A formal account of Socratic-style argumentation [1,2]

Martin W.A. Caminada

Institute of Information & Computing Sciences, Utrecht University, PO Box 80.089, 3508 TB Utrecht, The Netherlands

Abstract

In traditional mathematical models of argumentation an argument often consists of a chain of rules or reasons, beginning with premisses and leading to a conclusion that is endorsed by the party that put forward the argument. In informal reasoning, however, one often encounters a specific class of counterarguments that until now has received little attention in argumentation formalisms. The idea is that instead of starting with the premisses, the argument starts with the propositions put forward by the counterparty, of which the absurdity is illustrated by showing their (defeasible) consequences. This way of argumentation (which we call S-arguments) is very akin to Socratic dialogues and critical interviews; it also has applications in modern philosophy. In this paper, various examples of S-arguments are provided, as well as a treatment of the problems that occur when trying to formalize them in existing formalisms. We also provide general guidelines that can serve as a basis for implementing S-arguments into various existing formalisms. In particular, we show how S-arguments can be implemented in Pollock's formalism, how they fit into Dung's abstract argumentation approach and how they are related to the issue of self-defeating arguments.

Key words: argumentation, dialogue, inference

1 Introduction

Over the last few years, many formalisms for defeasible argumentation have been defined. In some approaches, like [Dun95], the internal structure of the

[1] Supported by the Netherlands Organisation for Scientific Research (NWO) under project number 612.060.005.

[2] Supported by the European Union under contract IST-002307 (ASPIC)

arguments is left abstract. In other approaches arguments consist of sets of assumptions [BDKT97,BH01], of lists [SL92,PS97] or of trees [Vre97]. In this paper, we will restrict ourselves to approaches that represent arguments as lists or trees.

Argumentation can also be seen from a dialectical perspective. A relatively straightforward approach is the argument-games as described by Prakken and Sartor [PS97], Prakken and Vreeswijk [VP00], or Dunne and Bench-Capon [DBC03]. In these approaches, argumentation takes place between two agents — a proponent and an opponent — who take turns in putting forward arguments. A thus played argumentation game can essentially serve as a proof theory for the underlying Dung-style semantics of the argumentation formalism.

There exists a close connection between argumentation in its dialectical form (like [VP00,BCAC05]) and persuasion dialogues (like [Mac79,WK95]). The main difference is that in an argument dialectical game, each argument is given as a whole, whereas in a dialogue game an argument can also gradually be "rolled off". Instead of stating the entire argument at once, one first claims the conclusion and when this is questioned, one repeatedly uses the "because" speech act to lay out the reasons, until the point is reached where the reasons are no longer disputed — for instance when one has reached the set of shared premisses. The connection between argumentation and dialogue has been explored by Prakken [Pra00]. Although our main interest is in argumentation, we will sometimes use the field of dialogue in order to illustrate specific concepts or intuitions.

One of the advantages of applying argumentation instead of (nonmonotonic) logic in general is that argumentation comes closer to how people actually reason. This is especially true if argumentation is seen from a dialectical perspective [PS97,VP00]. An interesting question, therefore, is to which extent the current generation of argumentation formalisms is able to capture the richness of human argumentation. That is, does the current generation of argumentation formalisms support the various types of arguments used in informal argumentation?

In this paper, we provide a treatment of one specific type of informal argument — a type that has been applied practically since antiquity — and show that this type of argument cannot properly be represented by many of today's argumentation formalisms (we have chosen Pollock's formalism as an example). We then provide a conceptual analysis of this specific type of informal reasoning and specify how to adjust existing tree and list based argumentation formalisms (for which we again have chosen Pollock's formalism as an example) to properly implement it.

This paper is structured as follows. In Section 2 (Socratic-style arguments) we provide an overview of the informal argument form under discussion and show its various applications. In Section 3 (Pollock's argumentation formalism) we identify some problems that occur when one tries to apply this kind of arguments in today's generation of list and tree based argumentation formalisms, for which we have chosen Pollock's formalism as an example. In Section 4 (Analysis) an analysis is provided of the concepts that are related to the argument form under discussion. In Section 5 (Formalization) we show how the argument form can be included in Pollock's argumentation formalism. The semantical issues are dealt with in Section 6. In Section 7 it is discussed how our particular argument form can deal with some subtle issues regarding self-defeat. The discussion is rounded off in Section 8.

2 Socratic-style arguments

Many formalisms of argumentation (such as [Vre93], [PS97] and [Pol95]) regard an argument as a structured chain of rules. An argument usually begins with one or more premises — statements that are simply regarded as true by all involved parties, such as directly observable facts. After this follows the repeated application of various rules, which generate new conclusions and therefore enable the application of additional rules. An example of such an argument is as follows:

> "Sjaak probably went to the soccer game, since people claim his car was parked nearby the stadium, and Sjaak is known to be a soccer fan."

$claimed(car\_at\_stadium),\ soccer\_fan,$
$claimed(car\_at\_stadium) \Rightarrow car\_at\_stadium,$
$car\_at\_stadium \wedge soccer\_fan \Rightarrow at\_game$

Arguments are often *defeasible*, meaning that the argument by itself is not a conclusive reason for the conclusions it brings about. Whether or not an argument should be accepted depends on its possible counterarguments. For the above argument, a possible counterargument could be:

> "Sjaak did not go to the soccer game, since his friends claim he was watching the game with them in a bar."

$friends\_claim(at\_bar),$
$friends\_claim(at\_bar) \Rightarrow at\_bar,$
$at\_bar \rightarrow \neg at\_game$

The issue of determining the arguments and conclusions that are considered to be *justified* then becomes a matter of weighing and evaluating the given arguments.

Most systems for formal argumentation take arguments to be grounded in the premises; that is, each rule of the argument is ultimately (directly or indirectly) based on premises only. In human argumentation, however, one can often observe arguments which are not based on premises only, but which are instead at least partly based on the conclusions of the other person's argument. As an illustration, consider the following example of a discussion between the opponent and proponent of a certain thesis:

> P: "Guus did not go to the game because his mobile phone record shows he was in his mother's house at the time of the game."
>
> $phone\_record$,
> $phone\_record \Rightarrow at\_mothers\_house(phone)$,
> $at\_mothers\_house(phone) \Rightarrow at\_mothers\_house(Guus)$,
> $at\_mothers\_house(Guus) \rightarrow \neg at\_game(Guus)$
>
> O: "Then he would not have watched the game at all, since his mother's TV has been broken for quite a while. Don't you think that's a little odd? Guus is known to be a soccer fan and would definitely have watched the game."
>
> $soccer\_fan(Guus)$,
> $at\_mothers\_house(Guus) \Rightarrow \neg watch\_game(Guus)$,
> $soccer\_fan(Guus) \Rightarrow watch\_game(Guus)$

Here, the opponent takes the propositions as uttered by the proponent as a starting point and then uses these to (defeasibly) derive a contradiction, thus illustrating the (implicit) absurdity of the proponent's original argument.

Socrates and the elenchus

The idea of taking the other party's opinion and then deriving a contradiction (or something else that is undesirable to the other party) is not new. One of the first well known examples of this style of reasoning can be found in the philosophy of Socrates, as written down by Plato. Socrates's form of reasoning — also called the elenchus — consists of letting the opponent make a statement, and then taking this statement as a starting point to derive more statements, each of which is committed by the opponent. The ultimate aim is to let the opponent commit himself to a contradiction, which shows that the

beliefs the opponent uttered in the dialogue cannot hold together and should therefore be rejected.

As an example of how Socrates's form of dialectical reasoning worked, consider the following dialogue, in which Socrates questions Menexenus about the nature of friendship [Pla10, pp. 212-213]

> (...) Answer me this. As soon as one man loves another, which of the two becomes the friend? the lover of the loved, or the loved of the lover? Or does it make no difference?
>
> None in the world, that I can see, he replied.
>
> How? said I; are both friends, if only one loves?
>
> I think so, he answered.
>
> Indeed! is it not possible for one who loves, not to be loved in return by the object of his love?
>
> It is.
>
> Nay, is it not possible for him even to be hated? treatment, if I mistake not, which lovers frequently fancy they receive at the hands of their favorites. Though they love their darlings as dearly as possible, they often imagine that they are not loved in return, often that they are even hated. Don't you believe this to be true?
>
> Quite true, he replied.
>
> Well, in such a case as this, the one loves, the other is loved.
>
> Just so.
>
> Which of the two, then, is the friend of the other? the lover of the loved, whether or not he be loved in return, and even if he be hated, or the loved of the lover? or is neither the friend of he other, unless both love each other?
>
> The latter certainly seems to be the case, Socrates.
>
> If so, I continued, we think differently now from what we did before. Then it appeared that if one loved, both were friends; but now, that unless both love, neither are friends.
>
> Yes, I'm afraid we have contradicted ourselves.

Socrates's method is that of asking questions. The questions, however, are often meant to direct the dialogue partner into a certain direction. It is the questions that force the dialogue partner to make certain inferences, as these seem to logically follow from the dialogue partner's own position. The inferences are not deductive, as they are usually based on common sense and what is reasonable. The inference is therefore more of a defeasible than of a strict nature.

Socrates's elenchus is not meant for the derivation of new facts. On the contrary, its purpose is primarily destructive, meant to destroy someone's pretension of knowledge [Nel94]. In "The Sophist", Plato provides the following definition of the elenchus [PlaBC]:

> They [those that apply the elenchus] cross-examine a man's words, when he thinks that he is saying something and is really saying nothing, and easily convict him of inconsistencies in his opinions; these they then collect by the dialectical process, and placing them side by side, show that they contradict one another about the same things, in relation to the same things, and in the same respect. He, seeing this, is angry with himself, and grows gentle towards others, and thus is entirely delivered from great prejudices and harsh notions, in a way that is most amusing to the hearer, and produces the most lasting effect to the person who is the subject of the operation.

The destruction of knowledge is best pursued by showing it to be incompatible with other knowledge, as argued by the Belgian scholar Chaïm Perelman [Per82, p. 24]:

> How do we disqualify a fact or truth? The most effective way is to show its incompatibility with other facts and truths which are more certainly established, preferably with a *bundle* of facts and truths which we are not willing to abandon.

Of course, an obvious way to show incompatibility is by means of a classical counterargument, but there are also forms of incompatibility that require argumentation beyond classical arguments.

Some modern examples

The kind of reasoning in which one confronts the other party with the (defeasible) consequences of its statements is still widely used in modern times. Consider the following dialogue between politician P and interviewing journalist J:

P: In two years time, the waiting lists in health care will be as good as resolved.

J: Then you are actually saying that the insurance fees will be increased, because the government has already decided not to put more money into the health care system, and you have promised not to lower the coverage of the standard insurance.

In general, one may say that many of today's interviews in which the interviewer takes a critical stance, the interviewer tries to force the interviewee to draw conclusions or make statements that the interviewee may wish to avoid.

