Abstract Argumentation and Explanation Applied to Scientific Debates

Dunja Šešelja (dunja.seselja@ugent.be) and Christian Straßer (christian.strasser@ugent.be)
Ghent University, Centre for Logic and Philosophy of Science

Abstract.
Abstract argumentation has been shown to be a powerful tool within many fields such as artificial intelligence, logic and legal reasoning. In this paper we enhance Dung's well-known abstract argumentation framework with explanatory capabilities. We show that an explanatory argumentation framework (EAF) obtained in this way is a useful tool for the modeling of scientific debates. On the one hand, EAFs allow for the representation of explanatory and justificatory arguments constituting rivaling scientific views. On the other hand, different procedures for selecting arguments, corresponding to different methodological and epistemic requirements of theory evaluation, can be formulated in view of our framework.

Keywords: abstract argumentation, explanation, scientific debates, scientific reasoning, formal modeling

1. Introduction

Formal theories of argumentation have been extensively researched within the fields of artificial intelligence, philosophy, logic and computer science. One of the most influential accounts of argumentation is Dung's abstract argumentation framework (see (Dung, 1993), (Dung, 1995)). The significance of Dung's approach derives from the fact that it abstracts away from the nature of arguments and argumentation rules, which allows the user to focus on the interplay of arguments rather than on their specific structure. More precisely, an argumentation framework (AF) consists of a set of arguments $\mathcal{A}$, which are taken to be abstract entities represented by alphabetical letters, and the binary (so-called *attack*) relation $\rightarrow$ defined over this set. AFs are used to select sets of arguments from $\mathcal{A}$ that satisfy certain standards of acceptability. Selection criteria are defined in order to explicate these standards: for instance selected sets of arguments are supposed to be non-conflicting and to be able to defend themselves against all argumentative attacks.[1] An extensive research on abstract argumentation has shown that such systems are capable of formalizing various approaches to nonmonotonic

[1] We give a formal account of this and other standard selection criteria in Section 3.

reasoning in the fields of artificial intelligence, logic programming and human reasoning. The fruitfulness of Dung's framework stems not only from its abstract character, but also from the fact that it is easily enhanceable with additional properties and useful in different application contexts.[2]

In this paper we will enhance AFs with explanatory features. The aim of this enhancement, which we will call an Explanatory Argumentation Framework (EAF), is, on the one hand, to equip AFs with tools that can model explanatory reasoning, and on the other hand, to demonstrate that abstract argumentation provides a useful formal framework for the modeling of scientific debates. The basic idea of our enhancement is to introduce to AFs a set of explananda and an explanatory relation. This will allow us to express certain notions, such as explanatory power and explanatory depth, in terms of our framework. Moreover, we will show that EAFs allow for a comparison of different sets of arguments in view of their explanatory virtues. Taking into account that scientific explanation is one of the key constituents of scientific reasoning, EAFs will turn out to be a handy modeling tool in fields dealing with the reconstruction and the modeling of scientific debates, such as the philosophy of science. To this end we will offer a set of criteria which are useful for the demarcation of rivaling scientific views in terms of arguments, as well as for an evaluation of such views in terms of their argumentative and explanatory properties. As a result we will be able to formulate new selection criteria, suitable for the modeling of argumentation and explanation in a scientific context. Finally, we will show that our approach may be embedded or linked to the argumentative shift in methodology that is associated with scholars such as Pera (see (Pera, 1994; Pera, 2000)) and Dascal (see (Dascal, 2000)). Since a number of enhancements developed for AFs can also be applied to EAFs, we will suggest that in this way abstract argumentation can provide an even more refined and more realistic modeling of scientific debates.

The paper is structured as follows. We begin in Section 2 by explicating the close relation between argumentation and explanation, on the basis of which we will motivate the significance as well as the structure of our framework. In Section 3 we introduce the basic notions of abstract argumentation. In Section 4 we present EAFs. In Section 5 we informally introduce criteria and selection procedures that allow for a more realistic representation of scientific reasoning than the standard

[2] For the enhancements that have been developed for AFs see Section 7.3. As for the different application fields, for instance, AFs have been used for an improved account of default reasoning (Bondarenko et al., 1997), (Dung and Son, 1996), as well as for multi-agent systems (Coste-Marquis et al., 2007), (Bench-Capon, 2003).

selections offered within Dung's abstract argumentation framework. In Section 6 we formally explicate the explanatory properties that have been previously introduced. Section 7 offers a discussion on some additional questions concerning the virtues of our framework. We show here that EAFs reflect some of the key ideas underlying rhetorically minded approaches to scientific rationality, and we point out the novelties of our framework, as well as possible enhancements of it. Section 8 concludes the paper.

2. Argumentation and Explanation

Explanation and argumentation have been studied in philosophy of science, epistemology and logic. While some authors have discussed the two in close relation, others have pointed out the need to distinguish them as two different processes of reasoning. In this section we will explicate the relation of argumentation and explanation in our framework and situate it within the broader context of the discussion on this matter.

2.1. The Goal-Directed Perspective

One way to look at the problem of distinguishing argumentation and explanation is to explicate what it is that explanations try to achieve. Hughes states that

> the purpose of an explanation is to show *why and how* some phenomenon occurred or some event happened; the purpose of an argument is to show *that* some view or statement is correct or true. Explanations are appropriate when the event in question is taken for granted, and we are seeking *to understand* why it occurred. Arguments are appropriate when we want to show that something is true, usually when there is some possibility of disagreement about its correctness. (Hughes, 1992, p. 76, italics added)

Thus, the goal of an explanation is to reach an understanding of the *why* or *how* something occurred, depending on the type of explanation. The occurrence itself is thereby taken for granted.

The quotation above suggests even more, namely that explanations are distinguished from arguments due to the different types of goals that the respective notions achieve. In contrast to explanations, arguments are justificatory, they show *that* something is the case and not why or how. Thus, the quotation suggests a clear distinction between arguments and explanations. However, we will subscribe in the following to the view that justificatory arguments are a certain subclass of arguments, and that explanations (in a strict sense) should be conceived of as a certain type of arguments as well.

2.2. Explanations as Arguments

The view that explanations are arguments has a long history. According to Hempel's covering law model of explanation, which is considered to be one of the origins of the contemporary study of explanation,[3] an explanation is an argument in which a sentence describing a phenomenon to be explained is derived from the class of those sentences which are adduced to account for this phenomenon, and which contain at least one law of nature (Hempel, 1965, p. 247). A similar view on explanations as arguments or argument patterns can be found in unificationist accounts of explanation (e.g. see (Kitcher, 1981), (Weber, 1999)). Moreover, the view that some arguments have an explanatory function is not foreign to the literature on argumentation either (e.g. see (Pera, 1994) p. 110, (Pera, 2000) p. 57).

In order to see which type of arguments explanations are we should first of all analyze the notion of an argument a bit more. Mayes in his (Mayes, 2000) distinguishes between two meanings of this term: a formal and an evidentiary one. In a broader, *formal sense*, an argument is a finite sequence of propositions (called premises) followed by a proposition (called conclusion), in which the premises are intended (or taken) to entail the conclusion (ibid., p. 363). In a narrower, *evidentiary sense*, we are speaking of a specific type of formal arguments, namely those in which premises provide a rational justification for believing the conclusion (ibid., p. 364). This is the sense in which Hughes uses this term and what we have called justificatory arguments. However, as Mayes points out, beside justificatory arguments there is another type of formal arguments: explanatory ones or simply, explanations. The basic difference between these two types of arguments, as we have already seen, is that while justificatory ones aim at justifying *that* something is the case, explanatory ones aim at answering the question *why* (or *how*) something is the case.[4]

[3] The contemporary study of explanation is usually seen as originating in (Hempel and Oppenheim, 1948), which was further developed in (Hempel, 1965).

