shown a theory of the flying abilities of birds, then she may have formed some prior degree of belief that birds are not penguins, which she applies to Tweety after hearing argument A. After subsequently being confronted with evidence about whether Tweety is a penguin in the form of arguments B and C, her degree of belief that Tweety is not a penguin may well have increased, so her degree of belief that Tweety can fly may have increased.

More generally, a problem with reasoning experiments like these is that it is often very hard to make the subjects stick to the information that was explicitly given; often the subjects will, either implicitly or explicitly, also take other beliefs and background information into account.

The observations about the importance of the context of argumentation can also be given a normative twist. Consider an application in which the two AFs in Figure 5 belong to two different stages in a testing process of a hypothesis, where a test takes the form of searching for a possible counterargument. Then AF_1 reflects the stage in which no test has yet been carried out while AF_2 reflects the stage in which one test has been carried out and argument A has passed the test in that counterargument B could be refuted. In such a context of application it seems rational to say that A is better justified in AF_1 than in AF_2 since it has passed more tests. A possible philosophical foundation of this approach can be found in Cohen's [6] theory of Baconian probability.

More generally, it can be concluded that a basic assumption underlying much recent work on degrees of justification lacks sufficient justification. This is not to say that the idea of degrees of justification makes no sense but only that this idea is arguably better developed while taking the nature of arguments and their attack relations and the context in which they are put forward into account.

5. Conclusion

In this paper we have discussed and illustrated some dangers of studying or manipulating abstract argumentation frameworks without regarding the nature of the arguments or attacks in the framework or the context that gave rise to the framework. It should be noted that in realistic applications of argumentation the original arguments will always be available for inspection. In reality, there are no abstract arguments: how could they else be recognised as arguments? Moreover, in realistic applications the context in which the arguments were put forward will always be known. For these reasons, focussing on abstract AFs is not a matter of dealing with incomplete information but instead of deliberately ignoring available information. But in evaluating arguments the appropriate level of abstraction cannot be determined before inspecting the arguments and the context in which they were put forward. This does not mean that abstract argumentation frameworks or their extensions and refinements should be abandoned. They can still be a very useful component in models of argumentation, provided that they are combined with principled accounts of the nature of arguments and their attack relations and (when relevant) of the contexts in which argumentation takes place.

References

[1] L. Amgoud and J. Ben-Naim. Ranking-based semantics for argumentation frameworks. In Liu W., Subrahmanian V.S., and Wijsen J., editors, *Scalable Uncertainty Management. SUM 2013*, number 8078 in Springer Lecture Notes in Computer Science, pages 134–147, Berlin, 2013. Springer Verlag.

[2] M. Caminada. Rationality postulates: applying argumentation theory for nonmonotonic reasoning. *IfColog Journal of Logics and Their Applications*, 8:2707 – 2733, 2017. Also to appear in Handbook of Formal Argumentation, College Publications, London.

[3] M. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171:286–310, 2007.

[4] M. Caminada and Y. Wu. On the limitations of abstract argumentation. In *Proceedings of the 23rd Benelux Conference on Artificial Intelligence (BNAIC-11)*, Gent, Belgium, 2011.

[5] C. Cayrol and M.-C. Lagasquie-Schiex. Bipolar abstract argumentation systems. In I. Rahwan and G.R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 65–84. Springer, Berlin, 2009.

[6] L.J. Cohen. *The Probable and the Provable*. Clarendon Press, Oxford, 1977.

[7] M. de Winter. Analysis and formal modelling of support relations in argument systems. Master's thesis, Cognitive Artificial Intelligence, Utrecht University, Utrecht, 2014.

[8] 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.

[9] D. Grossi and S. Modgil. On the graded acceptability of arguments. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 868–874, 2015.

[10] A. Hunter. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54:47–81, 2013.

[11] S. Modgil. Revisiting abstract argumentation. In E. Black, S. Modgil, and N. Oren, editors, *Second International Workshop, TAFA 2013, Beijing, China, August 3-5, 2013, Revised Selected papers*, number 8306 in Springer Lecture Notes in AI, pages 1–15, Berlin, 2014. Springer Verlag.

[12] S. Modgil and H. Prakken. Resolutions in structured argumentation. In B. Verheij, S. Woltran, and S. Szeider, editors, *Computational Models of Argument. Proceedings of COMMA 2012*, pages 310–321. IOS Press, Amsterdam etc, 2012.

[13] S. Modgil and H. Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument and Computation*, 5:31–62, 2014.

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

[15] H. Prakken. Some reflections on two current trends in formal argumentation. In *Logic Programs, Norms and Action. Essays in Honour of Marek J. Sergot on the Occasion of his 60th Birthday*, pages 249–272. Springer, Berlin/Heidelberg, 2012.

[16] H. Prakken. Historical overview of formal argumentation. *IfColog Journal of Logics and Their Applications*, 8:2183 – 2262, 2017. Also to appear in Handbook of Formal Argumentation, College Publications, London.

[17] I. Rahwan, M.I. Madakkatel, J.-F. Bonnefon, R.N. Awan, and S. Abdallah. Behavioural experiments for assessing the abstract semantics of reinstatement. *Cognitive Science*, 34:1483–1502, 2010.

[18] W. Twining. Necessary but dangerous? Generalisations and narrative in argumentation about "facts" in criminal process. In M. Malsch and F. Nijboer, editors, *Complex Cases. Perspectives on the Netherlands Criminal Justice System*, pages 69–98. Thela Thesis, Amsterdam, 1999.