On a more philosophical level, James Skidmore discusses the issue of *transcendental arguments*, which are meant to combat various forms of (philosophical) scepticism. The aim of a transcendental argument is "to locate something that

the sceptic must presuppose in order for her challenge to be meaningful, then to show that from this presupposition it follows that the skeptic's challenge can be dismissed." [Ski02, p. 121] Skidmore gives various (rather long) examples of these kind of arguments — we will not repeat them here.

To summarize, the technique of using statements from the other party's argument against him is still common in modern times, both in popular as well as in philosophical argumentation. It is our opinion that therefore the question of how these arguments can be formally modelled is a relevant one.

3 Pollock's argumentation formalism

In this section, we examine the problems that one encounters when trying to apply Socratic-style arguments (S-arguments) using today's formalisms for defeasible reasoning. The main focus of this section is on the argumentation formalism of John Pollock [Pol87,Pol92,Pol95]. We have chosen Pollock's formalism because it is well-known and is rich enough to deal with rebutting and undercutting defeaters, as well as with differences in rule-strength. Nevertheless, the problems that are discussed in this section also play a role in other formalisms (like default logic [Rei80] and the formalism of Prakken and Sartor [PS97]), as explained in [Cam04]. During his years of research, Pollock has produced different versions of his formalism. In this thesis, we focus on two of these versions:

- the one based on grounded semantics [Pol87,Pol92] (Section 3.1)
- the one resulting from an analysis of self-defeating arguments [Pol95] (Section 3.3)

We start with a summary of Pollock's grounded semantics based system. Notice that this summary is partly based on the Handbook of Philosophical Logic [PV02].

3.1 Pollock's grounded semantics based system

In the system of Pollock, arguments are constructed by means of reasons. Pollock distinguishes two kinds of reasons: conclusive and prima facie.

Conclusive reasons are reasons that logically entail their conclusions. A conclusive reason is any valid form of first order deduction. The following are examples of conclusive reasons.

$\{p,\ p \supset q\}$ is a conclusive reason for q
$\{\exists x : Px\}$ is a conclusive reason for $\neg\forall x : \neg Px$

Prima facie reasons, at the other hand, are not necessarily valid in first order logic; they only create a presumption in favor of their conclusion. This presumption can be defeated by other reasons, depending on the strength of the conflicting reasons. Pollock distinguishes several kinds of prima facie reasons, for instance principles of perception, such as:

$\lceil x$ appears to me as $Y \rceil$ is a prima facie reason for believing $\lceil x$ is $Y \rceil$

Notice that $\lceil . \rceil$ stands for the *objectification* operator. With this operator, expressions in the meta-language are translated into expressions in the object-language.

Another general source of prima facie reasons is the statistical syllogism:

If $(r > 0.5)$ then $\lceil x$ is an F and $prob(G/F) = r \rceil$ is a prima facie reason of strength r for believing $\lceil x$ is a $G \rceil$.

Other sources of prima facie reasons are also available [Pol95].

Pollock's notion of an argument is made formal in the following definition (taken from [PV02]) which is essentially an argument-based interpretation of [Pol95].

Definition 1 *Let* `INPUT` *be a consistent set of first-order formulas. An* argument *based on* `INPUT` *is a finite sequence* $\sigma_1, \ldots \sigma_n$*, where each* σ_i *is a line of argument. A* line of argument σ_i *is a triple* $\langle X_i, p_i, \nu_i \rangle$*, where* X_i*, a set of propositions, is the set of* suppositions *of* σ_i*,* p_i *is a proposition, and* ν_i *is the* degree of justification *of* σ_i*. A new line of argument is obtained from the earlier lines of argument according to one of the following rules of argument formation.*

Input. *If* p *is in* `INPUT` *and* σ *is an argument, then for any* X *it holds that* $\sigma, \langle X, p, \infty \rangle$ *is an argument.*

Reason. *If* σ *is an argument,* $\langle X_1, p_1, \eta_1 \rangle, \ldots, \langle X_n, p_n, \eta_n \rangle$ *are members of* σ*, and* $\{p_1, \ldots, p_n\}$ *is a reason of strength* ν *for* q*, and for each* $i, X_i \subseteq X$*, then* $\sigma, \langle X, q, \min\{\eta_1, \ldots, \eta_n, \nu\} \rangle$ *is an argument.*

Supposition. *If* σ *is an argument,* X *a set of propositions and* $p \in X$*, then* $\sigma, \langle X, p, \infty \rangle$ *is also an argument.*

Conditionalization. *If* σ *is an argument and some line of* σ *is* $\langle X \cup \{p\}, q, \nu \rangle$*, then* $\sigma, \langle X, (p \supset q), \nu \rangle$ *is also an argument.*

Dilemma *If* σ *is an argument and some line of* σ *is* $\langle X, p \vee q, \nu \rangle$*, and some line of* σ *is* $\langle X \cup \{p\}, r, \mu \rangle$*, and some line of* σ *is* $\langle X \cup \{q\}, r, \xi \rangle$*, then* $\sigma, \langle X, r, \min\{\nu, \mu, \xi\} \rangle$ *is also an argument.*

Pollock [Pol95] notes that other inference rules could be added as well.

The addition of argument formation rules like *supposition*, *conditionalization* and *dilemma* makes it possible to construct *suppositional arguments*, in addition to linear arguments. The idea of suppositional reasoning is to "suppose" something that is not derived from other information, draw conclusions from it, and then "discharge" the supposition to obtain a conclusion that no longer depends on the supposition. The way in which Pollock's system deals with suppositional reasoning is very similar to the use of assumptions in natural deduction. As a result of this, each line of inference contains an associated set of suppositions.

Pollock's formalism is one of the very few nonmonotonic logics that allow for suppositional reasoning. An example of the usefulness of suppositional arguments is that it enables "reasoning by cases", which is left unsupported in most other logics for nonmonotonic reasoning. For instance, if Dutch people usually like ice-skating, Norwegian people usually like ice-skating, and Sven is either Dutch or Norwegian, then it seems a reasonable conclusion that, presumably, Sven likes ice-skating.

Suppositional reasoning, however, is outside of the scope of the current paper, which mainly focusses on linear (non-suppositional) arguments. An example of such an argument is the following.

$\texttt{INPUT} = \{\texttt{kiss}(\texttt{Mary}, \texttt{John}), \texttt{looks\_cold}(\texttt{Mary})\}$
$\texttt{PFREASONS} = \{\texttt{looks\_cold}(\texttt{X}) \Rightarrow^{0.8} \texttt{has\_cold}(\texttt{X}),$
$\texttt{kiss}(\texttt{X}, \texttt{Y}) \wedge \texttt{has\_cold}(\texttt{X}) \Rightarrow^{0.6} \texttt{contaminated}(\texttt{X})\}$

1. $\langle \emptyset, \texttt{looks\_cold}(\texttt{Mary}), \infty \rangle$	(INPUT)
2. $\langle \emptyset, \texttt{has\_cold}(\texttt{Mary}), 0.8 \rangle$	(1, first prima facie reason)
3. $\langle \emptyset, \texttt{kiss}(\texttt{Mary}, \texttt{John}), \infty \rangle$	(INPUT)
4. $\langle \emptyset, \texttt{has\_cold}(\texttt{Mary}) \wedge \texttt{kiss}(\texttt{Mary}, \texttt{John}), 0.8 \rangle$	(2, 3 conclusive reason)
5. $\langle \emptyset, \texttt{contaminated}(\texttt{John}), 0.6 \rangle$	(4, second prima facie reason)

Throughout this paper, we will make use of many examples in order to illustrate concepts and potential problems. In order to keep the treatment concise, we hereby introduce an abbreviated notation for Pollock-style arguments. The main idea is to represent an argument as a chained sequence of reasons, instead of a "natural deduction style" derivation. The above argument can be represented as follows.

$\texttt{looks\_cold}(\texttt{Mary})^{\infty}, \quad \texttt{looks\_cold}(\texttt{Mary})^{\infty} \Rightarrow^{0.8} \texttt{has\_cold}(\texttt{Mary})^{0.8}, \quad \texttt{kiss}(\texttt{Mary}, \texttt{John})^{\infty},$
$\texttt{kiss}(\texttt{Mary}, \texttt{John})^{\infty}, \texttt{has\_cold}(\texttt{Mary})^{0.8} \rightarrow (\texttt{kiss}(\texttt{Mary}, \texttt{John}) \wedge \texttt{has\_cold}(\texttt{Mary}))^{0.8},$
$(\texttt{kiss}(\texttt{Mary}, \texttt{John}) \wedge \texttt{has\_cold}(\texttt{Mary}))^{0.8} \Rightarrow^{0.6} \texttt{contaminated}(\texttt{John})^{0.6}$

The above notation is a condensed version of the natural deduction style notation. A line obtained by applying a reason is represented by the respective reason (where "⇒" stands for prima facie reasons and "→" for conclusive reasons). Lines containing statements from `INPUT` are represented simply by the respective `INPUT`-statement. In order to keep the treatment concise, we often omit strict reasons that only build a conjunction of their premisses. Strengths of statements and prima facie reasons are omitted iff every prima facie reason has the same strength. Notice that our abbreviated argument-notation is not automatically suitable to represent suppositional reasoning. For our limited purposes, however, it will do.

Now that the notion of an argument has been explained, we can continue with Pollock's definition of defeat. For now, we use Pollock's old version of defeat [Pol92], as this is relatively simple to deal with. Although Pollock sometimes defines defeat in terms of inference graphs, we will instead use the equivalent argument interpretation of Prakken and Vreeswijk [PV02].

Definition 2 (rebut) *An argument σ* rebuts *an argument η iff:*

(1) η contains a line of the form $\langle X, q, \alpha\rangle$ that is obtained by the argument formation rule Reason *from some earlier lines $\langle X_1, p_1, \alpha_1\rangle, \ldots, \langle X_n, p_n, \alpha_n\rangle$ where $\{p_1, \ldots, p_n\}$ is a prima facie reason for q, and*
(2) σ contains a line of the form $\langle Y, \neg q, \beta\rangle$ where $X \subseteq Y$ and $\beta \geq \alpha$.

Definition 3 (undercut) *An argument σ* undercuts *an argument η iff:*

(1) η contains a line of the form $\langle X, q, \alpha\rangle$ that is obtained by the argument formation rule Reason *from some earlier lines $\langle X_1, p_1, \alpha_1\rangle, \ldots, \langle X_n, p_n, \alpha_n\rangle$ where $\{p_1, \ldots, p_n\}$ is a prima facie reason for q, and*
(2) σ contains a line of the form $\langle Y, \neg\lceil\{p_1, \ldots, p_n\} \rightarrow q\rceil, \beta\rangle$ where $Y \subseteq X$ and $\beta \geq \alpha$.

In the above definition $\lceil\{p_1, \ldots, p_n\} \Rightarrow q\rceil$ is a translation of "$\{p_1, \ldots p_n\}$ is a prima facie reason for q" into the object language.

Definition 4 (defeat) *An argument σ* defeats *an argument η iff σ rebuts or undercuts η.*

Regarding justified arguments, Pollock uses the following inductive definition [Pol87].

