Spurious preferences in structured argumentation: a preliminary analysis

Pietro Baroni[1,*], Federico Cerutti[1] and Massimiliano Giacomin[1]

[1]DII - University of Brescia, Italy

Abstract

The phenomenon of spurious preferences in argumentation can be described as an unjustified unequal treatment of some arguments emerging in a context where equal treatment would be expected. Using the *ASPIC*$^+$ formalism as a basis, we provide an emblematic example of spurious preference and introduce a basic requirement of spurious preference avoidance for a suitable family of argumentation theories. We then show that a variant of *ASPIC*$^+$, introduced to deal with problems concerning multiple contradictories, satisfies this requirement.

Keywords

Structured argumentation, Preferences, Justification status

1. Introduction

The phenomenon of spurious preferences in argumentation can be intuitively described as an unjustified unequal treatment of some arguments, with respect to justification status evaluation, which emerges as a byproduct of the formal representation adopted in a reasoning context where an equal treatment would be expected. This paper is devoted to a preliminary analysis of this phenomenon. We first exemplify how it may occur in the context of the well-known *ASPIC*$^+$ formalism and then show that, for a given family of argumentation theories, it is avoided by a variant of *ASPIC*$^+$, called *ASPIC*^{R}, introduced in [1]. The paper is organised as follows. After recalling the necessary background in Section 2, we illustrate the phenomenon of spurious preferences in Section 3. We then introduce a basic requirement of spurious preference avoidance in Section 4, and show in Section 5 that *ASPIC*^{R} satisfies it. Section 6 concludes.

2. Background

We briefly review Dung's theory of argumentation frameworks.

Definition 1. *An argumentation framework (AF) is a pair* $\mathcal{F} = \langle A, \twoheadrightarrow \rangle$*, where* A *is a set of arguments and* $\twoheadrightarrow \subseteq (A \times A)$ *is a binary relation on* A*.*

When $(\alpha, \beta) \in \twoheadrightarrow$ (also denoted as $\alpha \twoheadrightarrow \beta$) we say that α attacks β. For a set $X \subseteq A$ and an argument $\alpha \in A$ we write $\alpha \twoheadrightarrow X$ if $\exists \beta \in X : \alpha \twoheadrightarrow \beta$ and $X \twoheadrightarrow \alpha$ if $\exists \beta \in X : \beta \twoheadrightarrow \alpha$, and we denote the arguments attacking X as $X^- \triangleq \{\alpha \in A \mid \alpha \twoheadrightarrow X\}$ and the arguments attacked by X as $X^+ \triangleq \{\alpha \in A \mid X \twoheadrightarrow \alpha\}$. An *extension-based* argumentation semantics σ specifies the criteria for identifying, for a generic AF, a set of *extensions*, where each extension is a set of arguments considered to be acceptable together. Given a generic argumentation semantics σ, the set of extensions prescribed by σ for a given AF $\mathcal{F}$ is denoted as $\mathcal{E}_\sigma(\mathcal{F})$. Several argumentation semantics are recalled in Definition 2, along with some basic underlying notions. For more details, the reader is referred to [2].

AI^3 *2024 - 8th Workshop on Advances in Argumentation in Artificial Intelligence*

*Corresponding author.

pietro.baroni@unibs.it (P. Baroni); federico.cerutti@unibs.it (F. Cerutti); massimiliano.giacomin@unibs.it (M. Giacomin)

0000-0001-5439-9561 (P. Baroni); 0000-0003-0755-0358 (F. Cerutti); 0000-0003-4771-4265 (M. Giacomin)

Definition 2. *Let $\mathcal{F} = \langle A, \twoheadrightarrow \rangle$ be an AF, $\alpha \in A$ and $X \subseteq A$. X is* conflict-free*, denoted as $X \in \mathcal{E}_{\mathrm{CF}}(\mathcal{F})$, iff $X \cap X^- = \emptyset$. α is* acceptable *with respect to X (or α is defended by X) iff $\{\alpha\}^- \subseteq X^+$. The function $F_{\mathcal{F}} : 2^A \to 2^A$ which, given a set $X \subseteq A$, returns the set of the acceptable arguments with respect to X, is called the* characteristic function *of $\mathcal{F}$. X is* admissible *(denoted as $X \in \mathcal{E}_{\mathrm{AD}}(\mathcal{F})$) iff $X \in \mathcal{E}_{\mathrm{CF}}(\mathcal{F})$ and $X \subseteq F_{\mathcal{F}}(X)$. X is a* complete *extension (denoted as $X \in \mathcal{E}_{\mathrm{CO}}(\mathcal{F})$) iff $X \in \mathcal{E}_{\mathrm{CF}}(\mathcal{F})$ and $X = F_{\mathcal{F}}(X)$. X is the* grounded *extension (denoted as $X = GR(\mathcal{F})$ or $X \in \mathcal{E}_{\mathrm{GR}}(\mathcal{F})$) iff X is the least fixed point of $F_{\mathcal{F}}$ (equivalently, the least complete extension). X is a* preferred *extension (denoted as $X \in \mathcal{E}_{\mathrm{PR}}(\mathcal{F})$) iff X is a maximal (with respect to set inclusion) admissible set. X is a* stable *extension (denoted as $X \in \mathcal{E}_{\mathrm{ST}}(\mathcal{F})$) iff $X^+ = A \setminus X$. X is a* semi-stable *extension (denoted as $X \in \mathcal{E}_{\mathrm{SST}}(\mathcal{F})$) iff it is a complete extension such that $X \cup X^+$ is maximal (wrt $\subseteq$) among all complete extensions.*

Argument justification status is defined on the basis of extension membership.

Definition 3. *Given a set S a justification labeling of S is a function $J : S \to \Sigma_J$, where $\Sigma_J = \{Sk, Cr, No\}$. Given an AF $\mathcal{F} = \langle A, \twoheadrightarrow \rangle$ and a semantics σ, the justification labeling of A according to σ is defined as follows*[1]*: $J(\alpha) = Sk$ iff $\alpha \in \bigcap_{\mathsf{E} \in \mathcal{E}_\sigma(\mathcal{F})} \mathsf{E}$; $J(\alpha) = Cr$ iff $\alpha \in \bigcup_{\mathsf{E} \in \mathcal{E}_\sigma(\mathcal{F})} \mathsf{E}$ and $\alpha \notin \bigcap_{\mathsf{E} \in \mathcal{E}_\sigma(\mathcal{F})} \mathsf{E}$; $J(\alpha) = No$ iff $\alpha \notin \bigcup_{\mathsf{E} \in \mathcal{E}_\sigma(\mathcal{F})} \mathsf{E}$.*

In words, we will say respectively that α is skeptically justified, credulously justified, and not justified. We now recall the essential notions of the *ASPIC*$^+$ formalism.

Definition 4. *An* argumentation system *is a tuple $AS = (\mathcal{L}, ^-, \mathcal{R}, n)$ where:*

1. *$\mathcal{L}$ is a logical language*
2. *$^-$ is a* contrariness *function from $\mathcal{L}$ to $2^{\mathcal{L}}$ such that: (i) φ is a* contrary *of ψ if $\varphi \in \overline{\psi}$, $\psi \notin \overline{\varphi}$; (ii) φ is a* contradictory *of ψ (denoted by $\varphi = -\psi$) if $\varphi \in \overline{\psi}$, $\psi \in \overline{\varphi}$; (iii) each $\varphi \in \mathcal{L}$ has at least one contradictory*
3. *$\mathcal{R} = (\mathcal{R}_S, \mathcal{R}_D)$ is a pair of sets 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 $\mathcal{L}$), and $\mathcal{R}_S \cap \mathcal{R}_D = \emptyset$*
4. *$n : \mathcal{R}_D \to \mathcal{L}$ is a naming convention for $\mathcal{R}_D$.*

In the following, given a set $S \subseteq \mathcal{L}$ with a little abuse of notation we will denote the set of its contraries and contradictories as $\overline{S} = \bigcup_{\varphi \in S} \{\psi \mid \psi \in \overline{\varphi}\}$. Given a rule $r = \varphi_1, \ldots, \varphi_n \to (\Rightarrow)\varphi$, we will say that φ is the consequent of the rule, denoted as $cons(r)$ and that $\{\varphi_1, \ldots, \varphi_n\}$ is the set of the antecedents of the rule denoted as $ant(r)$.