[4] It is important to notice that sometimes we can determine whether a given argument is justificatory or explanatory only by taking into account the given context, which reveals the intention of the speaker. For example, an elliptically expressed argument "Shops are closed today because it's a public holiday." – could in one context be an explanation given in reply to the question "Why are shops closed today?", where the fact that shops are closed is taken for granted for both participants involved in the conversation. In some other context though, the same

In this sense, an explanation is a formal argument consisting of an explanans and an explanandum, where the former one offers the causes or the governing law of the latter one and thus provides a better understanding of it. That is, premises of an explanatory argument represent an explanans from which a conclusion, representing an explanandum, can be inferred on the basis of a certain inference relation (such as deduction, induction, etc.).

2.3. The Processual Character of Explanations

Let us in the following put more emphasis on the notion of understanding. By offering an explanation to an explainee, the explainer tries to make the explainee understand why/how/etc. the explanandum occurred. However, nothing guarantees that after offering an explanatory argument, the explainee has actually reached the point of understanding. Often an explanatory argument needs to be complemented by a dialogical process that clarifies certain open questions or doubts on part of the explainee. Thus, we can perceive explanations in a broader sense to be an argumentative process aiming at the explainee's understanding of the given phenomenon. Such a processual character of explanations has been emphasized, for example, by Schurz who speaks of 'explanatory episodes', characterized as relations between two cognitive systems communicating with each other in order to achieve a better understanding of the phenomena in question (Schurz, 1991).

An explanatory episode is considered to be a process which includes not only explanatory arguments but may also include justificatory arguments, where the task of the latter ones is to further substantiate the former. Upon hearing an explanation, the explainee may request further clarification and may express his doubt for some of the arguments by either challenging (some of) them with counter-arguments or by requesting further clarification. Consequently, the explainer may have to justify claims constituting her explanation. Thus, arguing is often a constitutive part of an explanatory process not only because the explainer may wish to explicate and strengthen her claims, but also because the validity of some of them may be brought into question in case the explainee does not find them sufficiently accurate, clarified, understandable, etc. Consequently, explanatory reasoning does not have to result only in knowledge accumulation, but may sometimes also include a revision and thus contraction of the knowledge base of the explainee (see (Schurz, 1991)).

argument could be expressed as a justification of the fact that shops are closed, where this fact is doubted by one of the participants.

Argumentation is thus a constitutive feature of explanatory reasoning. Together with Mayes we can say that, "until an explanatory hypothesis has been independently established through argument, it lacks the power to support anything at all", and the other way around, "until a justified belief has been adequately explained it lacks the power to support anything at all" (Mayes, 2000, p. 375). Let us take a closer look at Mayes' description of such an interactive relation:

> Explanation is a process that is triggered by a certain kind of input, viz., a surprising fact, a salient feature of our environment that we have somehow failed to predict. (E.g., the car wont start). ... Explaining a fact involves the formation of a causal hypothesis (The battery is dead.). This possible cause is the output of the explanatory process. But for any given fact there will always be a number of possible causes. Hence, the process of explanation will be useful as a way of gaining predictive control over our environment only if it is supported by another process whose function is to determine which, if any, of the possible causes should be accepted. (ibid., p. 378).

2.4. Explanation and Argumentation in the Context of Scientific Reasoning

In this paper we will primarily focus on the modeling of scientific explanations, or more precisely, scientific explanatory reasoning. In addition to the dynamics of explanation and argumentation which has to be taken into account in such a modeling, it is important to notice that a bilateral relation, involving one explainer and one explainee, is not the only possible situation in an explanatory process. This is, for instance, the case in scientific contexts where a number of scientists can participate in a discussion on a certain explanatory issue. In such situations, the explanation proposed by one scientist (or a group of scientists) undergoes a critical assessment by the other members of the given scientific community. Moreover, different scientists may offer different, mutually rivaling explanations. As a result, arguments used in explanatory reasoning will be open for criticism in terms of counterarguments, while explanations will be open for a comparison with other alternative explanations.

Thus, on the basis of the points presented in this section we can conclude that an appropriate modeling of scientific explanatory reasoning should allow for the following three properties:[5]

[5] Even though these properties are important for the modeling of scientific explanatory reasoning, they are not restricted to it. Similar kind of requirements may be posed on the modeling of other explanatory contexts such as e.g. expert systems.

1. a dynamic view on explanatory reasoning, involving both justificatory and explanatory arguments;
2. the possibility of expressing criticism in terms of counterarguments and alternative explanations;
3. the possibility of multiple participants in an explanatory process.

In this paper we will offer a framework that can satisfy all three of these requirements. First of all, rooting our framework in Dung's account of abstract argumentation allows for an abstract notion of an argument, which can be seen as corresponding to an argument in a formal sense. Consequently, both justificatory and explanatory arguments can be represented as argumentative letters in general. Second, the dynamics of abstract argumentation, based on the attack relation between arguments, allows for a modeling of counterarguments and alternative explanations. Finally, as we will demonstrate in our examples, an abstract argumentation system allows for the input from multiple parties to be represented in an explanatory process, which further contributes to its fitness for the modeling of scientific explanatory reasoning and scientific debates.

Before we introduce our framework, let us give a summary of the main concepts of Dung's abstract argumentation.

3. Abstract Argumentation

Let us first have a look at the classical definition of argument systems introduced by Dung in (Dung, 1995). We have a set of arguments and an attack relation between them. The abstractness of the framework concerns both elements. On the one hand, we do not reveal the concrete structure of the given arguments, but represent them by abstract letters. On the other hand, we do not reveal the concrete nature of the attack relation.

Definition 1. An *argumentation system (AF)* is a pair $(\mathcal{A}, \rightarrow)$ where $\mathcal{A}$ is a set of arguments, and $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is a relation between arguments. The expression $a \rightarrow b$ is pronounced as "a attacks b" and $\rightarrow$ is called the *attack relation.*

The central notion of AFs is acceptability. We are interested in selecting sets of arguments, let us call them *A-sets*, which satisfy criteria of acceptability.[6] For example, the selected arguments should be at least conflict-free or should be able to defend themselves from all the attacks by other arguments. Applied to scientific discourse, an A-set

[6] These were introduced by Dung as so-called "extensions".

represents a collection of arguments that satisfies a certain, for instance, methodological virtue. The following definitions introduce the standard selection criteria for A-sets.[7]

Definition 2. Given an argumentation framework (AF) $(\mathcal{A}, \rightarrow)$ and an A-set $A \subseteq \mathcal{A}$ we define:

(i) A defends the argument a iff every attacker of a is attacked by a member of A.

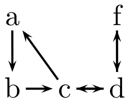
(ii) A is *conflict-free* iff no argument in A attacks an argument in A.

(iii) A is said to be *defended* if it is conflict-free and every argument in A is defended by A.[8]

(iv) We call maximal (w.r.t. $\subseteq$) defended A-sets *preferred A-sets.*

Example 1.

We will demonstrate the concepts just introduced with the attack-diagram to the right. The table lists the A-sets belonging to selections based on the different criteria:



conflict-free	defended	preferred
$\emptyset, \{a\}, \{b\}, \{c\}, \{d\}, \{f\},$ $\{a,d\}, \{a,f\}, \{b,d\}, \{b,f\}, \{c,f\}$	$\emptyset, \{d\}, \{f\}, \{a,d\}$	$\{a,d\}, \{f\}$

4. Enriching Abstract Argumentation with Explanations

In this section we will define explanatory argumentation frameworks (EAFs) and some basic notions that can be expressed by them.

4.1. Explanatory Argumentation Frameworks (EAFs)

In order to equip argumentation frameworks with explanatory capabilities we extend them with the following three elements:

1. A *set of explananda* $\mathcal{X}$: we interpret elements of the set $\mathcal{X}$ as statements describing a state of affairs which is considered to be requiring an explanation by all the parties involved in the given dispute or which is within the explanatory scope of a given discipline. This could be a

[7] Many other selection criteria for A-sets have been proposed in the literature (such as being "stable" and "complete" in (Bondarenko et al., 1997), being "semi-stable" in (Caminada, 2006), being "ideal" in (Dung et al., 2007), etc.). In order to make the technical level of the paper not too involving we stick to the selection criteria introduced in Definition 2. Generalizations of our framework for other selection criteria are straight-forward.