Definition 5 ([Pol87])

- *All non-selfdefeating arguments are* in at level 0.
- *An argument is* in at level $n+1$ *($n > 0$) iff it is in at level 0 and it is not defeated by any argument that is in at level n.*

- *An argument is* justified *iff there is an m such that for every $n \geq m$ the argument is in at level n.*

If self-defeating arguments are not taken into consideration then this definition is, as shown by Dung, equivalent to the definition of grounded semantics [Dun95] under the condition that every argument is defeated by at most a finite number of counterarguments.

3.2 Examples

We are now ready to treat some informal examples and discuss how Pollock's formalism deals with them. Each of the following examples begins with a small natural language conversation between a proponent (P) and an opponent (O) of a certain statement, followed by an attempt to formalize P and O's arguments in Pollock's formalism.

Example 6 (classical rebut)
P: I don't think it will rain this afternoon ($\neg$ra). The weather is sunny now (sn), so it will probably also be like this in the afternoon (sa).
O: But the weather forecast predicted rain (wfpr).
INPUT $= \{\mathtt{sn}, \mathtt{wfpr}\}$
PFREASONS $= \{\mathtt{sn} \Rightarrow \mathtt{sa},\ \mathtt{sa} \Rightarrow \neg\mathtt{ra},\ \mathtt{wfpr} \Rightarrow \mathtt{ra}\}$
P: $\mathtt{sn},\ \mathtt{sn} \Rightarrow \neg\mathtt{ra}$
O: $\mathtt{wfpr},\ \mathtt{wfpr} \Rightarrow \mathtt{ra}$

Here, the argument of O rebuts the argument of P.

Example 7 (classical undercut)
P: It probably rained last night (rln) because the streets are wet (sw) when I opened the curtains this morning, and I can't think of any other reason why they would be wet.
O: The streets are wet because a water pipeline bursted last night (wpb).
INPUT $= \{\mathtt{sw}, \mathtt{wpb}\}$
PFREASONS $= \{\mathtt{sw} \Rightarrow \mathtt{rln},\ \mathtt{wpb} \Rightarrow \neg\lceil\mathtt{sw} \Rightarrow \mathtt{rnl}\rceil\}$
P: $\mathtt{sw},\ \mathtt{sw} \Rightarrow \mathtt{rln}$
O: $\rightarrow \mathtt{wpb},\ \mathtt{wpb} \Rightarrow \neg\lceil\mathtt{sw} \Rightarrow \mathtt{rnl}\rceil\}$

Here, the argument of O undercuts the argument of P.

Example 8 (self-defeat)
P: John goes out with his friends every night (go), so he's probably bachelor (b). He also wears a ring on his hand (r), so he's probably married (m). So, he is both married and not, and therefore the earth is flat (fe).
O: Give me a break...

$\texttt{INPUT} = \{\texttt{go},\ \texttt{r},\ \texttt{b} \supset \neg\texttt{m}\}$
$\texttt{PFREASONS} = \{\texttt{go} \Rightarrow \texttt{b},\ \texttt{r} \Rightarrow \texttt{m}\}$
P: $\texttt{g},\ \texttt{g} \Rightarrow \texttt{b},\ \texttt{r},\ \texttt{r} \Rightarrow \texttt{m},\ \texttt{b} \supset \neg\texttt{m},\ \texttt{b} \wedge (b \supset \neg\texttt{m}) \rightarrow \neg\texttt{m},\ \texttt{m} \wedge \neg\texttt{m} \rightarrow \texttt{fe}$
O: —

Here, P puts forward an argument that is self-defeating (it rebuts itself). In Pollock's formalism, as well as in several other formalisms for formal argumentation, such an argument is automatically rejected without the need for serious counterargument (see Definition 5). The idea is that someone who contradicts himself should not be taken seriously.

Example 9 *(Ajax-Feijenoord; Socratic-style undercut)*

P: *"There is a threat that tonight's soccer game will lead to riots (t), because Ajax plays against Feijenoord (af)."*

O: *"Don't worry, if there is really a threat, then the government will of course send extra police, which will then make sure that tonight's game no longer causes a security threat."*

$\texttt{INPUT} = \{\texttt{af}\}$

$\texttt{PFREASONS} = \{\texttt{af} \Rightarrow \texttt{t},\ \texttt{t} \Rightarrow \texttt{p},\ \texttt{p} \Rightarrow \neg\lceil\texttt{af} \Rightarrow \texttt{t}\rceil\}$

P: $\texttt{af},\ \texttt{af} \Rightarrow \texttt{t}$ *(A_1)*

O: $\texttt{af},\ \texttt{af} \Rightarrow \texttt{t},\ \texttt{t} \Rightarrow \texttt{p},\ \texttt{p} \Rightarrow \neg\lceil\texttt{af} \Rightarrow \texttt{t}\rceil$ *(A_2)*

Here, the opponent confronts the proponent with the (defeasible) consequences of his own reasoning which undercuts the original argument. Thus, O's informal argument is essentially an Socratic-style argument (the same will hold in examples 10 until 12). In Pollock's original formalism (as well as in many others) the only way to represent this undercutting argument is by means of a self-defeating argument, which in Pollock's formalism (and, again, also in many others) is automatically made "out". The question, however, is whether the informal argument is in essence also self-defeating. In section 4 it will be argued that this is not the case and that Pollock's formalism (as well as many others) simply lacks the constructs needed to properly model this kind of arguments.

Example 10 *(shipment of goods; Socratic-style rebut)*

P: *"The shipment of goods must have arrived in the Netherlands by now (*$\texttt{a}$*), because we placed an order three months ago (*$\texttt{tma}$*)"*

O: *"I don't think so. If the goods would really have arrived in the Netherlands, then there would be a customs declaration (*$\texttt{cd}$*), and I can't see any such declaration in our information system (*$\neg\texttt{is}$*)."*

$\texttt{INPUT} = \{\texttt{tma},\ \neg\texttt{is}\}$

$\texttt{PFREASONS} = \{\texttt{tma} \Rightarrow \texttt{a},\ \neg\texttt{is} \Rightarrow \neg\texttt{cd},\ \texttt{a} \Rightarrow \texttt{cd}\}$

P: $\texttt{tma},\ \texttt{tma} \Rightarrow \texttt{a}$ (A_1)

O: $\texttt{tma},\ \texttt{tma} \Rightarrow \texttt{a},\ \texttt{a} \Rightarrow \texttt{cd},\ \neg\texttt{is},\ \neg\texttt{is} \Rightarrow \neg\texttt{cd}$ (A_2)

Here, the opponent again confronts the proponent with the (defeasible) consequences of his own reasoning. This time, the consequences are an inconsistency (cd and ¬cd). The result is again a self-defeating argument.

Example 11 *(tax relief; Socratic-style rebut)*

P: *"Next year, we are going to get a tax-relief (*tr*), because our politicians promised so (*pmp*)."*

O: *"But in the current situation, you can only implement a tax-relief by accepting a significant budget deficit (*bd*), which means we will also get a huge fine from Brussels (*fb*). There goes our tax-relief."*

$\texttt{INPUT} = \{\texttt{pmp}\}$

$\texttt{PFREASONS} = \{\texttt{pmp} \Rightarrow \texttt{tr},\ \texttt{tr} \Rightarrow \texttt{bd},\ \texttt{bd} \Rightarrow \texttt{fb},\ \texttt{fb} \Rightarrow \neg\texttt{tr}\}$

P: $\texttt{pmp},\ \texttt{pmp} \Rightarrow \texttt{tr}$ (A_1)

O: $\texttt{pmp},\ \texttt{pmp} \Rightarrow \texttt{tr},\ \texttt{tr} \Rightarrow \texttt{bd},\ \texttt{bd} \Rightarrow \texttt{fb},\ \texttt{fb} \Rightarrow \neg\texttt{tr}$ (A_2)

Here, the (defeasible) consequences of proponent's argument are again inconsistent. Notice that although example 10 could theoretically be dealt with by allowing the controversial principle of default contraposition (which would enable the construction of a counterargument "$\neg\texttt{is},\ \neg\texttt{is} \Rightarrow \neg\texttt{cd},\ \neg\texttt{cd} \Rightarrow a$") such an approach is not possible in example 11. Even if one allows default contraposition, all possible counterarguments will still be self-defeating, and will therefore automatically be "out".

3.3 Pollock's new system

Pollock is one of the few researchers in the field of defeasible reasoning who has given special attention to the issues related to self-defeating arguments. One particular example that is treated by Pollock is the following [Pol91].

Example 12 (pink elephant)
*Robert says the elephant besides him looks pink (*rselp*).*

The fact that Robert says the elephant looks pink is a reason to believe that it is looks pink (`elp`*).*
The fact that the elephant looks pink is a reason to believe that it is pink (`eip`*).*
Robert becomes unreliable in the presence of pink elephants (`ruppe`*).*

$\texttt{INPUT} = \{\texttt{rselp},\ \texttt{ruppe},\ \texttt{eip} \wedge \texttt{ruppe} \supset \neg(\texttt{rselp} \Rightarrow \texttt{elp})\}$
$\texttt{PFREASONS} = \{\texttt{rselp} \Rightarrow \texttt{elp},\ \texttt{elp} \Rightarrow \texttt{eip}\}$

We can then construct the following arguments:

P: $\texttt{rselp},\ \texttt{rselp} \Rightarrow \texttt{elp}$
O: $\texttt{rselp},\ \texttt{rselp} \Rightarrow \texttt{elp},\ \texttt{elp} \Rightarrow \texttt{eip},\ \texttt{ruppe},\ \texttt{eip} \wedge \texttt{ruppe} \supset \neg(\texttt{rselp} \Rightarrow \texttt{elp}),\ \texttt{eip} \wedge \texttt{ruppe} \wedge (\texttt{eip} \wedge \texttt{ruppe} \supset \neg(\texttt{rselp} \Rightarrow \texttt{elp})) \rightarrow \neg(\texttt{rselp} \Rightarrow \texttt{elp})$

The key point to notice about the above example is that it concerns a self-defeating argument that has an undercutter that undercuts one of the rules it is based on (one could say that "the argument's head bites its body"). In this respect, Pollock's pink elephant example is similar to our Ajax-Feijenoord example, although the latter is somewhat simpler. Pollock argues that in the pink elephant example, not only `eip` but also `elp` should be prevented from becoming justified. In this respect, Pollock shares the same intuition as us. The problem, however, is that the only classical argument defeating subargument $\texttt{rselp},\ \texttt{rselp} \Rightarrow \texttt{elp}$ is self-defeating, and in Pollock's original formalism, self-defeating arguments cannot prevent other arguments from becoming justified. How does one deal with this problem?

At first, Pollock tries to find the solution by generalizing the notion of self-defeat [Pol91], but later he retreats from this approach and instead tries to solve it using a multiple status assignment [Pol95]. Prakken and Vreeswijk [PV02] show that this approach basically boils down to implementing Dung's preferred semantics [Dun95]. Another difference is that in Pollock's new approach, only the *last* conclusion of an argument can be used to defeat arguments.