Closure under transposition of strict rules is a desirable property as it ensures (together with other conditions) that an argumentation system satisfies some *rationality postulates* [3] (we do not recall the relevant details as not necessary for this paper).

Definition 5. *Given an argumentation system $AS = (\mathcal{L}, ^-, \mathcal{R}, n)$, the set of strict rules $\mathcal{R}_S$ is* closed under transposition *iff if $\varphi_1, \ldots, \varphi_n \to \psi \in \mathcal{R}_S$ then, for $i = 1 \ldots n$, $\varphi_1, \ldots, \varphi_{i-1}, -\psi, \varphi_{i+1}, \ldots, \varphi_n \to -\varphi_i \in \mathcal{R}_S$. Given a set of strict rules $\mathcal{R}_S$, its closure under transposition is defined as $Cl_{tr}(\mathcal{R}_S) \triangleq \mathcal{R}_S \cup \bigcup_{r \in \mathcal{R}_S} tr(r)$, where for any $r = \varphi_1, \ldots, \varphi_n \to \psi, tr(r) = \bigcup_{i=1\ldots n} \{\varphi_1, \ldots, \varphi_{i-1}, -\psi, \varphi_{i+1}, \ldots, \varphi_n \to -\varphi_i\}$.*

A knowledge base is a subset of $\mathcal{L}$ including certain (called *axioms*) and defeasible (called *ordinary*) premises. It gives rise to the notion of argumentation theory. Arguments are built from a knowledge base using rules.

[1] We avoid reference to $\mathcal{F}$ and σ in J for ease of notation. Moreover, with respect to the traditional notion of justification we keep skeptical and credulous justification disjoint for reasons which will be clear later.

Definition 6. *A* knowledge base *in an argumentation system $AS = (\mathcal{L}, ^-, \mathcal{R}, n)$ 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*). The tuple $AT = (AS, \mathcal{K})$ is called an* argumentation theory*.*

Definition 7. *An argument α on the basis of a knowledge base $\mathcal{K}$ in an argumentation system $(\mathcal{L}, ^-, \mathcal{R}, n)$ is:*

1. *φ if $\varphi \in \mathcal{K}$ with: $Prem(\alpha) = \{\varphi\}$; $Conc(\alpha) = \varphi$; $Sub(\alpha) = \{\varphi\}$; $Rules(\alpha) = \emptyset$; $Top(\alpha) =$ undefined.*
2. *$\alpha_1, \ldots, \alpha_n \rightarrow (\Rightarrow)\ \psi$ if $\alpha_1, \ldots, \alpha_n$ are arguments such that there exists a strict (defeasible) rule $Conc(\alpha_1), \ldots, Conc(\alpha_n) \rightarrow (\Rightarrow)\ \psi$ in $\mathcal{R}_S$ ($\mathcal{R}_D$) with: $Prem(\alpha) = Prem(\alpha_1) \cup \ldots \cup Prem(\alpha_n)$; $Conc(\alpha) = \psi$; $Sub(\alpha) = Sub(\alpha_1) \cup \ldots \cup Sub(\alpha_n) \cup \{\alpha\}$; $Rules(\alpha) = Rules(\alpha_1) \cup \ldots \cup Rules(\alpha_n) \cup \{Conc(\alpha_1), \ldots, Conc(\alpha_n) \rightarrow (\Rightarrow)\ \psi\}$; $Top(\alpha) = Conc(\alpha_1), \ldots, Conc(\alpha_n) \rightarrow (\Rightarrow)\ \psi$; $DefRules(\alpha) = \{r \mid r \in Rules(\alpha) \cap \mathcal{R}_D\}$; $StRules(\alpha) = \{r \mid r \in Rules(\alpha) \cap \mathcal{R}_S\}$.*

For any argument α, $Prem_n(\alpha) = Prem(\alpha) \cap \mathcal{K}_n$; $Prem_p(\alpha) = Prem(\alpha) \cap \mathcal{K}_p$. α is: strict *if $DefRules(\alpha) = \emptyset$,* defeasible *if $DefRules(\alpha) \neq \emptyset$;* firm *if $Prem(\alpha) \subseteq \mathcal{K}_n$;* plausible *if $Prem(\alpha) \not\subseteq \mathcal{K}_n$;* finite *if $Rules(\alpha)$ is finite.*

Notation 1. *Some further notations are useful. Given $S \subseteq \mathcal{L}$, $S \vdash \varphi$ denotes that there exists a strict argument α such that $Conc(\alpha) = \varphi$, with $Prem(\alpha) \subseteq S$. $S \vdash_{min} \varphi$ denotes that $S \vdash \varphi$ and $\nexists T \subsetneq S : T \vdash \varphi$. Given a set of arguments X, $Prem(X) \triangleq \bigcup_{\alpha \in X} Prem(\alpha)$, and similarly for $Conc(X)$, $Sub(X)$, $Rules(X)$, $Top(X)$, $DefRules(X)$, $StRules(X)$.*

Three kinds of attack between arguments are considered.

Definition 8. *An argument α* attacks *an argument β iff α* undercuts*,* rebuts*, or* undermines *β where: α* undercuts *β (on β') iff $Conc(\alpha) \in \overline{n(r)}$ for some $\beta' \in Sub(\beta)$ such that $r = Top(\beta')$ is defeasible. α* rebuts *β (on β') iff $Conc(\alpha) \in \overline{\varphi}$ for some $\beta' \in Sub(\beta)$ of the form $\beta''_1, \ldots, \beta''_n \Rightarrow \varphi$. In such a case α* contrary-rebuts *β iff $Conc(\alpha)$ is a contrary of φ. α* undermines *β (on β') iff $Conc(\alpha) \in \overline{\varphi}$ for some $\beta' = \varphi$, $\varphi \in Prem_p(\beta)$. In such a case α* contrary-undermines *β iff $Conc(\alpha)$ is a contrary of φ.*

In some cases, attack effectiveness depends on a preference ordering $\preceq$ over arguments (assumed to be a preorder as in [4]). As usual $\alpha \prec \beta$ iff $\alpha \preceq \beta$ and $\beta \not\preceq \alpha$; $\alpha \simeq \beta$ iff $\alpha \preceq \beta$ and $\beta \preceq \alpha$. Effective attacks give rise to defeat.

Definition 9. *Let α attack β on β'. If α undercuts, contrary-rebuts, or contrary-undermines β on β', then α* preference-independent *attacks β on β', otherwise α* preference-dependent *attacks β on β'. Then, α* defeats *β iff for some β' either α* preference-independent *attacks β on β' or α* preference-dependent *attacks β on β' and $\alpha \not\prec \beta'$.*

Then a *structured argumentation framework* (SAF) can be defined from an argumentation theory,[2] using the attack relation. Using the defeat relation, an argumentation framework is then derived from a SAF.