[8] Defended A-sets are also often labeled "admissible".

certain natural or social phenomenon, an experimental result, etc. In accordance with the standard view on explanations which take the explanandum as indisputable in character (in contrast to the conclusions of evidentiary arguments), we assume that the set of explananda consists of facts which are considered to be indisputable in the given field. For example, an explanandum can be a description of, or a reference to a certain observation or an experimental result.[9]

2. The second element we need to introduce is an *explanatory relation* $\dashrightarrow$ which holds between:

(a) an argument and an explanandum, i.e., $\dashrightarrow \subseteq \mathcal{A} \times \mathcal{X}$ where $\mathcal{A}$ is the set of arguments of a given AF and $\mathcal{X}$ is the set of explananda;
(b) between two arguments, i.e., $\dashrightarrow \subseteq \mathcal{A} \times \mathcal{A}$.

Where $a \in \mathcal{A}$ and $x \in \mathcal{A} \cup \mathcal{X}$ we designate "$a \dashrightarrow x$" as "a explains x". While the explanatory relation between an argument and an explanandum links phenomena requiring explanation with the reasons which should allow for their better understanding, the explanatory relation between arguments themselves allows for explanations to be deepened. In other words, argument b can be used to explain one of the premises of argument a (which may itself be used to explain explanandum e) or the link between the premises and the conclusion. The former case corresponds to Thagard's idea of a deepening of scientific explanations or Bermúdez's notion of a vertical explanation.[10] The latter case can occur in explanatory situations typical for everyday language, didactic situations or oral disputes, in which arguments are usually expressed in an elliptic manner so that the link between premises and the conclusion might not be sufficiently clear. In accordance with the abstract character of abstract argumentation frameworks, we treat the explanatory relation in an abstract manner as well.

3. We introduce the third element that simplifies the modeling of scientific debates while making it at the same time more accurate. Sometimes two arguments a and b are based on incompatible presuppositions or premises. It is important to notice that this does not necessarily indicate that a attacks b or vice versa. This is often the case with alternative explanations of a certain phenomenon. For instance, some geologists in the first half of the twentieth century explained the origin of mountains by the idea of continental drift (g), while some other geologists explained it by the thesis of the earth's contraction (c). Although g and c were clearly incompatible (see Example 2, Section 4.2),

[9] Nevertheless, sometimes there are disputes on what is to count as a *valid* or *important* explanandum in a given scientific field. For the possibility of enhancing our framework so that it can allow for such disputes, see Section 7.3.

[10] We will present both notions in Section 6.

g in itself was not sufficient to attack c (and vice versa): just naming an alternative explanation is not considered as a counter-argument in a scientific debate. And indeed, counter-arguments against both ideas were established on independent grounds. For instance, contractionists attacked the theory of drifters by pointing out that it cannot account for a mechanism that would enable continents to plough through the dense seafloor. In order to model such incompatibilities between arguments we introduce another relation, the *incompatibility relation* $\sim$.

In conclusion, we define:

Definition 3. An Explanatory Argumentation Framework (EAF) is a tuple $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$, where $\langle \mathcal{A}, \rightarrow \rangle$ is an AF, $\mathcal{X}$ is a set of *explananda*, $\dashrightarrow \subseteq (\mathcal{A} \times \mathcal{X}) \cup (\mathcal{A} \times \mathcal{A})$, and $\sim \subseteq \mathcal{A} \times \mathcal{A}$ is a symmetric relation.

We call $\dashrightarrow$ the *explanatory relation* and the elements of $\dashrightarrow$ *atomic explanations*. The elements of $\mathcal{X}$ are denoted by $e, e_1, e_2, ...$, and the elements of $\mathcal{A}$ by $a, b, c, d, f, g, ...$ Moreover, $\sim$ is the *incompatibility relation* and in case $a \sim b$, a and b are said to be incompatible.

Concerning the selection criteria introduced in Definition 2 little adjustment is needed. The only change concerns the notion of conflict-freeness: In the remainder of the paper we call an A-set A *conflict-free* iff no argument in A attacks *or is incompatible* with an argument in A. As before, A is said to be *defended* if it is conflict-free and every argument in A is defended by A.

4.2. Basic Definitions

In order to introduce some basic notions it is useful to first define some basic graph-theoretic concepts.

Definition 4. A *directed graph (digraph)* is an ordered pair $G = \langle V, \rightsquigarrow \rangle$ where V is a set and $\rightsquigarrow$ is a binary relation on V, $\rightsquigarrow \subseteq V \times V$. The elements of V are called vertices and the elements of $\rightsquigarrow$ are called arrows. $G' = \langle V', \rightsquigarrow' \rangle$ is a *sub-graph* of G iff $V' \subseteq V$, $\rightsquigarrow' \subseteq V' \times V'$ and $\rightsquigarrow' \subseteq \rightsquigarrow$. G' is a *proper* sub-graph of G iff it is a sub-graph of G and $V' \subset V$ or $\rightsquigarrow' \subset \rightsquigarrow$. We say that there is a *path from x_1 to x_n in* G iff there are $x_1, \ldots, x_n \in V$ for which $(x_i, x_{i+1}) \in \rightsquigarrow$ for all $1 \leq i < n$. G is *circular* iff there is an $x \in V$ for which there is a path from x to x. Where $V' \subseteq V$, we define $\rightsquigarrow_{V'} =_{\text{df}} \{(x, y) \in \rightsquigarrow \mid x, y \in V'\}$.

We now introduce definitions that characterize explanations in EAFs.

Definition 5. Let $\mathsf{A} = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$ be an EAF, $e \in \mathcal{X}$, and $a \in \mathcal{A}$.

(i) We call a sub-graph $X = \langle A, \dashrightarrow_A \rangle$ of $\langle \mathcal{A}, \dashrightarrow_{\mathcal{A}} \rangle$ an *explanation of e* iff there is a unique argument $a \in A$ such that (i) $a \dashrightarrow e$ and (ii) there is a path in X from every $a' \in A \setminus \{a\}$ to a.
We say that the explanation X is circular if X is a circular graph.
We use the following writing conventions: In the case that an explanation X only consists of a path $\langle P, \dashrightarrow_P \rangle$, where $P = \{a_1, \ldots, a_n\}$ and $\dashrightarrow_P = \{(a_{i+1}, a_i) \mid 1 \leq i < n\}$, we abbreviate X by $\langle a_1, \ldots, a_n \rangle$.
We sometimes write $X[e]$ for X in order to indicate that X explains e.

(ii) An explanation $\langle A, \dashrightarrow_A \rangle$ is *conflict-free* iff A is conflict-free.

(iii) An explanation $X[e]$ is *deeper* than an explanation $X'[e]$ iff X' is a proper sub-graph of X. We write $X' \prec X$. We say that X' is a *sub-explanation* of X.

(iv) $X[e]$ and $X'[e]$ are *alternative explanations* of e iff neither $X \prec X'$ nor $X' \prec X$.

(v) An explanation $\langle B, \dashrightarrow_B \rangle$ is *offered by an A-set A* iff $B \subseteq A$. We define the set of all explananda for which A offers an explanation by

$$\epsilon(A) =_{\mathrm{df}} \{e \in \mathcal{X} \mid \text{there is an explanation } X[e] \text{ offered by } A\}.$$

Example 2. We give an example of a scientific debate, on the basis of which we can clarify the notions that have so far been introduced, and which will serve to show why and how our framework can be useful in an evaluation of scientific theories. The following arguments, which correspond to the EAF given in Figure 2, are central to (though they do not exhaust) an important discussion in the geological sciences around the 1920's. This debate marked the beginning of a scientific revolution which was initiated by Alfred Wegener's theory of continental drift (henceforth, the Drift), and resulted in the theory of plate tectonics (see e.g. (Le Grand, 1988)). Wegener started off by suggesting that his theory is superior compared to two already existing theories – contractionism and permanentism.