Under preferred semantics, O's argument (Pink elephant example) is not part of any extension, since it defeats itself and has no other defeaters. As O's argument defeats P's argument, the latter is also not in any preferred extension. Thus, P's argument cannot be ultimately undefeated, and `elp` neither `eip` is justified.

There is some controversy about whether `elp` should or should not be defeated outright. Prakken and Vreeswijk argue that although `eip` should be defeated, `elp` could also be undefeated, because its only defeater is a self-defeating argument, and `eip` is not a deductive consequence of `elp` [PV02]. Perhaps the best way to see why `elp` should be defeated is by means of a dialogue.

P: The elephant besides Robert looks pink, because Robert says so.

O: But if the elephant looks pink, then it probably also *is* pink, don't you think?

P: (cannot give any good reason for denying this inference) Euhh, yes...

O: But you know that Robert becomes unreliable in the presence of pink elephants, so how can you maintain that the elephant looks pink in the first place?

P: (understands that his statement `elp` has lost grounds) Euhh...

The point is that a rational agent that beliefs `elp` and allows for its beliefs to be critically questioned, will soon find out that its belief `elp` lacks any solid base. As this holds for any rational agent with this belief, `elp` should not be justified.

Unfortunately, there also exist examples that are handled in a somewhat less intuitive way by Pollock's new system. Take, for instance, the tax-relief example.

Example 13 (tax relief, continued)
$\texttt{INPUT} = \{\texttt{pmp}\}$
$\texttt{PFREASONS} = \{\texttt{pmp} \Rightarrow \texttt{tr},\ \texttt{tr} \Rightarrow \texttt{bd},\ \texttt{bd} \Rightarrow \texttt{fb},\ \texttt{fb} \Rightarrow \neg\texttt{tr}\}$
Here, there exists the following argument.

$\texttt{pmp}(1),\ \texttt{pmp} \Rightarrow \texttt{tr}(2),\ \texttt{tr} \Rightarrow \texttt{bd}(3),\ \texttt{bd} \Rightarrow \texttt{fb}(4),\ \texttt{fb} \Rightarrow \neg\texttt{bd}(5)$

Here, argument $(1,2,3,4,5)$ *defeats all of its subarguments that contain the prima facie reason* $\texttt{pmp} \Rightarrow \texttt{tr}$ *(including itself), and argument* $(1,2)$ *defeats argument* $(1,2,3,4,5)$*. There exists only one preferred extension, which contains argument* $(1,2,3,4,5)$ *as well as all its subarguments. Thus, in Pollock's terminology [Pol95],* `pmp`*,* `tr`*,* `bd`*,* `fb` *are ultimately undefeated and* $\neg\texttt{tr}$ *is defeated outright.*

The fact that in Pollock's new system `tr`, `bd` and `fb` are justified means that the tax-relief problem is dealt with in a structurally different way than the pink elephant problem, even though both of them are based on self-defeating arguments of the type "head bites body". Furthermore, one could describe the same kind of small conversation in which a person is confronted with the untenability of its standpoint. Therefore, we believe both examples should be dealt with in a uniform way.

4 Analysis

At first sight, implementing Socratic-style arguments appears to involve all kinds of difficulties related to the handling of self-defeating arguments. However, this does not necessarily need to be the case.

Before providing a technical solution, we will first provide an analysis of informal Socratic-style argumentation. For our purposes, the most appropriate way to do so is by means of semi-formal dialogues, as these are close to how people actually argue.

Notice that the aim of this section is not to provide a fully defined dialogue system for Socratic reasoning — although the task of doing so may be an interesting topic for future research. Instead, the main objective of our treatment of Socratic dialogues is to informally illustrate some of the concepts that play a role in them. This discussion then serves as a basis for stating the principles on which a fully formal notion of Socratic-style arguments will be based (Section 5).

In the following examples, we use the dialogue moves as has been described by [Mac79]. To enhance the readability of the examples, we also use an explicit "concede" statement, with with a party indicates agreement with the other party. To illustrate the workings of a dialogue system take the tax-relief example mentioned earlier. Suppose the proponent wants to defend that there will be a tax-relief and the opponent asks for the reasons for this but does not argue against it. Then the dialogue would look as follows:

Example 14

P:	*claim* $\mathtt{tr}$	$C_P(\mathtt{tr})$	*"I think that* $\mathtt{tr}$*."*
O:	*why* $\mathtt{tr}$		*"Why do you think so?"*
P:	*because* $\mathtt{pmp} \Rightarrow \mathtt{tr}$	$C_P(\mathtt{pmp}, \mathtt{tr})$	*"Because of* $\mathtt{pmp}$*."*
O:	*concede* $\mathtt{tr}$	$C_O(\mathtt{tr})$	*"OK, you are right."*

Each move in a dialogue game consists of a speech act, like claim (for claiming a proposition), why (for questioning a proposition), because (for supporting a proposition) or concede (for admitting a proposition endorsed by the other party). A central notion in a dialogue system is that of a *commitment*. A commitment is a party's "official" standpoint in the dialogue, it is what the party is bound to defend when it is questioned or attacked [WK95].

In the above dialogue the opponent concedes the main claim, so the proponent wins the dialogue. If, during the cause of a dialogue, parties can confront each other with the (defeasible) consequences of their opinions, then a different

dialogue may result:

Example 15

P:	*claim* $\texttt{tr}$	$C_P(\texttt{tr})$	*"I think that* $\texttt{tr}$*."*
O:	*but-then* $\texttt{tr} \Rightarrow \texttt{bd}$	$C_O(C_P(\texttt{bd}))$	*"Then you implicitly also hold that* $\texttt{bd}$*."*
P:	*concede* $\texttt{bd}$	$C_P(\texttt{tr}, \texttt{bd})$	*"Yes I do."*
O:	*but-then* $\texttt{bd} \Rightarrow \texttt{fb}$	$C_O(C_P(\texttt{fb})$ [3]	*"Then you implicitly also hold that* $\texttt{fb}$*."*
P:	*concede* $\texttt{fb}$	$C_P(\texttt{tr}, \texttt{bd}, \texttt{fb})$	*"Yes I do."*
O:	*but-then* $\texttt{fb} \Rightarrow \neg\texttt{tr}$	$C_O(C_P(\neg\texttt{tr}))$	*"Then you implicitly also hold that* $\neg\texttt{tr}$*."*
P:	*concede* $\neg\texttt{tr}$	$C_P(\texttt{tr}, \texttt{bd}, \texttt{fb}, \neg\texttt{tr})$	*"Oops, you're right; I caught myself in..."*

Here, much akin to a Socratic dialogue, the opponent wins the dialogue because it forces the proponent to commit himself to an inconsistency.

A key feature in the above dialogue is the *but-then* statement, with which the opponent confronts the proponent with the defeasible consequences of the proponent's commitments. A but-then statement is a special form of claim, in which the speaker does not become committed himself to the consequent of the rule being claimed applicable. In general, in order to use a "but-then $\varphi_1 \wedge \ldots \wedge \varphi_n \Rightarrow \phi$", the other party has to be committed to $\varphi_1 \wedge \ldots \wedge \varphi_n$. The immediate aim of a but-then statement is to commit him to ϕ as well. The final aim is then to get the other party to the point where it is obvious that his commitments are inconsistent.

Notice that the immediate effect of a but-then statement is a nested commitment, as is for instance shown on the second line of the above dialogue. Although this may appear odd at first, it is in fact the most appropriate way to describe the effects of the but-then statement in terms of commitments. When O says: "if you endorse $\texttt{tr}$ then you actually also endorse $\texttt{bd}$, don't you?" then what is it that O becomes committed to? The first thing to notice is that O does not necessarily endorse $\texttt{bd}$ himself, so it does not hold that $C_O(\texttt{bd})$. Furthermore, it goes too far to immediately have P committed to $\texttt{bd}$; the rule "$\texttt{tr} \Rightarrow \texttt{bd}$" is defeasible and P may defend himself by giving a reason (an undercutter) why this rule does not apply (an example of this will be treated further on). Therefore, it also does not hold that $C_P(\texttt{bd})$. The only thing that can be said regarding the but-then statement is that O claims the $\texttt{bd}$ is implicitly endorsed by P. Therefore, it holds that $C_O(C_P(\texttt{bd}))$.

An interesting question is how the style of reasoning of the "because" statement can be compared with that of the "but-then" statement (see also figure

[3] we no longer explicitly mention $C_O(C_P(\texttt{bd}))$ since $C_P(\texttt{bd})$

1):

(1) With the because statement, reasoning goes *backwards*; the party being questioned tries to find reasons to support its thesis. With the but-then statement, on the other hand, reasoning goes *forward*; the party being questioned can be forced to make additional reasoning steps.
(2) With the because statement, the *proponent* of a thesis (like ϕ in figure 1) tries to find a path (or tree) from the premises to ϕ (the opponent's task is then to try to defeat this path). With the but-then statement, on the other hand, it is the *opponent* of the thesis that tries to find a path (or tree).
(3) The path (or tree) constructed using because statements should ultimately originate from statements that are accepted to be *true* (such as premises), whereas the path constructed using but-then statements should ultimately lead to statements that are considered *false* (contradictions)
(4) With a successfully constructed because path (or tree), but the proponent and opponent become committed to the propositions on the path, whereas with a successfully constructed but-then path (or tree), it is possible that only the proponent becomes committed to the propositions on the path.

Fig. 1. because and but-then

In the above analysis, it appears that an opponent of ϕ has two options: either trying to construct a but-then path from ϕ, or trying to prevent the proponent from successfully constructing an undefeated because path. These strategies can sometimes also be combined.

The use of a but-then statement does not automatically lead to a new commitment on the side of the other party. Sometimes, it can be successfully argued why the counterparty does not have to become committed. To illustrate why, consider again the tax-relief example, but now with the extra information that France and Germany also have a budget deficit, and therefore have an interest in softening up the rule that a budget deficit leads to fines.[4] Thus, the rule $\texttt{bd} \Rightarrow \texttt{fb}$ can now be undercut.

Example 16

[4] And as everybody knows, France and Germany usually get their way in the EU...