Definition 10. *Let $AT = (AS, \mathcal{K})$ be an argumentation theory. A* structured argumentation framework *(SAF), defined by AT is a triple $(\mathsf{S}, \mathsf{C}, \preceq)$ where S is the set of all finite arguments constructed from $\mathcal{K}$ in AS (called the set of arguments on the basis of AT), $\mathsf{C} \subseteq \mathsf{S} \times \mathsf{S}$ is such that $(\alpha, \beta) \in \mathsf{C}$ iff α attacks β, and $\preceq$ is an ordering on S.*

Definition 11. *Let $\Delta = (\mathsf{S}, \mathsf{C}, \preceq)$ be a SAF, and $\mathsf{D} \subseteq \mathsf{S} \times \mathsf{S}$ be the defeat relation according to Definition 9. The AF corresponding to Δ is defined as $\mathcal{F}_\Delta = (\mathsf{S}, \mathsf{D})$.*

Given an argumentation semantics σ, the justification status of arguments in S according to σ is determined by the set of extensions $\mathcal{E}_\sigma(\mathcal{F}_\Delta)$ according to Definition 3.

[2] We do not consider here the alternative notion of c-structured argumentation framework in [5].

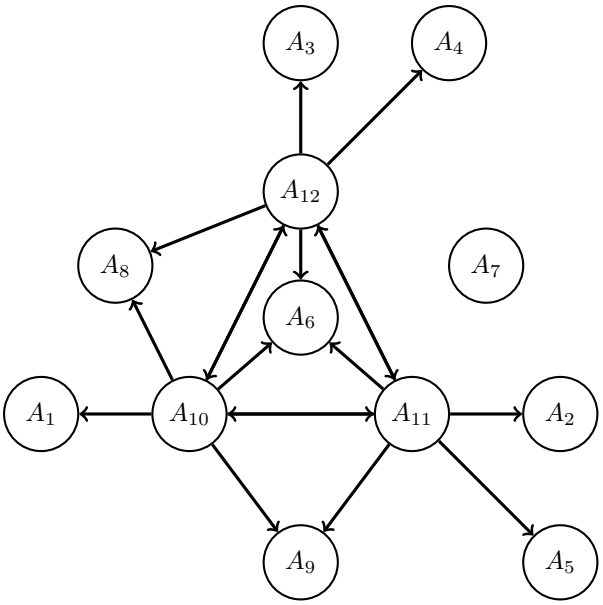

Figure 1: The argumentation framework corresponding to the first version of the taxpayers' list example.

3. Uncovering spurious preferences

To illustrate the phenomenon of spurious preferences, let us consider the following example (where we assume a simple language consisting just of a set of symbols and their negation). You have some uncertain evidence about the birthdate (b), birthplace (p) and domicile (d) of a person. From the known birthdate, it can be derived with certainty that the person is over 18 (m), and from the known birthplace it can be derived with certainty that the person is a citizen of a given country (c) (because *ius soli* holds in the country where the birthplace is located). Finally, from the domicile, age majority, and citizenship, it can be derived that the person must be included in the taxpayers' list (ω). However, it turns out with certainty that the person is not included in the list ($\neg\omega$). Thus, some of the uncertain evidence has to be rejected. In the absence of any preference among them, the three pieces of evidence are equal candidates to be retracted, and this is, in fact, what happens with the formalization of the example that we describe below.

The three evidences are represented by ordinary premises $\mathcal{K}_p = \{d, b, p\}$ while the certain fact is represented by an axiom $\mathcal{K}_n = \{\neg\omega\}$ and the domain knowledge is represented through strict rules $\mathcal{R}_{Sb} = \{b \to m; p \to c; d, c, m \to \omega\}$.

Closure under transposition (which is required to ensure the satisfaction of rationality postulates) leads to consider the following additional set of strict rules: $\mathcal{R}_{St} = \{\neg m \to \neg b; \neg c \to \neg p; d, c, \neg\omega \to \neg m; d, m, \neg\omega \to \neg c; c, m, \neg\omega \to \neg d\}$, thus $\mathcal{R}_S = \mathcal{R}_{Sb} \cup \mathcal{R}_{St}$.

The following arguments are then built: $A_1 = d$; $A_2 = b$; $A_3 = p$; $A_4 = A_3 \to c$; $A_5 = A_2 \to m$; $A_6 = A_1, A_4, A_5 \to \omega$; $A_7 = \neg\omega$; $A_8 = A_1, A_4, A_7 \to \neg m$; $A_9 = A_1, A_5, A_7 \to \neg c$; $A_{10} = A_4, A_5, A_7 \to \neg d$; $A_{11} = A_8 \to \neg b$; $A_{12} = A_9 \to \neg p$.

Since no defeasible rules are involved, all attacks have the form of undermining. Focusing on ordinary premises first, we have that: A_1 is undermined by A_{10}, A_2 is undermined by A_{11} and A_3 is undermined by A_{12}. Then, considering the use of ordinary premises in the construction of other arguments, the following attacks also occur: A_{10} undermines A_6, A_8, A_9, A_{11}, A_{12}; A_{11} undermines A_5, A_6, A_9, A_{10}, A_{12}; A_{12} undermines A_4, A_6, A_8, A_{10}, A_{11}. The relevant argumentation framework is shown in Figure 1.

It can be then observed that the core of the framework consists of the three arguments A_{10}, A_{11}, A_{12}, which are the only sources of attacks and mutually attack each other. As a consequence, it can be seen that in the case of stable, preferred, or semi-stable semantics, we get the same three extensions: $\mathsf{E}_1 =$

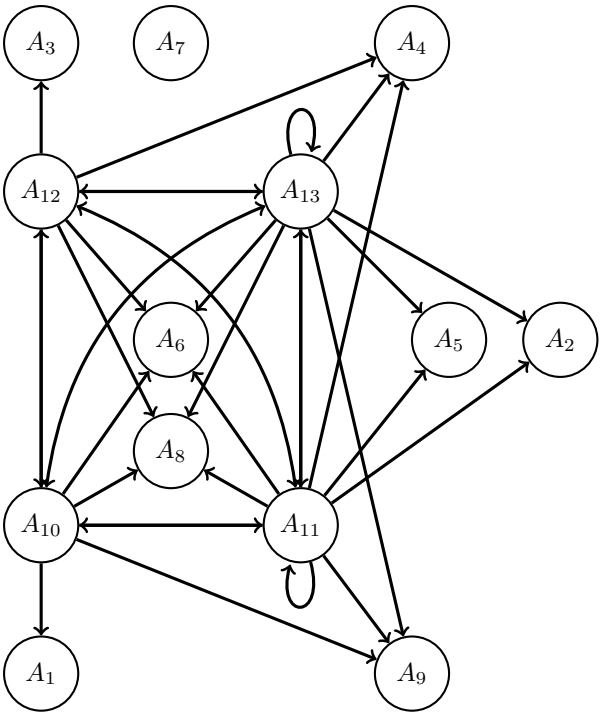

Figure 2: The argumentation framework corresponding to the second version of the taxpayers' list example.

$\{A_2, A_3, A_4, A_5, A_7, A_{10}\}$; $\mathsf{E}_2 = \{A_1, A_3, A_4, A_7, A_8, A_{11}\}$; and $\mathsf{E}_3 = \{A_1, A_2, A_5, A_7, A_9, A_{12}\}$.

Each of these extensions corresponds to the rejection of one of the three ordinary premises (A_1, A_2, and A_3, respectively) corresponding to the three uncertain pieces of evidence, which reflects the absence of any preference among them: according to Definition 3 any of them is credulously justified, while none of them is skeptically justified.

Consider now a slight variation of the above example, namely a country where *ius soli* has been introduced at a certain date (which we assume to be before the known birthdate of the person). In this case, the rule that derives citizenship will use both the birthplace and the birthdate as premises. Intuitively, this small variation in the structure of the strict rules should not affect the result: the three uncertain pieces of evidence should still be regarded to be equal candidates for rejection, given that, together, they give rise to a contradiction with a certain fact and that there is no preference among them. Somehow surprisingly, this is not the case, as illustrated below.