Scientific debates are often very technical and complex. Hence, in order to follow them a scholar has to be sufficiently familiar with the involved topics. We chose this example since the technical complexity of the given arguments allows for a representation that is understandable and transparent also for scholars that are not already familiar with the subtleties of the research in geology at that time and it is thus ideal as a running example for demonstrating our framework. We begin with an excerpt of the arguments given by drifters and contractionists, which will be further extended at a later point in this paper (see Example 3). The explananda are as follows:[11]

e_1 (*fossils*) Similar kinds of fossils were found on different continents.[12]

e_2 (*orogeny*) There are mountains and mountain chains on continents.

e_3 (*glaciation*) There is an evidence of glaciation which took place in the late Paleozoic in the southern continents (the so-called Southern Glaciation or the late Paleozoic glaciation).

The following arguments were offered:

a (*land bridges*) In the past, the continents were apart like nowadays, but connected by land bridges. This is how different species of flora and fauna were distributed to different parts of the world.

b (*no land bridges nowadays*) The hypothesis of the land bridges is not plausible since it is not clear how such land bridges would have disappeared throughout the history.

c (*contraction*) Vertical displacements of the otherwise unmovable earth's crust result from the contraction of the earth, which causes shrinking and lateral compression in the crust. That is why some rocks (such as mountains) became elevated while some others (such as the land bridges) subsided into the ocean.

d (*cooling*) The earth is contracting due to its cooling.

f (*drift-paleontology*) Continents were once connected into a super-continent, before they drifted away from each other. Different species of flora and fauna were distributed over different continents in this way.

g (*drift-orogeny*) Drifting of continents results in the leading edge of the continent being compressed and folded upwards due to the resistance of the seafloor. Consequently, mountains are being formed along the leading coastlines of a drifting continent, or result from two continents colliding against each other.

h (*drift-glaciation*) The nowadays southern continents were once a part of a super-continent, and positioned more in the north. That

[11] Even though we will, for the sake of simplicity, focus on some arguments exchanged between the drifters and the contractionists, it is important to notice that the permanentist side could easily be included in our example, and that EAFs are suitable for the modeling of any number of parties involved in an explanatory process.

[12] This is a simplified version of the actual explanandum, which states a peculiar distribution of Cambrian trilobites – fossil arthropods that lived 500 to 600 million years ago (see (Gould, 1977)); we will make similar simplifications of other explananda and arguments constituting this example in order to avoid burdening the reader with too many technical details.

is why glaciation could occur on them in Paleozoic, before they drifted to the south.

i (*drift*) The earth consists of concentric shells, the density of which increases from the crust to the core, so that the continents float on and extend into the ocean floors. This is why the continents, pulled by a particular (currently unknown) force, could drift away from their original locations where they once formed a super-continent.

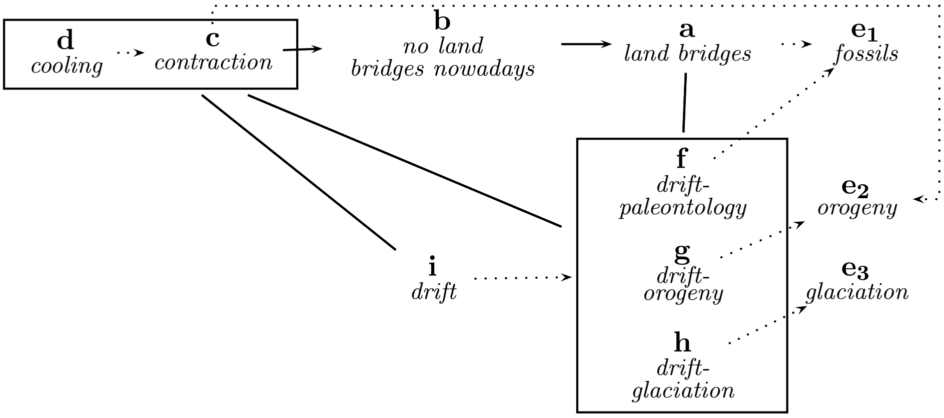


Figure 2. The EAF of Example 2. Solid arrows represent the attack relation, dotted arrows represent the explanatory relation, and solid lines represent the incompatibility relation. Solid lines from the box around arguments f, g and h to argument a and the box around d and c indicate that all three arguments are incompatible with a, d and c. The explanatory arrow from i to the box around f, g and h indicates that each of the three arguments is explained by i.

Let us focus on maximal conflict-free sets of arguments that are able to defend themselves in order to gain a first approximative representation of the rivaling scientific views.[13] Hence, we are interested in *preferred* A-sets. In this example we have two such A-sets: $A_1 = \{f, h, g, i\}$ and $A_2 = \{a, c, d\}$, corresponding to the two represented rivaling views in geology: Drift and contractionism. Next, we have two atomic explanations of e_1: a and f, two atomic explanations of e_2: c and g, and one atomic explanation of e_3: h. Each of them is a sub-explanation of the following explanations, resp.: $X_1[e_1] = \langle a \rangle$, $X_2[e_1] = \langle f, i \rangle$, $X_3[e_2] = \langle c, d \rangle$, $X_4[e_2] = \langle g, i \rangle$, and $X_5[e_3] = \langle h, i \rangle$ all of which are conflict-free and non-circular. By Definition 5iv, $\langle g, i \rangle$ is deeper than $\langle g \rangle$ alone. Explanations X_2, X_4 and X_5 are offered by the drifters, i.e. A_1, and explanations X_1 and X_3 are offered by the contractionists, i.e. A_2. Notice that the two preferred A-sets, A_1 and

[13] We will offer more realistic selection procedures for the representation of scientific views in Section 5.

A_2, offer explanations for different sets of explananda: while A_2 offers explanations for e_1 and e_2, A_1 offers explanations for e_1, e_2 and e_3. Hence, while the two A-sets are in view of their argumentative properties equivalent (i.e. they are both maximally defended and conflict-free), their explanatory power is different. Since a difference in the explanatory power of A-sets can play an important role in the evaluation of scientific theories, we will introduce criteria for selecting A-sets in view of their explanatory features in Sections 5 and 6.

This example also demonstrates the usefulness of our incompatibility relation. We take the contractionists' arguments a, c and d to be incompatible with all of the explanations given in view of the Drift (f,g,h and i) since they assume mutually incompatible explanatory mechanisms.[14] Obviously, argument d refers to the level of cooling that can account for the level of contraction needed to explain the formation of mountains (e_2), and not to a more moderate version of cooling and contracting, which would not be able to explain such a phenomenon and which would be compatible with the Drift. Notice that without our incompatibility relation, the arguments of the two sides would have to be modeled either as formally unrelated in terms of EAFs or as related in terms of bidirectional attacks (in place of the incompatibility relations). However, in the first case, it would be impossible to distinguish between the two rivaling scientific views, since for instance $\{a, c, d, f, g, h, i\}$ would be a conflict-free A-set. The second option would allow for the distinction between the rivaling views, but it would have some other implausible results. For example, argument b would be taken as defended from c by any of the Drift arguments f, g, h, i merely due to the fact that an alternative explanation has been proposed, which would be counterintuitive.

5. Towards a More Realistic Modeling of Scientific Debates

5.1. Criteria for the Modeling of Scientific Debates

As we have seen in Section 4, certain criteria for A-sets (such as being conflict-free or defended) are useful for the representation of opposing views in scientific debates. However, for a more realistic modeling of

[14] For example, one of the reasons for this incompatibility (which for the sake of simplicity we have kept out of our examples) lies in the fact that Drift relied on the principle of isostasy, which implied that continents and ocean floors had to be different either in structure or in composition, and which conflicted with the contractionists' idea of interchangeability of oceans and continents (Oreskes, 1999, p. 21-55).

scientific views and their evaluation, we need to add few more criteria and modify some of those that have already been introduced. We will show that some of the key epistemic values relevant in the evaluation of scientific theories can be expressed in terms of our framework. On the basis of them, we will be able to formulate selection types for A-sets that reflect certain methodological and epistemic preferences scientists or philosophers may have when evaluating theories in view of the available arguments. We will propose two procedures for such selections, which are more apt for this purpose than the standard criteria introduced in Section 3.