P:	*claim* $\mathtt{tr}$	$C_P(\mathtt{tr})$
O:	*but-then* $\mathtt{tr} \Rightarrow \mathtt{bd}$	$C_O(C_P(\mathtt{bd}))$
P:	*concede* $\mathtt{bd}$	$C_P(\mathtt{tr}, \mathtt{bd})$
O:	*but-then* $\mathtt{bd} \Rightarrow \mathtt{fb}$	$C_O(C_P(\mathtt{fb}))$
P:	*claim* $\neg\lceil \mathtt{bd} \Rightarrow \mathtt{fb} \rceil$	$C_P(\mathtt{tr}, \mathtt{bd}, \neg\lceil \mathtt{bd} \Rightarrow \mathtt{fb} \rceil)$
O:	*why* $\neg\lceil \mathtt{bd} \Rightarrow \mathtt{fb} \rceil$	$C_O(C_P(\mathtt{fb}))$
P:	*because* $\mathtt{bd}(\mathtt{F}) \wedge \mathtt{bd}(\mathtt{G}) \Rightarrow \neg\lceil \mathtt{bd} \Rightarrow \mathtt{fb} \rceil$	$C_P(\mathtt{tr}, \mathtt{bd}, \neg\lceil \mathtt{bd} \Rightarrow \mathtt{fb} \rceil, \mathtt{bd}(\mathtt{F}) \wedge \mathtt{bd}(\mathtt{G}))$
O:	*retract* $C_P(\mathtt{fb})$*, concede* $\mathtt{tr}$	$C_O(\mathtt{tr})$

Here, the opponent again tries to construct a successful but-then path. This path, however, is undercut by the proponent. What happens next depends on the nature of the dialogue. When backtracking is allowed, the opponent may pursue another strategy. When backtracking is not allowed, the opponent loses the game.

As for the effects of the but-then statement on the commitments in the dialogue the following general remarks can be made:

(1) A but-then statement is in essence a special form of a claim statement. A claim statement has as effect that a new commitment comes into existence, and such should also be the case for a but-then statement.
(2) But-then statements do not in general create unnested commitments (at least, not immediately). Suppose party O utters "but-then $\varphi_1 \wedge \ldots \wedge \varphi_n \Rightarrow \phi$". This does of course not mean that O becomes committed to ϕ (so we don't have $C_O(\phi)$. It also does not mean that P is actually committed to ϕ (that is, we don't automatically have $C_P(\phi)$), because P may avoid commitment by successfully defending ψ_i $(1 \leq i \leq m)$ The only thing thing that can be said is that O feels that P is implicitly committed to ϕ (so $C_O(C_P(\phi))$), but whether P is actually committed to ϕ is still open for discussion.
(3) In general, the party that makes a claim bears the responsibility of defending this claim. For instance, if P utters "claim ϕ" then upon P rests the task of defending ϕ. Similarly, if O utters "but-then $\varphi_i \wedge \ldots \wedge \varphi_n \Rightarrow \phi$" then upon O rests the task of defending $C_P(\phi)$ by making sure that P cannot avoid the conclusion ϕ. If O is unable to do so, it can loose the dialogue game.

To summarise: nested commitment is quite a natural concept to use in dialogues to enable parties to be confronted with the consequences of their own commitments.

As an aside, it may be interesting to compare the but-then statement with the resolve statement of MacKenzie's DC [Mac79]. In DC, if party A claims proposition q ("claim q"), which is then questioned ("why q") by party B, and party B is committed to p, from which q directly follows, then party A may utter "resolve $p \supset q$", which forces party B to become committed to q as well (or alternatively, B may retract its commitment to p). [5] See the following example.

P:	claim p	$C_P(p)$
O:	concede p	$C_O(p)$
P:	claim $p \supset q$	$C_P(p,\ p \supset q)$
O:	concede $p \supset q$	$C_O(p,\ p \supset q)$
P:	claim q	$C_P(p,\ p \supset q,\ q)$
O:	why q	[unchanged]
P:	resolve "If $p \wedge (p \supset q)$ then q"	[unchanged]
O:	concede q	$C_O(p,\ p \supset q,\ q)$

One obvious difference between the resolve and the but-then statement is that after a successful resolve statement *both* parties are committed to the proposition in question, whereas with a successful but-then statement it is possible that only one party becomes committed. Furthermore, the but-then statement also has an inherently defeasible nature; it is possible that the other party has a reason (exception) against applying the rule in question. This reason can then be discussed in the remainder of the dialogue. In short, although the resolve statement can be sufficient for classical, monotonic reasoning, for defeasible reasoning it is desirable to have a statement that has special, nonmonotonic properties. We think that the but-then statement fulfills this requirement.

There also exists an interesting difference between the but-then statement and the claim statement. In theory, it would be possible to replace a statement "but-then `tr` $\Rightarrow$ `bd`" by "claim `tr` $\supset$ `bd`" (for instance in example 15) and then question the other party regarding `bd`. The difference, however, is that of burden of proof. With a claim statement, the burden of proof is on the party issuing the claim, whereas with a but-then statement, it is the *other* party who is responsible to give explicit reasons why the rule in question would not be applicable to the current situation. In example 15, this would mean that

[5] Another use of MacKenzie's resolve statement is to notice the other party that its commitments are inconsistent, thus forcing the other party to retract one or more of them.

the proponent would have to come up with reasons against the applicability of $\texttt{tr} \Rightarrow \texttt{bd}$, if he/she wants to avoid becoming committed to bd. This is quite different from the situation where $\texttt{tr} \supset \texttt{bd}$ would simply be claimed, given all kinds of possibilities to contest this claim. The but-then statement is most appropriate in the case of a (defeasible) rule whose *general* applicability is held by both the proponent and the opponent.

4.1 Self-defeat

The treatment of the but-then statement and the notion of nested commitments is interesting not only because these are issues worthwhile of a treatment in itself, but also because they can shed new light on the issue of self-defeat. Recall that in Section 3.1 Socratic-style arguments were represented as self-defeating arguments. The discussion above, however, gives reason to believe that this may not be the right approach. Our point is not so much based on the technical difficulties related to self-defeating arguments, but is an intuitive one: the Socratic-style counterarguments, as they occur in actual human-to-human conversations, are essentially not self-defeating.

To see why this is, consider again the tax-relief dialogue (example 15) and ask oneself the question: does O contradict himself? Although O manages to let P commit himself to a contradiction, O itself is not at any moment committed to an inconsistency. It merely confronts P with the consequences of its own reasoning without endorsing these consequences himself.

More general, consider the case of a critical interview. The interviewer may question his guest into the direction of an inconsistency without endorsing this inconsistency himself. Or consider Socrates, who proclaimed to have no knowledge at all. During the discussions with his fellow citizens he was rarely reported to take any position himself. It were his discussion partners who committed themselves to inconsistencies, not Socrates himself.

It is therefore our view that, if Socratic-style reasoning is to be modelled by formal argumentation, the resulting Socratic-style arguments should not be self-defeating.

4.2 Principles of S-arguments

A Socratic-style argument (S-argument) can preliminary be defined as an argument that illustrates the problematic nature of another argument by assuming one or more of the other argument's conclusions. An example argument discourse featuring an S-argument is the following:

Example 17 (shipment of goods, continued)
P: tma, tma ⇒ a
O: ¬is, ¬is ⇒ ¬cd, a, a ⇒ cd

Another example of such an argument-game is the following:

Example 18 (Ajax-Feijenoord, continued)
P: af, af ⇒ t
O: t, t ⇒ p, p ⇒ ¬⌈af ⇒ t⌉

A *foreign commitment* of an S-argument is a proposition that is a conclusion of another argument that is used as an assumption in the S-argument. Foreign commitments are called like that because they are based on the actual commitments in the other arguments.

In example 17, proposition a in O's argument is a foreign commitment that has its origin in P's argument. In example 18, proposition t in O's argument is a foreign commitment that has its origin in P's argument.

A conclusion of an argument is *fc-based* iff it is based on one or more foreign commitments. For instance, in example 17 O's proposition cd is fc-based, while O's proposition ¬cd is not fc-based. We use the convention of representing fc-based proposition in gray.

Notice that the proposed principles for formalizing S-arguments do not allow for an unlimited depth of nesting of commitments. That is, we do not allow utterances like "I endorse that you endorse that I endorse...". These constructs are rarely seen in actual discussions and in our view not worthwhile the effort of overcoming the technical difficulties associated with them (see [Cam04] for a discussion). We will therefore limit the depth of nested commitments to at most two. This can be implemented by requiring that every foreign commitment has its origin in another argument's conclusion that is itself not fc-based.

With respect to the above discussion, the following general principles regarding S-arguments can be stated:

(1) An S-argument contains at least one foreign commitment, that is, a conclusion imported from another argument. All foreign commitments should have their origin in conclusions (of the other argument) that are themselves not fc-based.
(2) An S-argument A_2 S-rebuts an argument A_1 if
    (a) all foreign commitments of A_2 have their origin in A_1
    (b) A_2 contains conflicting conclusions, of which at least one conclusion is fc-based
(3) An S-argument A_2 S-undercuts an argument A_1 if

(a) all foreign commitments of A_2 have their origin in A_1
(b) A_2 contains an fc-based conclusion that undercuts some rule in A_1

5 Formalization

As shown in [Cam04], S-arguments can be implemented in various formalisms for defeasible reasoning, including an argument-theoretic version of default logic [Rei80] and the formalism of Prakken and Sartor [PS97]. In order to keep things concise we will, however, restrict ourselves to the formalism of Pollock.

As for the system of Pollock, the first thing to notice is that S-arguments are in fact based on a specific kind of suppositional reasoning. With an S-argument one supposes one or more of the commitments of the other party in order to derive something that undermines the other party's position (either a contradiction or an undercutter of the other party's original argument). Unfortunately, this particular form of suppositional reasoning is not supported in Pollock's framework for defeasible reasoning. This can be illustrated using the following example.

Example 19
$\texttt{INPUT} = \{p, \neg r\}$
PFREASONS $= \{p \Rightarrow q,\ q \Rightarrow r\}$

There now exists an argument for q (argument I):

1. $\langle \emptyset, p, \infty \rangle$ *($p \in$ INPUT)*
2. $\langle \emptyset, q, \alpha \rangle$ *(p is a prima facie reason for q)*

An S-argument would then suppose q and derive a contradiction (argument II).

1. $\langle \{q\}, q, \infty \rangle$ *(supposition)*
2. $\langle \{q\}, r, \alpha \rangle$ *(q is a prima facie reason for r)*
3. $\langle \emptyset, \neg r, \infty \rangle$ *($\neg r \in$ INPUT)*

Argument II, however, is self-defeating and does not prevent argument I from becoming justified, neither in Pollock's old system, nor in Pollock's new system.

So, even though Pollock's system supports suppositional reasoning, it does not support the specific suppositional reasoning required for S-arguments. It is clear that in order to allow for S-arguments, Pollock's framework needs to be extended.

In order to incorporate S-arguments in Pollock's framework, one of the things that is needed is the ability to represent foreign commitments, as well as a mechanism to determine whether a certain conclusion is fc-based or not. An obvious choice would be to use the facilities for suppositional reasoning for this. The precise definitions of suppositional reasoning would then have to be adjusted to suit the special requirements of Socratic-style argumentation. In order to keep things simple and to keep focus on our main point, we assume that all classical (non S) arguments are linear, thus limiting suppositional reasoning to S-arguments only.

A general principle of the now coming formalization is that each line only contains the suppositions (foreign commitments) that it is actually based on. For this, it is necessary to deviate from Pollock's original definition of an argument. First of all, the **input** rule should produce a line with an empty supposition, the **foreign commitment** rule should produce a line whose supposition is a singleton, and the **reason** rules should take the union of the suppositions of the lines they use. For the application of conclusive reasons we require that the entire antecedent is actually needed to support the consequent.