With respect to the representation of the original example, everything remains the same with the exception of the strict rule $p \to c$, which becomes $b, p \to c$. This gives rise to $\mathcal{R}'_{Sb} = \{b \to m; b, p \to c; d, c, m \to \omega\}$.

Closure under transposition of $\mathcal{R}'_{Sb}$ gives rise to $\mathcal{R}'_{St} = \{\neg m \to \neg b; p, \neg c \to \neg b; b, \neg c \to \neg p; d, c, \neg\omega \to \neg m; d, m, \neg\omega \to \neg c; c, m, \neg\omega \to \neg d\}$, and the overall set of strict rules is given by $\mathcal{R}'_S = \mathcal{R}'_{Sb} \cup \mathcal{R}'_{St}$.

The following arguments are then built (for simplicity of notation, we use the same names as above since the distinction is clear from the context): $A_1 = d$; $A_2 = b$; $A_3 = p$; $A_4 = A_2, A_3 \to c$; $A_5 = A_2 \to m$; $A_6 = A_1, A_4, A_5 \to \omega$; $A_7 = \neg\omega$; $A_8 = A_1, A_4, A_7 \to \neg m$; $A_9 = A_1, A_5, A_7 \to \neg c$; $A_{10} = A_4, A_5, A_7 \to \neg d$; $A_{11} = A_8 \to \neg b$; $A_{12} = A_2, A_9 \to \neg p$; $A_{13} = A_3, A_9 \to \neg b$.

The attacks occurring directly on ordinary premises are the following: A_1 is undermined by A_{10}; A_2 is undermined by A_{11} and A_{13}; A_3 is undermined by A_{12}. Then, according to the use of the premises, the following attacks also occur: A_{10} undermines A_6, A_8, A_9, A_{11}, A_{12}, A_{13}; A_{11} and A_{13} undermine A_4, A_5, A_6, A_8, A_9, A_{10}, A_{11}, A_{12}, A_{13}; finally A_{12} undermines A_4, A_6, A_8,A_{10}, A_{11}, A_{13}.

The relevant argumentation framework is shown in Figure 2.

The core of the resulting argumentation framework consists of the four arguments A_{10}, A_{11}, A_{12},

and A_{13} which are the only sources of attacks. The four arguments are all mutually attacking each other, moreover A_{11} and A_{13} are self defeating. It turns out that according to the preferred, stable and semi-stable semantics, there are two extensions: one where A_{10} is accepted while A_{11}, A_{12}, and A_{13} are rejected, and one where A_{12} is accepted while A_{10}, A_{11}, and A_{13} are rejected. Taking into account the attack relations involving other arguments it turns out that the two extensions are as follows: $\mathsf{E}_1 = \{A_2, A_3, A_4, A_5, A_7, A_{10}\}$ and $\mathsf{E}_2 = \{A_1, A_2, A_5, A_7, A_9, A_{12}\}$.

It turns out that, focusing on the ordinary premises, A_1 and A_3 are alternatively rejected, while A_2 is always accepted, which corresponds to a sort of implicit preference for argument A_2, which turns out to be skeptically justified, with respect to arguments A_1 and A_3, which are credulously justified. This implicit preference can be regarded as an accidental side effect of the structure of the set of strict rules, and has therefore a rather dubious conceptual status. For this reason, we call this implicit preference *spurious*.

Indeed, the fact that the information on birthdate is involved in two intermediate reasoning steps turns out to give it a privileged status that appears to be accidental and unjustified. We regard spurious preferences as an interesting but undesired behavior, and we develop a relevant preliminary investigation in the next sections.

4. A basic requirement of spurious preference avoidance

A broad analysis of spurious preferences and the cases where they may arise appears to be a challenging task that is beyond the limits of this paper. As a preliminary contribution in this direction, in this section we define a simple reference context, which includes the motivating example as an instance, where it is possible to provide a formal notion of spurious preference. Accordingly, we will formulate a requirement of spurious preference avoidance, which is meant to support the comparison of different formalisms in situations like the one described in Section 3.

To this purpose, we resort to a notion of core argumentation theory, namely an argumentation theory which contains the essential elements for the representation of a reasoning case but may need to be completed with other elements, derivable from the core, before being used for the construction of arguments and their assessment. The idea is that different completions of the same core may behave differently with respect to spurious preference avoidance. In particular, we refer to a family of core argumentation theories called SSDOP (Simple Strict Derivation from Ordinary Premises), defined as follows.

Definition 12. *Let $AS = (\mathcal{L}, ^-, \mathcal{R}, n)$ be an argumentation system and $\mathcal{K}$ a knowledge base in AS. An argumentation theory $AT = (AS, \mathcal{K})$ is said to be an instance of the SSDOP family if the following conditions hold:*

- *the language $\mathcal{L}$ consists of the closure of a given set Σ of symbols and their negation, namely $\mathcal{L} = \Sigma \cup \{\neg s \mid s \in \Sigma\}$;*
- *the contrariness function coincides with the classical notion of negation: for every $s \in \Sigma$, $\overline{s} = \{\neg s\}$ and $\overline{\neg s} = \{s\}$;*
- *$\mathcal{R} = (\mathcal{R}_{Sb}, \emptyset)$, namely the set of defeasible rules is empty;*
- *$\nexists r, r' \in \mathcal{R}_{Sb} : cons(r) \in \overline{cons(r')}$, namely no contradiction can be derived using the strict rules only;*
- *$\mathcal{K}_n = \{\neg\omega\}$ for some $\omega \in \Sigma$ that will be called* contradiction focus*;*
- *$\forall r \in \mathcal{R}_{Sb}, ant(r) \cap \{\omega, \neg\omega\} = \emptyset$;*
- *the set of ordinary premises $\mathcal{K}_p$ satisfies the following conditions*
  - *$|\mathcal{K}_p| \geq 2$;*
  - *$\mathcal{K}_p \cap \{\omega, \neg\omega\} = \emptyset$;*
  - *$\nexists p, p' \in \mathcal{K}_p : p \in \overline{p'}$;*
  - *$\nexists r \in \mathcal{R}_{Sb} : cons(r) \in \mathcal{K}_p \cup \overline{\mathcal{K}_p}$*

- *there is an argument α such that $Prem(\alpha) = \mathcal{K}_p$, $Conc(\alpha) = \omega$, and there is no argument α' such that $Prem(\alpha') \subsetneq \mathcal{K}_p$, $Conc(\alpha') = \omega$;*
- *for every $p_1, p_2 \in \mathcal{K}_p$, $p_1 \simeq p_2$.*

It is easy to see that the examples presented in Section 3 are SSDOP instances. The idea is to encompass cases where a set $\mathcal{K}_p$ of ordinary premises gives rise, through strict rules, to a contradiction with a certain fact, and all ordinary premises are equal candidates to be rejected to avoid this contradiction. The assumptions on $\mathcal{K}_p$ are meant to ensure multiple choices ($|\mathcal{K}_p| \geq 2$) and that there are no other possible reasons to discard some elements of $\mathcal{K}_p$.

In a SSDOP instance, the main focus of the evaluation is represented by the ordinary premises, as they are the only defeasible elements. To capture this central aspect and to abstract away formal aspects which are specific of different variants of *ASPIC*$^+$ to be compared, we introduce a generic notion of evaluation mechanism, incorporating all steps leading from a SSDOP instance to the evaluation of the justification status of the premises.

Definition 13. *An SSDOP evaluation mechanism E is a function which, given a SSDOP instance AT with ordinary premises $\mathcal{K}_p$, returns a justification labeling of $\mathcal{K}_p$, denoted as E_{AT}.*

The evaluation mechanism is semantics-dependent in *ASPIC*$^+$. It consists of applying closure under transposition to the set of strict rules[3], then carrying out the reasoning steps corresponding to Definitions 7-11, and finally applying a chosen semantics σ to derive the justification labeling of ordinary premises from $\mathcal{E}_\sigma(\mathcal{F}_\Delta)$.