It is important to mention though that it is beyond the scope of this paper to finally settle the question, which criteria (and combinations thereof) most adequately capture the methodological and epistemic standards used in theory evaluation, either descriptively or normatively. Many new criteria have been studied since Dung developed his AFs, which refine and optimize the first generation of selection criteria in many ways.[15] It is a task left for future research to clarify which (combinations of) criteria are the most suitable for the modeling of the notions of acceptability underlying theory choice. What we want to present here is rather a general directive for how this research may proceed and in which way notions developed in terms of EAFs can be useful for this task.

The criterion of conflict-freeness, introduced in the previous section, is a minimal requirement that should be satisfied by A-sets representing a given scientific view. In view of this criterion we can then distinguish between mutually rivaling scientific views. Another epistemic standard significant in the evaluation of scientific theories is their explanatory power.

Explanatory power. The explanatory power of an A-set – usually also referred to as explanatory *scope* or explanatory *breadth* – is given by the set of explananda which are explained by its constituting arguments. We are interested in sufficiently explanatory powerful (conflict-free) A-sets. There are two ways of comparing the explanatory power that may be relevant in the assessment of scientific theories. On the one hand, if an A-set has a (clearly) smaller explanatory scope compared to another A-set, then it is usually considered to be a suboptimal candidate in the context of theory acceptance (other things being equal). On the other hand, when we are evaluating whether a new scientific theory is worthy of pursuit, it may be enough for an A-set to have some novel explanations, i.e. to explain certain explananda that are not explained by any alternative A-set. Therefore it will be

[15] See Footnote 7.

useful to introduce two ways of comparing the explanatory power of A-sets: on the one hand in a *quantitative* sense and on the other hand in a *qualitative* sense. We will discuss them in more detail in Section 6.1. Since the aim of this section is to present the main idea underlying our new criteria, we will use only a simplified version of the latter comparison type which we informally define as follows:

We say that an A-set A_1 is explanatory more powerful than an A-set A_2 iff the set of explananda for which the arguments in A_1 offer an explanation is a proper super-set of the set of explananda for which A_2 offers an explanation (i.e., $\epsilon(A_2) \subset \epsilon(A_1)$).

Example 3. In order to get a more accurate picture of the discussion in geological sciences presented in Example 2, we extend it with some additional arguments. The EAF corresponding to the example is given in Figure 3.

j (***mechanism-problem***) It is not at all clear how the continental drift can occur, since continents cannot simply plough through the dense seafloor.

k (***radioactivity***) Due to the discovery of radioactive material in the earth's crust, which produces heat when decaying, we can claim that the earth cannot be cooling, at least not to such an extent that would account for the origin of higher mountain chains.

l (***why contracting?***) It is not plausible to assume that the earth is contracting unless we know the causes of such a process, and no such cause seems to exist.

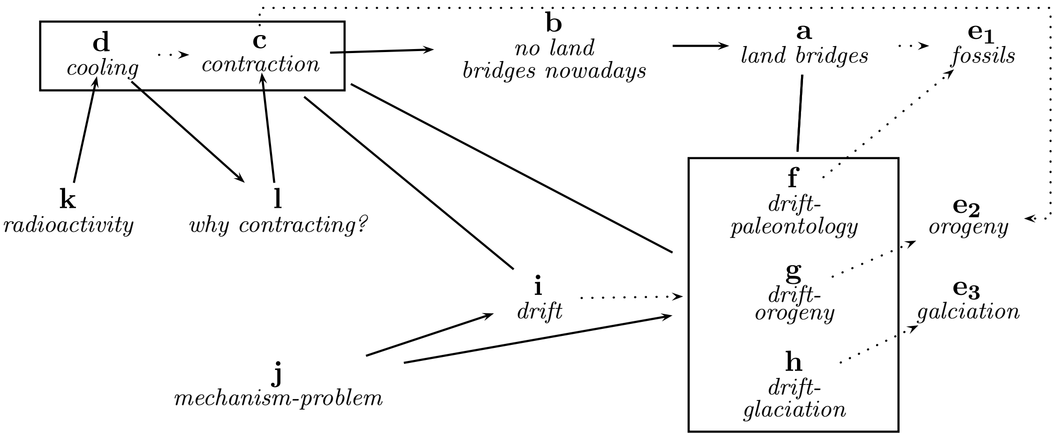


Figure 3. The EAF from Example 3.

The most explanatory conflict-free A-sets are various super-sets of $A_1 = \{f, g, h\}$, e.g. $A_2 = \{f, g, h, i, l\}, A_3 = \{b, f, g, h, k, l\}, A_4 = \{b, f, g, h\}$, etc. Note that any conflict-free super-set of $\{a, c\}$ (i.e. A-sets representing contractionism) only explains $\{e_1, e_2\}$ and is hence explanatory weaker than $A_1, \ldots, A_4$. Indeed, given the arguments introduced so far, the Drift offers a broader explanatory scope than Contractionism.

However, it is important to notice that in this example none of the most explanatory A-sets, such as $A_1, \ldots, A_4$, are defended in a strict sense: after all neither of these sets is able to defend itself from the attack by j. The only preferred A-set is $A'_1 = \{b, j, k, l\}$. Nevertheless, this set has no explanatory power with respect to the given explananda. Note that the two preferred A-sets from Example 2 – $\{f, g, h, i\}$ and $\{a, c, d\}$ – both offer a greater explanatory power than A'_1. However, they are not anymore selected, since they are not able to defend themselves from all the attacks. This situation is not atypical in science since many theories that are accepted or pursued are confronted with objections and criticisms of various kinds, from which they cannot always immediately be defended. This is especially so in the context of pursuit where we are primarily interested in a theory that can offer explanations for certain phenomena in spite of having some open problems. For example, during the confrontation of the above mentioned geological theories, the fact that none of them was resistant against criticism was not a sufficient reason for abandoning their further pursuit. Hence, in certain epistemic contexts we might wish to lower our standard of defense, that requires from an A-set to defend itself from all attacks.

Weakening the Standard of Defense. The idea of this weakening is to say that an A-set A_1 is more defended than another A-set A_2 iff A_2 is attacked by more arguments against which it cannot defend itself than A_1.[16]

Let us take a look at the most explanatory A-sets from Example 3. A_2 is attacked by both j and d. Similarly, A_4 is attacked by j and c. The A-sets A_3, $A_5 = \{f, g, h, i\}$, and $A_6 = \{f, g, h, i, k\}$ fair better: they are only attacked by j. The latter sets belong to the most defended of the most explanatory A-sets.

Explanatory Depth. In addition to comparing sets of arguments in view of their explanatory power, we can also compare them in view of their *explanatory depth.* For instance, both $\{f, g, h\}$ and $\{f, g, h, i\}$ have the same explanatory breadth since both offer explanations for all the shared explananda e_1, e_2 and e_3. However, $\{f, g, h, i\}$ is explanatory deeper than $\{f, g, h\}$ since i explanatory deepens arguments f, g and

[16] This is still, of course, a very rough account of the degree of being defended. It would get more realistic if we took into consideration a weighting of the attacks, since some attacks may be considered as more severe than others (see also our Discussion in Section 7.3).

h. Note that the latter are not shared explananda but theory-internal parts of the Drift. Since we are interested in representing a scientific view as consisting not only of the arguments that directly explain the shared explananda (e.g. $\{f, g, h\}$), but also the arguments that explanatory deepen the former, this criterion could be of use as well. We will give a more precise definition of this notion in Section 6.2.

5.2. Selection Procedures for New Types of A-Sets

On the basis of the newly introduced criteria, we are now able to express selection procedures that correspond to certain types of epistemic evaluation of scientific theories, that is, views of scientists participating in scientific debates.