Another design consideration is that the foreign commitments are conclusions of the argument being attacked. Furthermore, the conclusions on which the foreign commitments are based should themselves not be based on foreign commitments.

Definition 20 *An* argument *based on* `INPUT` *is a finite sequence* $\sigma_1, \ldots, \sigma_n$*, where each* σ_i *is a line of argument. A line of argument is a triple* $\langle X_i, p_i, \nu_i \rangle$*, where* X_i *is the set of* foreign commitments *of* σ_i*,* p_i *is a proposition, and* ν_i *is the degree of justification of* σ_i*. A new line of argument is obtained from earlier lines of argument according to the following rules of argument formation.*

Input. *If* p *is in* `INPUT` *then it holds that* $\sigma, \langle \emptyset, p, \infty \rangle$ *is an argument.*

Prima facie reason. *If* σ *is an argument,* $\langle X_1, p_1, \eta_1 \rangle, \ldots, \langle X_n, p_n, \eta_n \rangle$ *are members of* σ *and* $\{p_1, \ldots, p_n\}$ *is a reason of strength* ν *for* q*, then* $\sigma, \langle X_1 \cup \ldots \cup X_n, q, \min\{\eta_1, \ldots \eta_n, \nu\} \rangle$ *is an argument.*

Conclusive reason. *If* σ *is an argument,* $\langle X_1, p_1, \eta_1 \rangle, \ldots, \langle X_n, p_n, \eta_n \rangle$ *are members of* σ *obtained by any rule of argument formation except* ***conclusive reason****,* $\{p_1, \ldots, p_n\}$ *is a conclusive reason for* q*, and for any* $p_i \in \{p_1, \ldots, p_n\} : \{p_1, \ldots, p_n\} - \{p_i\}$ *is not a conclusive reason for* q*, then* $\sigma, \langle X_1 \cup \ldots \cup X_n, q, \min\{\eta_1, \ldots, \eta_n\} \rangle$ *is an argument.*

Foreign commitment. *If* σ *is an argument and* p *is a proposition, then* $\sigma, \langle \{p\}, p, \infty \rangle$ *is also an argument.*

To avoid redundancy, we require that different lines always have different propositions.

Definition 21 *An argument* σ classically rebuts *an argument* η *iff:*

(1) η *contains a line of the form* $\langle \emptyset, q, \alpha \rangle$ *that is obtained by the argument formation rule* prima facie reason *from some earlier lines* $\langle \emptyset, p_1, \alpha_1 \rangle, \ldots, \langle \emptyset, p_n, \alpha_n \rangle$ *where* $\{p_1, \ldots p_n\}$ *is a prima facie reason for* q*, and*
(2) σ *contains a line of the form* $\langle \emptyset, \neg q, \beta \rangle$ *where* $\beta \geq \alpha$*.*

Definition 22 *An argument* σ classically undercuts *an argument* η *iff:*

(1) η *contains a line of the form* $\langle X, q, \alpha \rangle$ *that is obtained by the argument formation rule* reason *from some earlier lines* $\langle X_1, p_1, \alpha_1 \rangle, \ldots, \langle X_n, p_n, \alpha_n \rangle$ *where* $\{p_1, \ldots p_n\}$ *is a prima facie reason for* q*, and*
(2) σ *contains a line of the form* $\langle \emptyset, \neg \lceil \{p_1, \ldots, p_n\} \Rightarrow q \rceil, \beta \rangle$ *where* $\beta \geq \alpha$*.*

Definition 23 *An argument* σ S-rebuts *an argument* η *iff:*

(1) σ *contains a line of the form* $\langle X_1, L, \beta_1 \rangle$ *and a line of the form* $\langle X_2, \neg L, \beta_2 \rangle$ *where* $X_1 \neq \emptyset$ *or* $X_2 \neq \emptyset$*, and*
(2) for each $f_i \in X_1 \cup X_2$ *it holds that there is a line in* η *of the form* $\langle \emptyset, f_i, \alpha_i \rangle$*, and*
(3) it holds that $\min\{\beta_1, \beta_2\} \geq \min\{\alpha_1, \ldots, \alpha_n\}$*.*

If argument σ *S-rebuts argument* η *and* $\min\{\beta_1, \beta_n\} = \min\{\alpha_1, \ldots, \alpha_n\}$ *then* η reverse S-rebuts σ*.*

The notion of reverse S-rebuttal in Definition 23 has been introduced to make S-rebutting symmetrical in the case of equal argument-strengths. This makes it similar to classical rebutting, which is also symmetrical when argument-strengths are equal.

Definition 24 *An argument* σ S-undercuts *an argument* η *iff:*

(1) η *contains a line of the form* $\langle X, q, \alpha \rangle$ *that is obtained by the argument formation rule* prima facie reason *from some earlier lines* $\langle X_1, p_1, \alpha_1 \rangle, \ldots, \langle X_n, p_n, \alpha_n \rangle$ *where* $\{p_1, \ldots p_n\}$ *is a prima facie reason for* q*, and*
(2) σ *contains a line of the form* $\langle Y, \neg \lceil \{p_1, \ldots, p_n\} \Rightarrow q \rceil, \beta \rangle$ *with* $Y \neq \emptyset$*, and*
(3) for each $f_i \in Y$ *it holds that there is a line in* η *of the form* $\langle \emptyset, f_i, \alpha_i \rangle$*, and*
(4) it holds that $\beta \geq \min\{\alpha_1, \ldots, \alpha_n\}$*.*

Definition 25 *An argument* σ defeats *an argument* η *iff:*

(1) σ *classically rebuts* η*, or*
(2) σ *classically undercuts* η*, or*
(3) σ *S-rebuts* η*, or*
(4) σ *reverse S-rebuts* η*, or*

(5) σ S-undercuts η.

The formalism of Definition 20 until 25 will be referred to as the S-enriched Pollock system.

It must be noticed that if one restricts oneself to linear arguments without foreign commitments, the S-enriched Pollock system is equivalent to Pollock's old system with linear arguments only. This can be seen as follows. First, if no foreign commitments are allowed then the remaining arguments are all linear (see Definition 20). Also, without foreign commitments, arguments cannot S-rebut, reverse S-rebut or S-undercut each other. Thus, on the level of a Dung-style argumentation framework, what the above definitions do is that they take the existing set of arguments and the defeat-relation, and add new arguments (S-arguments) and extend the defeat-relation. In this way, the S-enriched set of arguments and defeat relation are each a superset of the original set of arguments and defeat relation (more on this in Section 6).

5.1 Examples

In order to see how the S-enriched system works, consider the following examples.

Example 26 (Ajax-Feijenoord, continued)
$\texttt{INPUT} = \{\texttt{af}\}$
$\texttt{PFREASONS} = \{\texttt{af} \Rightarrow \texttt{t},\ \texttt{t} \Rightarrow \texttt{p},\ \texttt{p} \Rightarrow \neg\lceil\texttt{af} \Rightarrow \texttt{t}\rceil\}$

P: $\texttt{af},\ \texttt{af} \Rightarrow \texttt{t}$
O: $\texttt{t},\ \texttt{t} \Rightarrow \texttt{p},\ \texttt{p} \Rightarrow \neg\lceil\texttt{af} \Rightarrow \texttt{t}\rceil$

Here, argument con is an undercutting S-argument against argument pro.

Example 27 (tax-relief, continued)
$\texttt{INPUT} = \{\texttt{pmp}\}$
$\texttt{PFREASONS} = \{\texttt{pmp} \Rightarrow \texttt{tr},\ \texttt{tr} \Rightarrow \texttt{bd},\ \texttt{bd} \Rightarrow \texttt{fb},\ \texttt{fb} \Rightarrow \neg\texttt{tr}\}$

P: $\texttt{pmp},\ \texttt{pmp} \Rightarrow \texttt{tr}$
O: $\texttt{tr},\ \texttt{tr} \Rightarrow \texttt{bd},\ \texttt{bd} \Rightarrow \texttt{fb},\ \texttt{fb} \Rightarrow \neg\texttt{tr}$

Here, the argument con is a rebutting S-argument against the argument pro.

6 Semantical issues

It is interesting to examine how the notion of S-arguments fits into Dung's abstract argumentation theory [Dun95]. A Dung-style argumentation framework consists of a pair $(Args, def)$ where $Args$ is a set of arguments and def is a defeat relation between the arguments. We write $A_2 \, def \, A_1$ (A_2 defeats A_1) when $(A_2, A_1) \in def$. Based on the notion of an argument framework, Dung defines several principles ("semantics") for determining whether an argument is overall justified. In this section we will focus on grounded semantics.

To illustrate the working of an argumentation framework, consider the following example.

$\texttt{INPUT} = \{\texttt{A}\}$

$\texttt{PFREASONS} = \{\texttt{A} \Rightarrow \texttt{B},\ \texttt{B} \Rightarrow \neg\texttt{A},\ \texttt{A} \Rightarrow \texttt{C}\}$

Now, if we assume a formalism whose arguments and defeat relation are purely classical,[6] this results in an argumentation framework containing arguments for each combination of prima facie reasons (we leave out arguments consisting of the same rules applied in different order).

Arguments:

(A_1) $\texttt{A}$

(A_2) $\texttt{A},\ \texttt{A} \Rightarrow \texttt{B}$

(A_3) $\texttt{A},\ \texttt{A} \Rightarrow \texttt{C}$

(A_4) $\texttt{A},\ \texttt{A} \Rightarrow \texttt{B},\ \texttt{B} \Rightarrow \neg\texttt{A}$

(A_5) $\texttt{A},\ \texttt{A} \Rightarrow \texttt{B},\ \texttt{A} \Rightarrow \texttt{C}$

(A_6) $\texttt{A},\ \texttt{A} \Rightarrow \texttt{B},\ \texttt{B} \Rightarrow \neg\texttt{A},\ \texttt{A} \Rightarrow \texttt{C}$

Defeat relation:
$A_4 \ def \ A_1,\ A_4 \ def \ A_2,\ A_4 \ def \ A_3,\ A_4 \ def \ A_4,\ A_4 \ def \ A_5,\ A_4 \ def \ A_6$
$A_6 \ def \ A_1,\ A_6 \ def \ A_2,\ A_6 \ def \ A_3,\ A_6 \ def \ A_4,\ A_6 \ def \ A_5,\ A_6 \ def \ A_6$

This argumentation framework is graphically depicted in figure 2.

Now, if the argumentation formalism is changed to include S-arguments, then this introduces a new class of arguments. Basically, one can now construct S-arguments that include foreign commitments from the classical arguments (as is stated in definition 23 In the above example, some of these new arguments are:

[6] For simplicity, we do not yet take any conclusive reasons into account in this example.