We can now introduce a basic requirement concerning spurious preference avoidance. First, we need to formalize the notion of modifying an argumentation theory by adding a premise to the antecedents of a strict rule.

Definition 14. *Given an argumentation system $AS = (\mathcal{L}, ^-, \mathcal{R}, n)$, where $\mathcal{R} = (\mathcal{R}_{Sb}, \emptyset)$ and a knowledge base $\mathcal{K}$ such that $AT = (AS, \mathcal{K})$ belongs to the SSDOP family, we say that $\mathcal{R}'_{Sb}$ is a P-addition of $\mathcal{R}_{Sb}$ iff $\exists r \in \mathcal{R}_{Sb}$ such that $\mathcal{R}'_{Sb} = (\mathcal{R}_{Sb} \setminus \{r\}) \cup \{r'\}$ where $cons(r') = cons(r)$ and $ant(r') = ant(r) \cup \{p\}$ for some $p \in \mathcal{K}_p$.*

With a little abuse of language, we will also say that AS' is a P-addition of AS when AS' is obtained from AS by replacing $\mathcal{R}_{Sb}$ with a P-addition $\mathcal{R}'_{Sb}$, and we will say that AT' is a P-addition of AT with the same meaning. It is easy to see that for any SSDOP instance AT, every P-addition of AT also belongs to the SSDOP family since the addition of a premise to any strict rule does not affect any of the conditions in Definition 12.

It is also easy to see that in Section 3, the first example corresponds to an instance of the SSDOP family, and the second example is a P-addition of the first one.

We can now specify a basic spurious preference avoidance requirement with reference to an evaluation mechanism. This requirement refers to the adoption of a multiple-status semantics and adheres to a credulous perspective: in a SSDOP instance each premise should be accepted in some scenario, but rejected in some other. Under a single-status semantics, like the grounded semantics, a skeptical treatment, where all the premises are rejected altogether, would be appropriate and a complementary analysis would need to be developed, which is left to future work.

Definition 15. *An argumentation theory $AT = (AS, \mathcal{K})$ which belongs to the SSDOP family is* Cr-premise-fair *with respect to an evaluation mechanism E iff for every ordinary premise $p \in \mathcal{K}_p$, $E_{AT}(p) = Cr$.*

Definition 16. *An evaluation mechanism E satisfies the requirement of basic spurious preference avoidance if given any argumentation theory AT which is Cr-premise-fair with respect to E, it holds that every P-addition of AT is Cr-premise-fair too.*

[3] In [5], the property of closure under contraposition is also considered to satisfy rationality postulates. We do not consider it here, as it is not constructive, leaving further analyses to future work.

In plain words, if the premises of a SSDOP instance AT are treated equally (with a credulous outcome), the equal treatment should be preserved in every P-addition of AT.

From the example in Section 3, it emerges then that the evaluation mechanism corresponding to *ASPIC*$^+$ does not guarantee basic spurious preference avoidance with preferred, stable, and semi-stable semantics. In Section 5, we show that a variant of *ASPIC*$^+$, introduced in [1] for different purposes, provides better guarantees.

5. ASPIC$^+$ revisited satisfies basic spurious preference avoidance

ASPIC$^+$ revisited (in the following *ASPIC*^{R}) was introduced in [1] with the main goal of addressing a technical problem affecting *ASPIC*$^+$ in presence of multiple contradictories, first evidenced in [6]. A thorough presentation of *ASPIC*^{R} is beyond the limits of the present paper. In this section, we recall the main differences with respect to *ASPIC*$^+$, which are relevant to the subsequent results, while omitting some unnecessary details.

One of the main standpoints of *ASPIC*^{R} is avoiding the requirement that strict rules are closed under transposition. Rather, the satisfaction of the rationality postulates is ensured (as shown in [1]) by an extended contrariness relation at the level of sets of language elements whose definition is based on a general notion of strict derivability.

Definition 17. *Given an argumentation system $AS = (\mathcal{L}, ^-, \mathcal{R}, n)$ the strict knowledge base $\mathcal{K}^*_{AS}$ for AS is given by $\mathcal{K}_n = \mathcal{L}$, $\mathcal{K}_p = \emptyset$ and the corresponding argumentation theory is defined as $AT^*_{AS} = (AS, \mathcal{K}^*_{AS})$. $S \vdash^* \varphi$ and $S \vdash^*_{min} \varphi$ denote respectively that $S \vdash \varphi$ and $S \vdash_{min} \varphi$ in AT^*_{AS}.*

Definition 18. *Given an argumentation theory $AT = (AS, \mathcal{K})$ with $AS = (\mathcal{L}, ^-, \mathcal{R}, n)$, let $\mathsf{EC}^1(AS)$, $\mathsf{EC}^2(AS)$, $\mathsf{EC}^3(AS)$ be the following subsets of $2^{\mathcal{L}} \times 2^{\mathcal{L}}$*

- $\mathsf{EC}^1(AS) = \{(\{\varphi\}, \{\psi\}) \mid \varphi \in \overline{\psi}\}$;
- $\mathsf{EC}^2(AS) = \{(S, \{\psi\}) \mid S \vdash^*_{min} \varphi \text{ and } \varphi \in \overline{\psi}\}$;
- $\mathsf{EC}^3(AS) = \{(S, T) \mid T \vdash^*_{min} \psi \text{ and } (S, \{\psi\}) \in \mathsf{EC}^1(AS) \cup \mathsf{EC}^2(AS)\}$.

Letting $\mathsf{EC}^{*m}(AS) = \mathsf{EC}^1(AS) \cup \mathsf{EC}^2(AS) \cup \mathsf{EC}^3(AS)$, and, for $S \subseteq \mathcal{L}$, $\widehat{S} = S \setminus \mathcal{K}_n$, the extended contrariness relation is defined as $\mathsf{EC}(AS) = \{(\widehat{S}, \widehat{T}) \mid (S, T) \in \mathsf{EC}^{*m}(AS)$ and $\forall (S', T') \in \mathsf{EC}^{*m}(AS)$ s.t. $\widehat{S'} \subseteq \widehat{S}$ and $\widehat{T'} \subseteq \widehat{T}, \widehat{S'} = \widehat{S}$ and $\widehat{T'} = \widehat{T}\} \subseteq 2^{\mathcal{L}} \times 2^{\mathcal{L}}$. U is a contrary of V if $(U, V) \in \mathsf{EC}(AS)$ and $(V, U) \notin \mathsf{EC}(AS)$; U is a contradictory of V if $(U, V) \in \mathsf{EC}(AS)$ and $(V, U) \in \mathsf{EC}(AS)$.

The various forms of attack are then revised, replacing Definition 8 with the following definition concerning attacks between a set of arguments and an argument.

Definition 19. *Given an argumentation theory $AT = (AS, \mathcal{K})$, a set of arguments X* attacks *an argument β iff X* undercuts, rebuts, *or* undermines *β where:*

- *X* undercuts *β (on β') iff for some $\beta' \in Sub(\beta)$ such that $r = Top(\beta') \in \mathcal{R}_D$, the following condition holds: $\exists T, U$ such that $T \cup U = Conc(X) \cup \{n(r)\}$, $(T, U) \in \mathsf{EC}(AS)$ and $n(r) \in U$.*
- *X* rebuts *β (on β') iff for some $\beta' \in Sub(\beta)$ of the form $\beta''_1, \ldots, \beta''_n \Rightarrow \varphi$ the following condition holds: $\exists T, U$ such that $T \cup U = Conc(X) \cup \{\varphi\}$, $(T, U) \in \mathsf{EC}(AS)$ and $\varphi \in U$. In this case X* contrary-rebuts *β if $(U, T) \notin \mathsf{EC}(AS)$.*
- *X* undermines *β (on β') iff for some $\beta' = \varphi$, $\varphi \in Prem_p(\beta)$ the following condition holds: $\exists T, U$ such that $T \cup U = Conc(X) \cup \{\varphi\}$, $(T, U) \in \mathsf{EC}(AS)$ and $\varphi \in U$. In this case X* contrary-undermines *β if $(U, T) \notin \mathsf{EC}(AS)$.*

As the effectiveness of some attacks depends on the preference relation, the notion of preference ordering needs to be generalized to sets of arguments.