Procedure 1. The underlying idea of this procedure is to select the argumentative core of the most explanatory scientific views or theories together with arguments that are used for attacking their rivals. It consists of the following steps:

1. Select all the conflict-free A-sets.
2. Out of these, select the most explanatory A-sets.
3. Out of these, select the most defended A-sets.
4. Out of these, select the maximal A-sets (w.r.t. set-inclusion).

Applied to our example, the procedure delivers the A-set $\{b, f, g, h, i, l, k\}$. First, by selecting conflict-free A-sets, we make sure that we distinguish between the rivaling theories. Second, we choose from these the most explanatory powerful ones. Third, by choosing the most defended ones of these, we select the least problematic of the explanatory powerful theories. Finally, by choosing the maximal ones from the latter selection, we make sure we include as many mutually compatible arguments as possible, thus also including those which are used to attack the rivaling theories.

Procedure 2. The idea underlying this procedure is to select the explanatory core of the most explanatory theories.

1. Select all the conflict-free A-sets.
2. Out of these, select the most explanatory A-sets.
3. Out of these, select the most defended A-sets.
4. Out of these, select the explanatory deepest A-sets.
5. Out of these, select the minimal A-sets (w.r.t. set-inclusion).

In the first three steps we proceed analogous to Procedure 1. By selecting the explanatory deepest A-sets we want to make sure we include the explanatory gist of the given theory. Finally, by choosing the minimal of these, we preserve only those arguments belonging to such an explanatory core, while disregarding, for example, the arguments used only for attacking the rivaling theories, i.e. arguments that don't have an explanatory or defensive function. Applied to our example this procedure delivers the A-set $\{f, g, h, i\}$ (as the reader can easily verify).

Even though both of our procedures prioritize the criterion of explanatory power to the criterion of defense, in some contexts we may wish to reverse this order, and even use the full defense criterion. For instance, when evaluating which theories should be accepted (and not only pursued), we may want to allow only for theories that can be fully defended. Such a procedure would begin with the selection of defended A-sets, followed by the selection of the most explanatory ones of these. Obviously, by different combinations of the criteria we may obtain different procedures suitable for different epistemic contexts.

Let us also remark that the procedures offered above are of a sequential or vertical nature: selections with respect to various criteria are applied step-wise. It is also possible to select "horizontally" by making use of weighting functions for A-sets. Let us give a simple example. Given an EAF $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$ and a conflict-free A-set A we define:

$$\pi(A) = \mu_d \frac{|\mathcal{A} \setminus \alpha(A)|}{|\mathcal{A}|} + \mu_e \frac{|\epsilon(A)|}{|\mathcal{X}|},$$

where $\alpha(A)$ is the set of attackers of A against which it cannot defend itself. Moreover, μ_d and μ_e are numerical weights that model the importance we attach to the criteria defendedness and explanatory power respectively. Let us apply π to some A-sets from Example 3 where $\mu_d = \mu_e = 1$: for the Drift we have e.g. $\pi(\{b, f, g, h, i, k, l\}) = \pi(\{i, f, g, h\}) = \frac{10}{11} + \frac{3}{3}$, while for contractionism we have $\pi(\{a, c, d, j\}) = \pi(\{a, c, d\}) = \frac{10}{11} + \frac{2}{3}$. Of course, horizontal and vertical selection mechanisms may be combined. For instance, we could first select A-sets that maximize π and then select the maximal ones out of these. In the example above we would end up with $\{b, f, g, h, i, k, l\}$.

In the following section we will give a more precise formal representation of the comparisons in view of explanatory power and explanatory depth, which will also allow for a refinement of different aspects of these two procedures.

6. A Formal Account of Explanatory Properties

6.1. Explanatory Power

Let us now properly define the two ways of comparing the explanatory power of A-sets that have been introduced in the previous section and point out possible refinements for each of them.

Definition 6. Comparing the explanatory power in a qualitative sense: A is explanatory stronger than A', in signs $A' \sqsubset_e A$, iff the set of explananda for which A offers an explanation is a super-set of the set of explananda for which A' offers an explanation: $\epsilon(A') \subset \epsilon(A)$. This notion was used in Section 5.

Definition 7. Comparing the explanatory power in a quantitative sense: A is explanatory stronger than A', in signs $A' \sqsubset_c A$, iff A explains numerically more explananda than A': $|\epsilon(A')| < |\epsilon(A)|$.[17]

Example 4. Let $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$ be an EAF where $\mathcal{X} = \{e_1, \ldots, e_{10}\}$. Let A and A' be preferred A-sets for which $\epsilon(A) = \{e_1, \ldots e_8\}$ and $\epsilon(A') = \{e_6, \ldots, e_{10}\}$. Note that $A' \sqsubset_c A$ since A explains 8 out of 10 explananda but A' only explains 5. However, $A' \not\sqsubset_e A$ since $\epsilon(A') \not\subset \epsilon(A)$.

Sometimes the comparative measures of explanatory power offered in Definitions 6 and 7 are too strict. To see this suppose that there is a third preferred A-set A'' in our Example 4 for which $\epsilon(A'') = \{e_4, \ldots, e_{10}\}$. Note that $A'' \sqsubset_c A$. However, A numerically explains only one explanandum more than A''. Often if the explanatory power of two theories is not very different this is not a sufficient reason for preferring one over the other.

For a more refined approach to representing both comparative notions of explanatory power, we generalize them by introducing a threshold value τ. For the quantitative notion we define $A \sqsubset_c^{\tau} A'$ iff $|\epsilon(A')| - |\epsilon(A)| > \tau$ where τ is a constant. Note that $\sqsubset_c^0$ is equivalent to $\sqsubset_c$. By introducing a threshold value τ, for instance 1, both A-sets A and A'' are equally explanatory strong with respect to $\sqsubset_c^1$. This introduces an interesting option for the modeling of states in science where different scientific groups are characterized by similarly strong, but nevertheless incompatible explanatory features. It is easy to see that $\sqsubset_e$ can be generalized in a similar way using threshold values.

[17] It is easy to see that $\sqsubset_c$ and $\sqsubset_e$ are strict preorders on $\wp(\mathcal{A}) \times \wp(\mathcal{A})$. (A strict preorder is a irreflexive and transitive binary relation.) Obviously $A \sqsubset_e A'$ implies $A \sqsubset_c A'$.

Example 5. Let us return to our Example 3. Take for instance the two conflict-free A-sets: $A_7 = \{a, c, d\}$ and $A_8 = \{b, f, g, h, i\}$. It is easy to see that A_8 has a greater explanatory power than A_7 since it offers an explanation of e_1, e_2, and e_3, while A_7 only offers an explanation of e_1 and e_2. Formally speaking, $A_7 \sqsubset_e A_8$ as well as $A_7 \sqsubset_c A_8$.

Nevertheless, the explanatory power of the two A-sets in question is similar (although in our simplified modeling $\epsilon(A_7) \subset \epsilon(A_8)$): A_8 explains only one explanandum more than A_7. By introducing a threshold $\tau = 1$, we obtain both extensions – A_7 and A_8 – as maximal elements of $\sqsubset_c^1$ (with respect to all conflict-free A-sets): due to their high and similar explanatory power they are both acceptable according to this notion. Such a rendering would correspond, for instance, to the view that geology in the first half of the twentieth century was in a multi-paradigm state where both contractionism and the Drift (as well as permanentism) were mutually rivaling paradigms (see e.g. (Stewart, 1990, p. 139)), in contrast to the above, strict rendering which corresponds to the preference on the Drift as a more explanatory powerful conception. This is due to the fact that the explanatory power of the two camps was very similar such that, given an appropriate threshold τ, also for a more complete and realistic modeling of this debate, A-sets representing both views would be selected.