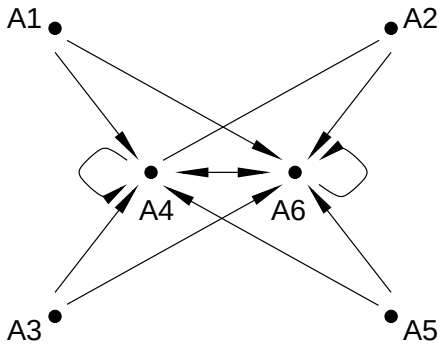


Fig. 2. Argumentation framework without S-arguments.

(A_7) B, B $\Rightarrow \neg$A, **A**

(A_8) B, B $\Rightarrow \neg$A, A

(A_9) $\neg$A, **A**

(A_{10}) $\neg$A, A

With the new arguments, the defeat relation is also extended:

A_2 *def* A_7, A_4 *def* A_7, A_6 *def* A_7, A_5 *def* A_7
A_7 *def* A_2, A_7 *def* A_4, A_7 *def* A_6, A_7 *def* A_5

A_2 *def* A_8, A_4 *def* A_8, A_6 *def* A_8, A_5 *def* A_8
A_8 *def* A_2, A_8 *def* A_4, A_8 *def* A_6, A_8 *def* A_5

A_4 *def* A_9, A_6 *def* A_9
A_4 *def* A_{10}, A_6 *def* A_{10}

The new S-enriched argumentation framework is depicted in figure 3.

Thus, one can see that the concept of S-arguments is compatible with the notion of a Dung-style argumentation framework. Essentially, what happens is that both the set of arguments and the defeat relation among these arguments are extended. The extension is done in such a way that if one would leave out the S-arguments, as well as each instance of the defeat relation involving at least one S-argument, the result would be an argumentation framework of an argumentation formalism that allows only classical arguments and classical defeat.

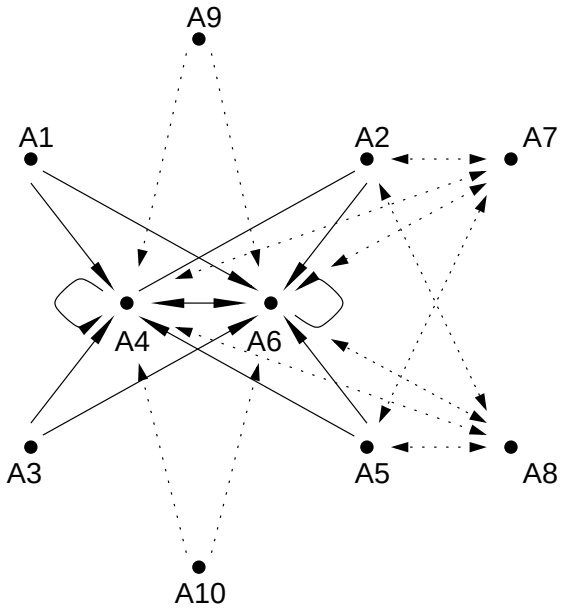


Fig. 3. Argumentation framework with S-arguments.

6.1 S-arguments and the notion of justified arguments

Given an argumentation framework, the next step is to determine which arguments are considered *justified*. As our aim is to stay relatively close to Pollock's original system, we assume grounded semantics, which can be summarized as follows.

Definition 28 *Let* AF $= (Args, def)$ *be an argumentation framework in which every argument is defeated by at most a finite number of arguments. Consider the following sequence of sets of arguments:*

- $F^0 = \emptyset$
- $F^{i+1} = \{A \in Args \mid$ *for each* $B \in Args$ *such that* B *def* A *there exists a* $C \in F^i$ *such that* C *def* $B\}$

An argument is justified *under grounded semantics iff it is in* $\cup_{i=0}^{\infty}(F^i)$.

This definition can be applied on figure 2 as follows.

$$
\begin{aligned}
F^0 &= \emptyset \\
F^1 &= \{A \in Args \mid A \text{ is acceptable with respect to } \emptyset\} \\
&= \{A \in Args \mid A \text{ has no arguments defeating it }\} \\
&= \{A_1, A_2, A_3, A_5\} \\
F^2 &= \{A \in Args \mid A \text{ is acceptable with respect to } F^1\} \\
&= \{A \in Args \mid \text{ every argument defeating } A \text{ is defeated by } \{A_1, A_2, A_3, A_5\}\}
\end{aligned}
$$

$$
\begin{aligned}
&= \{A_1, A_2, A_3, A_5, A_6\} \\
F^{i+1} &= F^i \text{ (with } i \geq 2)
\end{aligned}
$$

This means that $\cup_{i=0}^{\infty} = \{A_1, A_2, A_3, A_5\}$, so A_1, A_2, A_3 and A_5 are considered justified under grounded semantics.

If we look at the S-enriched formalism, on the other hand, then a different outcome results. Intuitively, with S-arguments, only $\mathtt{A}$ and $\mathtt{C}$ should become justified, since the argument for $\neg\mathtt{A}$ ($\mathtt{A},\ \mathtt{A} \Rightarrow \mathtt{B},\ \mathtt{B} \Rightarrow \neg\mathtt{A}$) is incoherent and the argument for $\mathtt{B}$ ($\mathtt{A},\ \mathtt{A} \Rightarrow \mathtt{B}$) has a S-counterargument ($\mathtt{B},\ \mathtt{B} \Rightarrow \neg\mathtt{A},\ \mathtt{A}$).

It is interesting to examine what happens when grounded semantics is applied straightforwardly to the S-enriched argumentation framework of figure 3.

$$
\begin{aligned}
F^0 &= \emptyset \\
F^1 &= \{A \in Args \mid A \text{ is acceptable with respect to } \emptyset\} \\
&= \{A \in Args \mid A \text{ has no arguments defeating it }\} \\
&= \{A_1, A_3, A_9, A_{10}\} \\
F^2 &= \{A \in Args \mid A \text{ is acceptable with respect to } F^1\} \\
&= \{A \in Args \mid \text{ every argument defeating } A \text{ is defeated by } \{A_1, A_3, A_9, A_{10}\}\} \\
&= \{A_1, A_3, A_9, A_{10}\} \\
F^{i+1} &= F^i \text{ (with } i \geq 2)
\end{aligned}
$$

At least, this result is partly in line with what one expects. A_2 and A_5 are not justified anymore because they now have S-counterarguments against them. A_1 and A_3, on the other hand, remain justified, as is desired. This result, however, comes with a price, since not only A_1 and A_3, but also the S-arguments A_9 and A_{10} become justified. Worse yet, if one would just take the conclusions of justified arguments (where the conclusions are simply all propositions derived in the argument, as is for instance the approach in [PS97]) then from $\neg\mathtt{A},\ \mathtt{A}$ being a justified argument, it would follow that both $\mathtt{A}$ and $\neg\mathtt{A}$ would become justified conclusions!

Applying (grounded) semantics to an S-enriched argumentation framework therefore involves two problems: (1) S-arguments that can become justified and (2) the conclusions of S-arguments that can become justified.

Let us first consider the point that S-arguments can become justified. In the examples in Section 5.1 and elsewhere, we saw that S-arguments are meant as *counter*arguments. That is, they are not meant to yield conclusions on their own, but instead to prevent other conclusions from becoming justified. An S-argument is context-dependent; it *needs* an argument that it defeats. One can

say that an S-argument does not have a meaning when standing on its own. A dialogue, for example, provides a proper environment for S-arguments to make sense. The semantics as stated by Dung, on the other hand, treat arguments as having an existence that is independent from any other argument. Dung's semantics tries to answer the question "Is S-argument A justified?", but this question does not make any sense if one considers that S-arguments cannot stand on their own. It is simply not applicable to call S-arguments "justified", "defeated" or "defensible", as these terms are applicable to classical arguments only. A possible solution would therefore be to reserve the terms "justified", "defeated" or "defensible" only to classical arguments, while leaving the rest of the specification of grounded semantics unchanged.

The second point to consider is that of justified conclusions. If our ultimate interest is in the justified conclusions, and we regard arguments only as a technical intermediate step to yield these conclusions, than in a certain sense it does not matter which arguments are justified and which are not, a long as we have the "right" justified conclusions. If we apply the notions of justified arguments as in Definition 28 without any changes, then there are two alternatives for defining when a conclusion is justified or not.

As the grounded extension is always conflict-free, it can easily be seen that the set of conclusions of classical arguments is always consistent. An obvious choice would therefore to take only the conclusions of the classical justified arguments as justified.

Definition 29 *A formula is* justified *iff it is a conclusion of a classical argument (that is, an argument without foreign commitments) that is justified under grounded semantics.*

7 On the issue of self-defeat

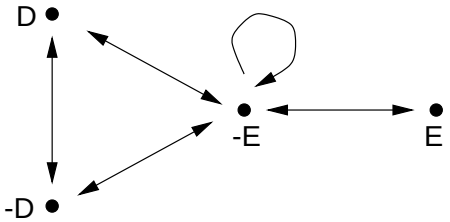
One particular subtle issue in formal argumentation is that of self-defeat. In Pollock's old system, self-defeating arguments are automatically rejected and cannot influence the status of other arguments (Definition 5). One of the reasons for this is informal: someone who contradicts himself should not be taken serious and should not be able to keep other arguments from becoming justified.

There also exists, however, a technical reason for "neutralizing" the effects of self-defeating arguments. Consider the following example.

Example 30
$\texttt{INPUT} = \{\texttt{A}, \texttt{B}, \texttt{C}\}$

$\texttt{PFREASONS} = \{\texttt{A} \Rightarrow \texttt{D},\ \texttt{B} \Rightarrow \neg\texttt{D},\ \texttt{C} \Rightarrow \texttt{E}\}$
There now exists an argument against E*:*
$\texttt{A},\ \texttt{A} \Rightarrow \texttt{D},\ \texttt{B} \Rightarrow \neg\texttt{D},\ \texttt{D} \wedge \neg\texttt{D} \rightarrow \neg\texttt{E}$

The interaction in Pollock's old system between the arguments for D, ¬D, ¬E and E is shown in figure 7. The argument for ¬E now prevents E from becoming justified, even though intuitively E has nothing to do with the conflict between D and ¬D.



The point is that, when one allows self-defeating arguments in combination with first-order conclusive reasoning, it is possible to combine two arguments that rebut each other into an argument that can defeat any arbitrary argument. This is especially problematic when grounded semantics is being applied (like in Pollock's old system), since the grounded extension would be empty. As is explained in [Cam05], this situation is somewhat better in Pollock's new (preferred semantics based) formalism, but is still not completely solved by it.

The approach of ruling out self-defeating arguments or neutralizing their effects is not just taken in Pollock's system, but also in a wide range of other formalisms for defeasible argumentation [PS97,GS04,BH01]. It is therefore interesting to examine how S-arguments deal with the issue of self-defeat.

The first thing to notice is that in our S-enhanced formalism, self-defeat is limited to classical rebutting and classical undercutting.

Lemma 31 *For each argument σ in the S-enriched Pollock system such that σ defeats itself, it holds that σ classically rebuts itself or σ classically undercuts itself.*

PROOF. The fact that σ defeats σ means (Definition 25) that either σ classically rebuts itself, σ classically undercuts itself, σ S-rebuts itself, σ reverse S-rebuts itself, or σ S-undercuts itself.