Definition 20. *Given a preorder $\preceq$ on a set of arguments X, we extend $\preceq$ to $2^X \times X$ as follows. An argument α is at least as preferred as a set of arguments Y, denoted $Y \preceq \alpha$, iff $\exists \beta \in Y$ such that $\beta \preceq \alpha$. α is strictly preferred to Y, denoted $Y \prec \alpha$, iff $\exists \beta \in Y$ such that $\beta \prec \alpha$, not strictly preferred to Y, denoted $Y \not\prec \alpha$ iff $\nexists \beta \in Y$ such that $\beta \prec \alpha$.*

On this basis, an extended notion of defeat is introduced, replacing Definition 9.

Definition 21. *Let the set of arguments Y attack an argument β on β' according to Definition 19. If Y undercuts, contrary-rebuts, or contrary-undermines β on β', then Y* preference-independent *attacks β on β', otherwise Y* preference-dependent *attacks β on β'. Then, Y* defeats *β iff either Y* preference-independent *attacks β on β' or Y* preference-dependent *attacks β on β' and $Y \not\prec \beta'$. Y* minimally defeats *β, denoted as $Y \rightsquigarrow \beta$, if Y defeats β and $\nexists Y' \subsetneq Y$ such that Y' defeats β.*

An AF based on the notion of defeat provided in Definition 21 is then defined to evaluate the justification status of arguments. The idea is that the framework nodes represent relevant sets of arguments. In particular, we need a node for each singleton corresponding to a produced argument, and a node for each set of ultimately fallible arguments (as per Definition 22) that minimally defeats some produced argument.

Definition 22. *Given an argumentation theory $AT = (AS, \mathcal{K})$ let S be the set of the arguments produced in AS on the basis of $\mathcal{K}$. The set of ultimately fallible arguments of AT is defined as $\mathsf{UF}(\mathsf{S}) \triangleq \mathcal{K}_p \cup \{\alpha \in \mathsf{S} \mid Top(\alpha) \in \mathcal{R}_D\}$.*

Definition 23. *Given an argumentation theory $AT = (AS, \mathcal{K})$ with ordering $\preceq$, let S be the set of the arguments produced in AS on the basis of $\mathcal{K}$. The set of relevant sets of arguments of AT, denoted as $\mathsf{RS}(AT)$, is defined as $\mathsf{RS}(AT) = \{\{\alpha\} \mid \alpha \in \mathsf{S}\} \cup \{X \mid X \subseteq \mathsf{UF}(\mathsf{S})$ and $\exists \beta \in \mathsf{S} : X \rightsquigarrow \beta\}$.*

Then, a relevant set of ultimately fallible arguments attacks another one simply if it minimally defeats one of its members.

Definition 24. *Let $X, Y \in \mathsf{RS}(AT)$ for an argumentation theory $AT = (AS, \mathcal{K})$ and S be the set of the arguments produced in AS on the basis of $\mathcal{K}$. X D-attacks Y, denoted as $\|X\| \twoheadrightarrow \|Y\|$, iff $X \subseteq \mathsf{UF}(\mathsf{S})$ and $\exists \alpha \in Y : X \rightsquigarrow \alpha$.*

The relevant set based AF is defined accordingly.

Definition 25. *Given an argumentation theory $AT = (AS, \mathcal{K})$, the RS-based argumentation framework induced by AT is defined as $\mathsf{RS}{-}\mathsf{F}(AT) = (\{\|X\| \mid X \in \mathsf{RS}(AT)\}, \twoheadrightarrow)$.*

Finally, given an argumentation semantics σ, the justification status of an argument α corresponds to the one of $\|\{\alpha\}\|$ in $\mathsf{RS}{-}\mathsf{F}(AT)$ according to Definition 3.

ASPIC^R provides a different evaluation mechanism for SSDOP argumentation theories, consisting in deriving the extended contrariness relation and then carrying out the reasoning steps established by Definitions 17-25 till the assignment of a justification status, as described above. As in the case of *ASPIC*$^+$, the evaluation mechanism is parametric with respect to the choice of an argumentation semantics σ. We now prove that *ASPIC*^{R} satisfies the requirement of basic spurious preference avoidance for preferred, stable and semi-stable semantics.

Proposition 1. *Any argumentation theory $AT = (AS, \mathcal{K})$ which belongs to the SSDOP family is Cr-premise-fair with respect to the evaluation mechanism provided by* ASPIC^R *under the choice of preferred semantics.*

Proof: Let $AS = (\mathcal{L}, ^-, \mathcal{R}, n)$ be an argumentation system and $\mathcal{K}$ a knowledge base in AS, such that the argumentation theory $AT = (AS, \mathcal{K})$ belongs to the SSDOP family. Assume $\mathcal{K}_p = P$ and for every $x \in P$, let us denote $P_x = P \setminus \{x\}$. From the conditions in Definition 12, we have that $P \vdash^*_{min} \omega$, from which, with reference to Definition 18, it follows that $(P, \{\neg\omega\}) \in \mathsf{EC}^2(AS)$ and $(\{\neg\omega\}, P) \in \mathsf{EC}^3(AS)$. Since $\neg\omega \in \mathcal{K}_n$, we get $\{(P, \emptyset), (\emptyset, P)\} \in \mathsf{EC}(AS)$. From $(\emptyset, P) \in \mathsf{EC}(AS)$, it follows that for every $x \in P$, P_x undermines x according to Definition 19 (with $T = \emptyset$ and $U = P$). By the conditions in Definition 12, $P_x \not\vdash x$ and to conclude that P_x defeats x according to Definition 21 we have to show that $\nexists P' \subsetneq P_x$ such that P' defeats x. To this purpose, let us refer to the elements of the extended contrariness relation of the form (S, T) such that $x \in T$, taking into account the conditions specified in Definition 12. First, $(\{\neg x\}, \{x\})$ is the only such pair in $\mathsf{EC}^1(AS)$ and we note $\neg x$ is not a premise nor an axiom and that there is no rule r such that $\neg x = cons(r)$. Therefore there cannot be any argument with conclusion $\neg x$. For the same reason, there is no pair $(S, \{x\})$ with $S \neq \{\neg x\}$ in $\mathsf{EC}^2(AS)$. Turning to $\mathsf{EC}^3(AS)$, it contains any pair (S, T) with $x \in T$ satisfying the third bullet of Definition 18. This means that $S \vdash^*_{min} \varphi$ and $T \vdash^*_{min} \psi$ with $\varphi \in \overline{\psi}$. However, Definition 12 prevents that a contradiction is derived through strict rules only, which means that S or T must be a singleton, i.e. $S = \{\varphi\}$ or $T = \{\psi\}$, and that, to have an attack according to Definition 19, φ or ψ must be an ordinary premise or an axiom. The fact that φ or ψ is an ordinary premise is prevented by the conditions in Definition 12 (no ordinary premise is a contrary of another premise, no rule can conclude an ordinary premise or a contrary of an ordinary premise). Since by Definition 19 arguments cannot be undermined on axioms, the remaining case is where φ is an axiom, i.e. $\varphi = \neg\omega$. We have already shown above that it cannot be the case that the pair $(\{\neg\omega\}, \{x\})$ belongs to $\mathsf{EC}^1(AS) \cup \mathsf{EC}^2(AS)$. The only possible case is $(\{\neg\omega\}, T) \in \mathsf{EC}^3(AS)$, for some T such that $T \vdash_{min} \omega$ and $x \in T$. From the conditions in Definition 12 the only set T satisfying these constraints is P, from which it follows, as desired, that $\nexists P' \subsetneq P_x$ such that P' defeats x. It follows that P_x defeats x according to Definition 21 (recall that all premises are equally preferred, hence $P_x \not\vdash x$). Note now that, by the conditions in Definition 12, $\mathsf{UF}(AT) = \mathcal{K}_p = P$. It follows that, for every $x \in P$, $P_x \in \mathsf{RS}(AT)$ according to Definition 23. Then, according to Definition 24, $\|P_x\| \twoheadrightarrow \|X\|$ for every X such that $X \in \mathsf{RS}(AT)$ and $x \in X$. This means in particular that for every $x, y \in P$ with $x \neq y$, there is a mutual attack between $\|P_x\|$ and $\|P_y\|$ in the argumentation framework $\mathsf{RS}{-}\mathsf{F}(AT)$.