6.2. Explanatory Depth

As we have seen in the previous section, we can also compare A-sets with respect to their explanatory depth. This is important, for instance, in cases in which we have two A-sets with the same explanatory power, but one of which offers a deeper explanation of some explananda than the other one. The formal definition of explanatory depth is as follows:

Definition 8. Given two A-sets A_1 and A_2, we say that A_2 is *explanatory at least as deep as* A_1, in signs $A_1 \sqsubseteq_d A_2$ iff for every explanation $X_1[e]$ of $e \in \epsilon(A_1)$ offered by A_1 there is an explanation $X_2[e]$ offered by A_2 such that $X_1 \prec X_2$ or $X_1 = X_2$. We say that A_2 is *explanatory deeper* than A_1, written $A_1 \sqsubset_d A_2$, iff $A_1 \sqsubseteq_d A_2$ but it is not the case that $A_2 \sqsubseteq_d A_1$.

For instance, in Example 3, $\{f, g, h, i\}$ is explanatory deeper than $\{f, g, h\}$.

In addition to their application in the selection procedures mentioned in the previous section, our criteria for explanatory power and explanatory depth may be interesting in capturing some philosophical notions as well. For example, Thagard's concepts of *broadening*

– explaining new facts, and *deepening* – explaining why the theory works (Thagard, 2007, p. 29) correspond to our notions of explanatory power and explanatory depth. Similarly, we can account for the notions of "horizontal" and "vertical explanations" (see (Bermúdez, 2005)) by representing them, respectively, with our notions of primary explanation and its deepening.[18]

7. Discussion

In this section we address questions that are relevant for situating our framework in the broader context of philosophy of science and methodology, Dung's abstract argumentation, as well as in the context of other formal accounts of explanatory reasoning. Therefore we shall clarify the relevance and the novelties of EAFs, as well as some possible enhancements.

7.1. EAFs and the Argumentative Shift in Methodology

Discussions in the field of philosophy of science and scientific methodology in the last couple of decades have witnessed a growing conviction that a rule-based algorithmic approach to theory appraisal is problematic. One possible attempt to preserve the normative idea of rationality in spite of abandoning the idea of a static, universally applicable scientific method can be found in more rhetorically minded approaches to scientific reasoning, such as Pera's (1994)[19] or Dascal's (2000). Instead of an algorithmic assessment of scientific theories, Pera and Dascal emphasize the evaluation in view of the argumentative context underlying the given episode in the history of science. Similarly, Longino points out that a "method must [. . .] be understood as a collection of social, rather than individual, processes, so the issue is the extent to which a scientific community maintains critical dialogue" (Longino, 1990, p. 76). While formal approaches to scientific reasoning have been mainly focused on the logical form of arguments (that is, the nature of the inference

[18] According to Bermúdez, a "horizontal explanation is the explanation of a particular event or state in terms of distinct (and usually temporarily antecedent) events or states." (Bermúdez, 2005, p. 32). For example, a horizontal explanation providing an answer to the question why the window broke when it did, might call upon the baseball's hitting it and a generalization about windows tending to break when hit by a baseball. However, if we ask why the mentioned generalization holds, that is, what features of the physical structure of glass make it fragile in such circumstances – we are asking for an explanation of the grounds of the given horizontal explanation. Such explanations Bermúdez calls vertical explanations (ibid, p. 32-33).

[19] Pera is inspired by Kuhn's notion of persuasion (Kuhn, 1962).

relation), both Pera and Dascal show that scientific debates (Dascal's *controversies*) are typically not resolved by the derivational reasoning that is characteristic for logic but rather by scientists exchanging arguments and trying to convince each other by giving reasons that substantiate their points:

> The contenders pile up arguments they believe increase the weight of their positions vis a vis the adversaries' objections, thereby leading, if not to deciding the matter in question, at least to tilting the 'balance of reason' in their favor. Controversies are neither 'solved' nor 'dissolved'; they are resolved. Their resolution may consist in the acknowledgment (by the contenders or by their community of reference) *that enough weight has been accumulated in favor of one of their positions, or in the emergence (thanks to the controversy) of modified positions acceptable to the contenders, or simply in the mutual clarification of the nature of the differences at stake.* (Dascal, 2000, p. 165, italics added)

Our account of EAFs is supposed to mirror the idea underlying such an argumentative approach to scientific controversies in a formal way. Various possible enhancements (which will be mentioned in subsection 7.3) allow for a framework that reflects such a rhetorically minded approach to scientific debates in a more refined way. However, in contrast to an informal analysis such as Pera's *Dialectics*, which deals with rhetorical aspects of arguments in scientific debates in terms of classifying argument types and explicating their roles, our approach abstracts from the concrete type of arguments by focusing only on their roles as being attacks or explanations. This allows us to inhabit a formal middle-ground for the modeling of the "tilting of the 'balance of reason'" by means of selection procedures defined with the help of our framework. Thus, EAFs (and more generally speaking, abstract argumentation) can be considered to complement the informal theories of argumentation by representing a formal tool that can serve to rationally reconstruct scientific debates from an argumentative point of view.

7.2. The Novelty of EAFs

What are the main novelties of our framework? This question should be answered in view of the research done in abstract argumentation frameworks, as well as in view of other formal accounts of explanatory reasoning.

With regard to the former, it could be argued that since our explanatory arrow is a kind of support relation, systems such as the bipolar one (see (Cayrol and Lagasquie-Schiex, 2005)), which also feature an argumentative support relation, might be sufficient to model the notions introduced by our framework (such as explanations, explanatory power, explanatory depth, etc.). Nevertheless, the presence of the set of explananda $\mathcal{X}$ in EAFs makes an important difference. Placing explananda outside of the set of arguments makes it possible not only to express notions such as the explanatory power of an A-set, but also to represent alternative explanations of the same phenomenon and to compare the explanatory virtues of different A-sets. Moreover, on the basis of such an enhancement we are able to formulate new selection types in view of explanatory properties of the arguments, which are more suitable for the evaluation of scientific views than the standard Dung's selection types. Thus, our explanatory relation cannot be substituted by the already existing support relation.

With regard to other formal accounts of explanatory reasoning, it is important to notice that there are different levels of abstraction on which formal representations can be based. First of all, we can formally analyze explanatory reasoning by focusing on the nature of the inferential relation present in explanations. This will give us, for instance, logical systems of abduction (see e.g. (Aliseda, 2006)). Next, we can obtain formal representations by abstracting away from the logical properties of explanatory reasoning and focusing on the explanatory coherence of the propositions constituting a certain cognitive system. Thagard's account of explanatory coherence (**?**) and its implementation in the computer program ECHO is an example of such an approach.[20] Finally, if we abstract away from the propositional level, we can represent explanatory reasoning in terms of arguments taken in an abstract sense of the term, that is, without analyzing the specific type of inferential relations involved in them. This kind of approach is the one employed in EAFs. An important merit of such an approach is that it allows for a transparent representation of scientific debates, while at the

[20] It is important to notice that even though both Thagard's explanatory coherence and our EAFs aim at modeling the comparison of cognitive systems in terms of their explanatory virtues, there is a number of differences between these two accounts. For example, while the basic unit of Thagard's coherentism is a proposition, our framework is constituted of arguments and explananda; while Thagard's notion of acceptability is defined in terms of explanatory coherence, we speak of different types of acceptability, defined in terms of defensibility and certain explanatory properties; while Thagard does not model the dynamics of argumentation and thus cannot model the idea of an argumentative defense, we can; while the distinction of the evaluation of theories in the context of pursuit and in the context of acceptance is not explicated in his account, we have shown that such a distinction can be made in EAFs; finally, our graphical representation is quite different from Thagard's and may be considered as more transparent when it comes to the representation of scientific debates.

same time offering handy tools for evaluating A-sets in view of their argumentative properties (e.g. being conflict-free and defended) and explanatory virtues. Moreover, by abstracting away from the specific nature of inferential relations in argumentative reasoning, we are able to model debates in which not every argument is necessarily based on a valid inference (and which may lead to it being attacked by a counterargument). Hence, by applying such an approach to scientific reasoning, we have not only introduced a novelty in the field of abstract argumentation, but also in the field of formal representations of scientific reasoning, which has, to our knowledge, so far not been linked with argumentation in an abstract sense of the term.[21]