Suppose σ S-undercuts itself. Then (Definition 24 (1)) σ contains a line of the form $\langle Y, \neg\lceil\{p_1, \ldots, p_n\} \Rightarrow q\rceil, \beta\rangle$ with $Y \neq \emptyset$. Then (Definition 24 (2)), σ also contains at least one line $\langle\{f_i\}, f_i, \infty\rangle$ for some $f_i \in Y$. But Definition 24 (3) also requires that σ contains a line $\langle\emptyset, f_i, \alpha_i\rangle$. This, however, conflicts with the definition of an argument (Definition 20), where it is required that different lines have different propositions. Contradiction.

Thus, σ cannot S-undercut itself. For similar reasons, σ also cannot S-rebut itself (which implies that σ also cannot reverse S-rebut itself). Thus, the only remaining possibilities are that σ classically rebuts itself, or that σ classically undercuts itself. $\Box$

The fact that self-defeat is limited to classical arguments means that every self-defeating argument is defeated by an argument that is itself undefeated, as is stated in the following theorem.

Theorem 32 *If an argument σ in the S-enriched Pollock system defeats itself, then there exists an argument η that defeats σ and is not defeated by any argument.*

PROOF. Let σ be an argument such that σ defeats itself. Then, according to lemma 31, σ either classically rebuts itself or classically undercuts itself. If σ classically rebuts itself then (Definition 21) σ contains a line $\langle \emptyset, q, \alpha \rangle$ and a line $\langle \emptyset, \neg q, \beta \rangle$. Then the argument $\eta = (\langle \{q\}, q, \infty \rangle, \langle \{\neg q\}, \neg q, \infty \rangle)$ S-rebuts σ and is not defeated by any argument. If σ classically undercuts itself, then (Definition 22) σ contains a line of the form $\langle \emptyset, \neg\lceil \{p_1, \ldots, p_n\} \Rightarrow q \rceil, \beta \rangle$. Then the one-line argument $\eta = (\langle \{\neg\lceil \{p_1, \ldots, p_n\} \Rightarrow q \rceil, \beta\}, \neg\lceil \{p_1, \ldots, p_n\} \Rightarrow q \rceil, \alpha \rangle)$ S-undercuts σ and is not defeated by any argument. $\Box$

Now again consider the criterion for determining the justified arguments. In Definition 5, self-defeating arguments were explicitly ruled out, so that they do not prevent other arguments from becoming justified. It is interesting to see that with S-arguments, it is no longer needed to rule out self-defeating arguments in order to neutralize their effects. Let us assume that in Definition 5 *all* arguments are in at level 0 (including any self-defeating arguments). Now suppose σ is a self-defeating argument. Then, according to theorem 32, there exists an argument η that defeats σ and is itself undefeated. The fact that η is not defeated by any argument means that η is *in* at every level. This, however, also means that σ is *out* and stays *out* at every level starting from level 1, and will thus not keep other arguments from becoming justified.

The point is that with S-arguments, there is no need for a "hacked" version of grounded semantics (like Definition 5) or for Pollock's new system, at least not in order to neutralize the effects of self-defeating arguments. With S-arguments, ordinary grounded semantics (as, for instance, stated in Definition 28) will do fine.

8 Summary and Conclusions

Overall, one can say that Socratic-style argumentation has been known since antiquity and is still in use today in philosophic as well as in everyday reasoning. Today's generation of formalisms for defeasible reasoning, however, does not support this kind of arguments. This has been shown to be true for Pollock's formalism (as shown in this paper) but can also be illustrated using default logic [Rei80] or the formalism of Prakken and Sartor [PS97] (as shown in [Cam04]).

The approach of modelling Socratic-style arguments by self-defeating arguments is undesirable both from technical and intuitive perspective. The technical problems of using self-defeating arguments are best illustrated by the figure on page 32. From a conceptual perspective, the modelling of Socratic-style arguments by self-defeating arguments is undesirable because S-arguments are essentially not self-defeating. This is best understood by studying Socratic-style argumentation in terms of (nested) commitments (Section 4). S-arguments can, however, play a role in "neutralizing" the undesirable effects of real self-defeating arguments, as shown by theorem 32.

The notion of S-arguments can be embedded in the argumentation formalism of John Pollock, as well as in other ones (see [Cam04]). On the level of a Dung-style argumentation framework the introduction of Socratic-style argumentation amounts to extending the set of arguments as well as extending the defeat relation to include these arguments. When determining justified conclusions from the justified arguments, however, one then has to be careful only to include conclusions of classical arguments.

8.1 Related research

Although the concept of Socratic-style reasoning is ubiquitous in informal argumentation, it has until now received surprisingly little attention from a formal perspective. One exception is the work of Joseph Fulda [Ful00], in which attention is paid to the process of legal cross-examination. Fulda describes cross-examination as "an opportunity to impeach evidence given by the witness during direct examination". The idea is that by asking a carefully selected series of questions, the witness can be led to commit himself to an inconsistency [7], thereby undermining his credibility. Fulda's treatment, how-

[7] The contradiction can either be immediate — in case the witness directly contradicts something he claimed earlier — or indirectly [Ful00, p. 338]: " (...) the testimony can be impeached by contradicting not just a proposition in the testimony space, but by contradicting *any logical consequence of any proposition in the*

ever, is quite brief, and no attention is being paid to the specific features of defeasible reasoning.

8.2 Future research

A possible interesting research topic would be the construction of a dialogue system for Socratic-style reasoning. Such a dialogue system would then make use of the *but-then* statement and the concept of nested commitments. One particular problem that needs to be dealt with, however, is that of the *relevance* of dialogue moves. This problem in fact also occurs in informal dialogue. When, in an Anglo-Saxon legal system, a lawyer cross-examines a witness it is sometimes not immediately clear what the point is that he or she wants to make. In that case, the counterparty can object at which the lawyer may have to explain privately to the judge his question technique as well as the relevance of it. As the problem of relevance plays a role in real-life dialogue and argumentation it should not come as a surprise to encounter the same problem when attempting to formalize this kind of dialogue. Joseph Fulda encounters basically the same problem when trying to determine a criterion that allows one to distinguish between those cross-examinations that are "to the point" and those that are not [Ful00]. Fulda concludes that this is only possible if the future part of the line of questioning is also taken into account. It thus seems that one possible way to deal with this issue is to have a third party to whom such a future line of questions (or a future line of but-then statements) could be entrusted, and who would then determine whether the questioner's dialogue steps can still be considered as relevant. The role of this third party would be comparable to that of a judge in informal cross-examinations. The question of how such should be concretely implemented is an issue still open for future research.

9 Acknowledgements

I would like to thank Leon van der Torre for helping to develop the initial idea, Henry Prakken for his useful comments, Reind van de Riet for his support and the anonymous referees for their many useful suggestions.

testimony space."

References

[BCAC05] T. Bench-Capon, K. Atkinson, and A. Chorley. Persuasion and value in legal argument. *Journal of Logic and Computation*, 15:1075–1097, 2005.

[BDKT97] A. Bondarenko, P.M. Dung, R.A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93:63–101, 1997.

[BH01] Ph. Besnard and A. Hunter. A logic-based theory of deductive arguments. *Artificial Intelligence*, 128 (1-2):203–235, 2001.

[Cam04] M.W.A. Caminada. *For the sake of the Argument. Explorations into argument-based reasoning.* Doctoral dissertation Free University Amsterdam, 2004.

[Cam05] M.W.A. Caminada. Contamination in formal argumentation systems. In *Proceedings of the 17th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC)*, pages 59–65, 2005.

[DBC03] P. E. Dunne and T. J. M. Bench-Capon. Two party immediate response dispute: Properties and efficiency. *Artificial Intelligence*, 149:221–250, 2003.

[Dun95] 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.

[Ful00] J. Fulda. The logic of "improper cross". *Artificial Intelligence and Law*, 8:337–341, 2000.

[GS04] A.J. García and G.R. Simari. Defeasible logic programming: an argumentative approach. *Theory and Practice of Logic Programming*, 4(1):95–138, 2004.

[Mac79] J. D. MacKenzie. Question-begging in non-cumulative systems. *Journal of Philosophical Logic*, 8:117–133, 1979.

[Nel94] L. Nelson. *De Socratische Methode.* Uitgeverij Boom, Amsterdam, 1994.

[Per82] Chaim Perelman. *The Realm of Rhetoric.* University of Notre Dame Press, Notre Dame, Indiana, 1982. translated by William Kluback.

[PlaBC] Plato. Sophist, 360 BC. translated by Benjamin Jowett.

[Pla10] Plato. Lysis. In E. Rhys, editor, *Socratic Discourses by Plato and Xenophon.* J.M. Dent & Sons ltd., London, 1910.

[Pol87] J. L. Pollock. Defeasible reasoning. *Cognitive Science*, 11:481–518, 1987.

[Pol91] J.L. Pollock. Self-defeating arguments. *Minds and Machines*, 1:367–392, 1991.

[Pol92] J. L. Pollock. How to reason defeasibly. *Artificial Intelligence*, 57:1–42, 1992.

[Pol95] J. L. Pollock. *Cognitive Carpentry. A Blueprint for How to Build a Person.* MIT Press, Cambridge, MA, 1995.

[Pra00] H. Prakken. On dialogue systems with speech acts, arguments, and counterarguments. In *Proceedings of the 7th European Workshop on Logic for Artificial Intelligence (JELIA-2000)*, number 1919 in Springer Lecture Notes in AI, pages 224–238, Berlin, 2000. Springer Verlag.

[PS97] H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7:25–75, 1997.

[PV02] H. Prakken and G. A. W. Vreeswijk. Logics for defeasible argumentation. In D. Gabbay and F. Günthner, editors, *Handbook of Philosophical Logic*, volume 4, pages 219–318. Kluwer Academic Publishers, Dordrecht/Boston/London, second edition, 2002.

[Rei80] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[Ski02] J. Skidmore. Skepticism about practical reasoning: transcendental arguments and their limits. *Philosophical Studies*, 109:121–141, 2002.

[SL92] G.R. Simari and R.P. Loui. A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence*, 53:125–157, 1992.

[VP00] G. A. W. Vreeswijk and H. Prakken. Credulous and sceptical argument games for preferred semantics. In *Proceedings of the 7th European Workshop on Logic for Artificial Intelligence (JELIA-00)*, number 1919 in Springer Lecture Notes in AI, pages 239–253, Berlin, 2000. Springer Verlag.

[Vre93] G. A. W. Vreeswijk. *Studies in Defeasible Argumentation.* PhD thesis, Dept. of Mathematics and Computer Science,Vrije Universiteit Amsterdam, 1993.

[Vre97] G. A. W. Vreeswijk. Abstract argumentation systems. *Artificial Intelligence*, 90:225–279, 1997.

[WK95] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning.* SUNY Series in Logic and Language. State University of New York Press, Albany, NY, USA, 1995.