We want now to show that a generic $\|P_x\|$ has no other attackers in $\mathsf{RS}{-}\mathsf{F}(AT)$. Suppose there is $X \in \mathsf{RS}(AT)$ such that $\|X\| \twoheadrightarrow \|P_x\|$. By Definition 24, this means that $\exists y \in P_x$ such that $X \rightsquigarrow y$. Now, since $y \in P$, by the same reasoning carried out above, the only possibility is that $X = P_y$.

It follows that for every $x \in P$, $\|P_x\|$ is admissible and hence there is a preferred extension E including $\|P_x\|$. By Lemma 3 of [1], for every $y \in P_x$ it follows then that $\|\{y\}\| \in \mathsf{E}$, while of course $\|\{x\}\| \notin \mathsf{E}$ since it is attacked by $\|P_x\|$. Since $|P| \geq 2$ it follows that, for every $x \in P$, there is at least one preferred extension including $\|\{x\}\|$ and one not including it, hence x is credulously justified and AT is Cr-premise-fair as desired. □

Proposition 2. *Any argumentation theory $AT = (AS, \mathcal{K})$ which belongs to the SSDOP family is Cr-premise-fair with respect to the evaluation mechanism provided by* ASPIC^R *under the choice of stable and semi-stable semantics.*

Proof: We use the same notation introduced in the proof of Proposition 1 and we show that preferred extensions coincide with stable extensions (and hence with semi-stable extensions) in the argumentation framework $\mathsf{RS}{-}\mathsf{F}(AT)$. The conclusion then follows from Proposition 1. From the proof of Proposition 1 we know that for every $x \in P$ it holds that $\|P_x\|$ is attacked by all and only the elements $\|P_y\|$ with $y \in P$ and attacks these elements in turn. It follows that every preferred extension E must include exactly one element $\|P_x\|$ for some $x \in P$. We want now to show that given a generic preferred extension E, for every argument α, $\|\alpha\|$ is either defended by (and hence included in) E or attacked by E. By the properties of $\mathsf{RS}{-}\mathsf{F}(AT)$ proved in [1], it then follows that for every $X \in \mathsf{RS}(AT)$ $\|X\|$ is either defended by (and hence included in) E or attacked by E, thus showing that E is also stable. Given

the conditions in Definition 12, any argument α is either strict and firm, with $Prem(\alpha) \subseteq \{\neg\omega\}$ or plausible, with $Prem(\alpha) \cap P \neq \emptyset$. If α is strict and firm, then clearly $\|\alpha\|$ is unattacked in $\mathsf{RS}{-}\mathsf{F}(AT)$ and belongs to every preferred extension. If α is plausible, we observe that it can only be attacked on its ordinary premises, and, from the proof of Proposition 1, we already know that, for each ordinary premise $y \in Prem(\alpha)$, $\|y\|$ can only be attacked by $\|P_y\|$, and hence the set of attackers of $\|\alpha\|$ in $\mathsf{RS}{-}\mathsf{F}(AT)$ is $\|\alpha\|^- = \{\|P_y\| \mid y \in Prem(\alpha)\}$. Given a preferred extension E, let x be the one and only premise such that $\|P_x\| \in \mathsf{E}$. We have then two cases: (i) if $x \in Prem(\alpha)$, then $\|P_x\|$ attacks $\|\alpha\|$ in $\mathsf{RS}{-}\mathsf{F}(AT)$; (ii) if $x \notin Prem(\alpha)$, then for every $y \in Prem(\alpha)$ it holds that $x \in P_y$ and thus $\|P_x\|$ attacks $\|P_y\|$ in $\mathsf{RS}{-}\mathsf{F}(AT)$, hence $\|P_x\|$ defends $\|\alpha\|$ as desired. □

The property of basic spurious preference avoidance follows from Propositions 1 and 2.

Theorem 1. *The evaluation mechanism provided by* ASPIC^R *under the choice of preferred, stable, and semi-stable semantics satisfies the basic spurious preference avoidance requirement.*

Proof: The conclusion follows from the fact that given an argumentation theory AT which belongs to the SSDOP family every P-addition of AT belongs to the SSDOP family too. From Propositions 1 and 2 it follows then that both AT and any P-addition of AT are Cr-premise-fair, thus complying with the requirement of basic spurious preference avoidance. □

As an illustration of the above result, we describe the behavior of *ASPIC*^{R} in the second version of the taxpayers' list example without closure under transposition, giving rise to arguments: $A_1 = d$; $A_2 = b$; $A_3 = p$; $A_4 = A_2, A_3 \rightarrow c$; $A_5 = A_2 \rightarrow m$; $A_6 = A_1, A_4, A_5 \rightarrow \omega$; $A_7 = \neg\omega$.

Concerning the $\vdash_{min}$ relation we have that $\{\varphi\} \vdash^*_{min} \varphi$ for every $\varphi \in \mathcal{L}$; $\{b\} \vdash^*_{min} m$; $\{b, p\} \vdash^*_{min} c$; $\{d, c, m\} \vdash^*_{min} \omega$; $\{d, c, b\} \vdash^*_{min} \omega$; $\{d, p, b\} \vdash^*_{min} \omega$. As to $\mathsf{EC}(AS)$ we get:

- $\mathsf{EC}^1(AS) = \{(\{\varphi\}, \{\neg\varphi\}), (\{\neg\varphi\}, \{\varphi\}) \mid \varphi \in \mathcal{L}\}$;
- $\mathsf{EC}^2(AS) \setminus \mathsf{EC}^1(AS) = \{(\{b\}, \{\neg m\}), (\{b, p\}, \{\neg c\}), (\{d, c, m\}, \{\neg\omega\}), (\{d, c, b\}, \{\neg\omega\}), (\{d, p, b\}, \{\neg\omega\})\}$;
- $\mathsf{EC}^3(AS)\setminus(\mathsf{EC}^1(AS)\cup\mathsf{EC}^2(AS)) = \{(\{\neg m\}, \{b\}), (\{\neg c\}, \{b, p\}), (\{\neg\omega\}, \{d, c, m\}), (\{\neg\omega\}, \{d, c, b\}), (\{\neg\omega\}, \{d, p, b\})\}$;
- $\mathsf{EC}(AS) = \{(\{\varphi\}, \{\neg\varphi\}), (\{\neg\varphi\}, \{\varphi\}) \mid \varphi \in \mathcal{L} \setminus \{\omega\}\} \cup \{(\emptyset, \{\omega\}), (\{\omega\}, \emptyset), (\{b\}, \{\neg m\}), (\{b, p\}, \{\neg c\}), (\{\neg m\}, \{b\}), (\{\neg c\}, \{b, p\}), (\{d, c, m\}, \emptyset), (\{d, c, b\}, \emptyset), (\{d, p, b\}, \emptyset), (\emptyset, \{d, c, m\}), (\emptyset, \{d, c, b\}), (\emptyset, \{d, p, b\})\}$.

According to Definition 19, from $(\emptyset, \{d, p, b\}) \in \mathsf{EC}(AS)$ we get that $\{A_1, A_2\}$ undermines A_3, $\{A_1, A_3\}$ undermines A_2, and $\{A_2, A_3\}$ undermines A_1. Taking into account the subargument relations, $\{A_1, A_2\}$ undermines also A_4 and A_6, $\{A_1, A_3\}$ undermines also A_4, A_5 and A_6, $\{A_2, A_3\}$ undermines also A_6. Definition 19 encompasses also the following attacks: $\{A_4, A_5\}$ undermines A_1 and A_6; $\{A_1, A_4\}$ undermines A_2, A_4, A_5 and A_6; $\{A_2, A_4\}$ undermines A_1 and A_6. They however are "filtered out" by Definition 23 since they involve arguments which are not ultimately fallible. All the attack relations listed above are also minimal defeats.

It follows that $\mathsf{RS}(AT) = \{\{A_1\}, \{A_2\}, \{A_3\}, \{A_4\}, \{A_5\}, \{A_6\}, \{A_7\}, \{A_1, A_2\}, \{A_1, A_3\}, \{A_2, A_3\}\}$.

Then, according to Definition 24, $\|\{A_1, A_2\}\|$ D-attacks the singletons $\|\{A_3\}\|$, $\|\{A_4\}\|$, $\|\{A_6\}\|$, and the relevant sets including them, namely $\|\{A_1, A_3\}\|$ and $\|\{A_2, A_3\}\|$. Similarly $\|\{A_1, A_3\}\|$ D-attacks $\|\{A_2\}\|$, $\|\{A_4\}\|$, $\|\{A_5\}\|$, $\|\{A_6\}\|$, and $\|\{A_2, A_3\}\|$ and $\|\{A_1, A_2\}\|$, while $\|\{A_2, A_3\}\|$ D-attacks$\|\{A_1\}\|$, $\|\{A_6\}\|$, $\|\{A_1, A_2\}\|$ and $\|\{A_1, A_3\}\|$. The resulting argumentation framework is shown in Figure 3. It has three preferred, stable and semi-stable extensions, namely $\mathsf{E}_1 = \{\|\{A_1, A_2\}\|, \|\{A_1\}\|, \|\{A_2\}\|, \|\{A_5\}\|, \|\{A_7\}\|\}$, $\mathsf{E}_2 = \{\|\{A_2, A_3\}\|, \|\{A_2\}\|, \|\{A_3\}\|, \|\{A_4\}\|, \|\{A_5\}\|, \|\{A_7\}\|\}$, $\mathsf{E}_3 = \{\|\{A_1, A_3\}\|, \|\{A_1\}\|, \|\{A_3\}\|, \|\{A_7\}\|\}$. As expected, each of the three defeasible premises is credulously justified.

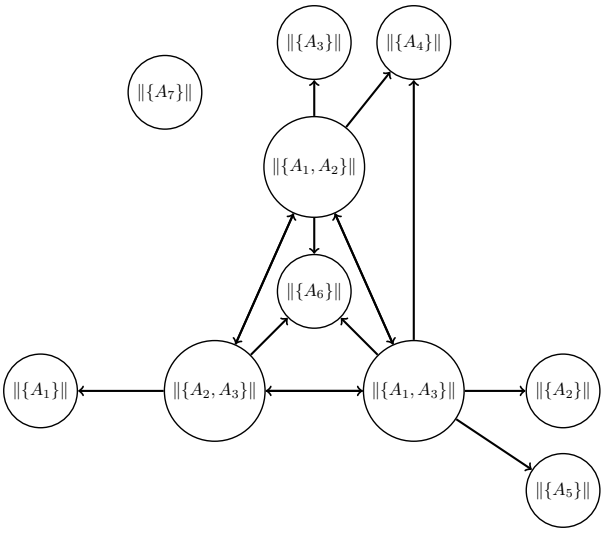

Figure 3: The argumentation framework generated by *ASPIC*^{R} for the second version of the taxpayers' list example.

6. Discussion and conclusions

Preferences have been considered in various ways both in structured and abstract argumentation [5, 7, 8]. However, to our knowledge, the issue of implicit undesired preferences emerging from the behavior of an argumentation system has not been considered before in the literature. After illustrating it in the context of *ASPIC*$^+$, we provided a requirement of basic spurious preference avoidance and showed that *ASPIC*^{R} satisfies it. This provides a language-independent approach to the problem of spurious preferences. It has to be noted that language-dependent solutions can also be considered. For instance, if one assumes a language equipped with the notion of logical conjunction, the second version of the example might use a rule like $b \wedge p \rightarrow c$ instead of $b, p \rightarrow c$. It can be seen that a spurious preference would not arise in this case. Pursuing a language-dependent solution would, however, be in contrast with the spirit of *ASPIC*$^+$ as a general framework not bound to a specific logical language nor to a specific interpretation thereof (see, for instance, the relevant remarks in [9]). Moreover, we suggest that it can be considered peculiar to get different results, in a relatively simple reasoning case, depending on the representation choice between $b, p \rightarrow c$ and $b \wedge p \rightarrow c$. We focused the analysis in this paper on the *ASPIC*$^+$ formalism, as a well-known and general approach to structured argumentation. As shown, the problem is not inherent to the *ASPIC*$^+$ formalism but can be regarded as a side effect of the use of closure under transposition of strict rules and can be avoided by using a different approach to ensure the satisfaction of rationality postulates. The present work represents an initial step in analysing the issue of spurious preferences. Among the many directions of future work, we mention the investigation about the occurrence of this problem in other structured argumentation formalisms like DeLP [10] or ABA [11] and the study of more general contexts with respect to SSDOP where spurious preferences should be avoided.

Acknowledgments

This work was supported by MUR project PRIN 2022 EPICA 'Enhancing Public Interest Communication with Argumentation' (CUP D53D23008860006) funded by the European Union - Next Generation EU, mission 4, component 2, investment 1.1.

References

[1] P. Baroni, M. Giacomin, B. Liao, A general semi-structured formalism for computational argumentation: Definition, properties, and examples of application, Artif. Intell. 257 (2018) 158–207.

[2] P. Baroni, M. Caminada, M. Giacomin, An introduction to argumentation semantics, Knowledge Engineering Review 26 (2011) 365–410.

[3] M. Caminada, L. Amgoud, On the evaluation of argumentation formalisms, Artif. Intell. 171 (2007) 286 – 310.

[4] H. Prakken, An abstract framework for argumentation with structured arguments, Argument & Computation 1 (2010) 93–124.

[5] S. Modgil, H. Prakken, A general account of argumentation with preferences, Artif. Intell. 195 (2013) 361 – 397.

[6] P. Baroni, M. Giacomin, B. Liao, Dealing with generic contrariness in structured argumentation, in: Proc, of the 24th Int. Joint Conference on Artificial Intelligence (IJCAI 2015), 2015, pp. 2727–2733.

[7] L. Amgoud, S. Vesic, Rich preference-based argumentation frameworks, International Journal of Approximate Reasoning 55 (2014) 585–606.

[8] S. Kaci, L. van Der Torre, S. Vesic, S. Villata, Preference in Abstract Argumentation, in: Handbook of Formal Argumentation, Volume 2, College Publications, 2021.

[9] H. Prakken, S.Modgil, Clarifying some misconceptions on the aspic$^+$ framework, in: Proc. of the 4th Int. Conf. on Computational Models of Argument (COMMA 2012), 2012, pp. 442–453.

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

[11] F. Toni, A tutorial on assumption-based argumentation, Argument & Computation 5 (2014) 89–117.