7.3. Enhancing EAFs

In order to allow for an even more realistic modeling of scientific debates, our framework can be enhanced in different ways. For example, we can easily introduce a support relation (presented in (Cayrol and Lagasquie-Schiex, 2005)) which runs over the set of arguments.[22] Another interesting enhancement would be to introduce a weighting on arguments (by means of values (Bench-Capon, 2002), (Bench-Capon, 2003), preferences (Amgoud and Cayrol, 1998), or audiences (Bench-Capon et al., 2007)), evidential support (Oren and Norman, 2008), or to introduce different types of joint attacks (Nielsen and Parsons, 2006) or fuzziness (Janssen et al., 2008) that allows, for instance, for the relaxation of the standards of being conflict-free and of being defended. Introducing nested attacks in a hierarchical manner into EAFs (see (Modgil, 2006; Modgil, 2009)) would allow for a formal distinction and an analysis of the interplay between arguments given on the object level and methodological arguments addressing the former ones. Moreover, our explanatory relation could be refined by formally distinguishing between two types of deepening explicated in Section 4.1. Similarly to the above mentioned nested attacks, we could allow for a nested explanatory relation: $a \dashrightarrow b$ would in that case indicate that a explains one of the premises of b, while $a \dashrightarrow (b \dashrightarrow c)$ would indicate that a explains the explanatory link between b and c.

[21] Even though in this paper we have been primarily concerned with scientific explanations, philosophy of science is not the only novel domain in which abstract argumentation in terms of EAFs may be fruitfully applied. Another interesting application field is expert systems. For example, Moulin et al. (Moulin et al., 2002) argue that in order to justify and "convince their users of the validity of their recommendations ..., artificial agents should also be equipped with explanation and argumentation capabilities." (p. 172).

[22] After all, there are types of argumentative supports that are not explanatory in nature and can hence not be represented by means of our explanatory relation.

Another possible enhancement is related to the property of our framework that explananda are not supposed to be a matter of dispute. Even though this is often so, there are rare cases in which not all involved parties agree about what is to count as a *significant* or even *valid* explanandum in the given field. For example, some geologists criticized Wegener for pointing out that his theory of continental drift explained the jigsaw-fit of continental coastlines, since such a match was, according to them, not at all obvious and did not represent a significant explanandum for geological sciences. In order to model such disputes, EAFs could be enhanced by allowing preferences or values on explananda, so that different opinions of scientists can be represented in that way. However, if we want to model cases in which explananda can be rejected as invalid on an argumentative basis, that is, if we want to allow for a dispute on explananda we could enhance EAFs by allowing attack relation to run not only over arguments, but also to go from arguments to explananda (that is, $\rightarrow \subseteq \mathcal{A} \times (\mathcal{A} \cup \mathcal{X})$). As a result, explananda could be both criticized and defended, while the modeling of explanatory virtues would have to be adjusted in order to account for the fact that different sets of explananda are acceptable with respect to different sets of arguments.

As we have mentioned in the introduction, an important virtue of abstract argumentation is that a basic framework can easily be enhanced in various ways. EAFs as presented in this paper clearly provide an idealized modeling with regard to the subtleties that occur in real scientific debates. By listing possible enhancements of EAFs we have suggested in which way a more realistic modeling could be obtained.

8. Conclusion

In this paper we have presented the Explanatory Argumentation Framework (EAF), obtained by enhancing Dung's Abstract Argumentation Framework with explanatory capabilities. We have motivated such an enhancement by pointing out, on the one hand, the close relation between argumentation and explanation, and on the other hand, the usefulness of our framework in the modeling of scientific debates. We have demonstrated that EAFs allow for a dynamic view on explanatory reasoning by involving both justificatory and explanatory arguments. The relation to scientific debates has been explicated by showing how multiple different scientific views can be modeled by making use of different criteria for selecting arguments. In this way EAFs are able (i) to model the criticism inherent to scientific debates in terms of counter-arguments, (ii) to model alternative competing explanations, and (iii) to evaluate and compare the explanatory features offered by the competing scientific views. Moreover, different selection procedures that can model different epistemic and methodological preferences regarding theory choice can be formulated in our framework.

References

Aliseda, A.: 2006, *Abductive Reasoning.* Springer.

Amgoud, L. and C. Cayrol: 1998, 'On the Acceptability of Arguments in Preference-based Argumentation.'. In: G. F. Cooper and S. Moral (eds.): *UAI.* pp. 1–7, Morgan Kaufmann.

Bench-Capon, T. J. M.: 2002, 'Value Based Argumentation Frameworks'. *CoRR* **cs.AI/0207059**. informal publication.

Bench-Capon, T. J. M.: 2003, 'Persuasion in Practical Argument Using Value-based Argumentation Frameworks'. *Journal of Logic and Computation* **13**, 429–448.

Bench-Capon, T. J. M., S. Doutre, and P. E. Dunne: 2007, 'Audiences in argumentation frameworks.'. *Artificial Intelligence* **171**(1), 42–71.

Bermúdez, J. L.: 2005, *Philosophy of Psychology: A Contemporary Introduction.* Routledge.

Bondarenko, A., P. M. Dung, R. A. Kowalski, and F. Toni: 1997, 'An Abstract, Argumentation-Theoretic Approach to Default Reasoning'. *Artificial Intelligence* **93**, 63–101.

Caminada, M.: 2006, 'Semi-Stable Semantics'. In: *Computational Models of Argument.* pp. 121–132, IOS Press.

Cayrol, C. and M.-C. Lagasquie-Schiex: 2005, 'On the Acceptability of Arguments in Bipolar Argumentation Frameworks'. In: L. Godo (ed.): *ECSQARU*, Vol. 3571 of *Lecture Notes in Computer Science.* pp. 378–389, Springer.

Coste-Marquis, S., C. Devred, S. Konieczny, M.-C. Lagasquie-Schiex, and P. Marquis: 2007, 'On the merging of Dung's argumentation systems'. *Artificial Intelligence* **171**, 730–753.

Dascal, M.: 2000, 'Epistemology and Controversies'. In: T. Y. Cao (ed.): *Philosophy of Science: Volume 10 of Proceedings of the Twentieth World Congress of Philosophy.* Philosophers Index Inc., pp. 159–192.

Dung, P. M.: 1993, 'On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning and Logic Programming.'. In: *International Joint Conference on Artificial Intelligence, Proceedings.* pp. 852–859.

Dung, P. M.: 1995, 'On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games'. *Artificial Intelligence* **77**, 321–358.

Dung, P. M., P. Mancarella, and F. Toni: 2007, 'Computing ideal sceptical argumentation'. *Artificial Intelligence* **171**, 642–674.

Dung, P. M. and T. C. Son: 1996, 'An argumentation-theoretic Approach to Reasoning with Specificity'. In: *Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning (KR'96).* Cambridge, Massachusetts, Morgan Kaufmann Publischers, Inc.

Gould, S. J.: 1977, *Ever Since Darwin*, Chapt. The validation of continental drift, pp. 160–167. Harvard University.

Hempel, C.: 1965, *Aspects of Scientific Explanation and other Essays in the Philosophy of Science.* Free Press.

Hempel, C. and P. Oppenheim: 1948, 'Studies in the Logic of Explanation'. *Philosophy of Science* **15**(2).

Hughes, W.: 1992, *Critical Thinking*. Broadview Press, Petersborough.

Janssen, J., M. D. Cock, and D. Vermeir: 2008, 'Fuzzy Argumentation Frameworks'. In: *Proceedings of IPMU 2008 (12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems)*. pp. 513–520.

Kitcher, P.: 1981, 'Explanatory Unification'. *Philosophy of Science* **48**, 507–531.

Kuhn, T.: 1962, *Structure of Scientific Revolutions*. The University of Chicago Press.

Le Grand, H. E.: 1988, *Drifting Continents and Shifting Theories*. Cambridge University Press.

Longino, H. E.: 1990, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, New Jersey: Princeton University Press.

Mayes, G. R.: 2000, 'Resisting Explanation'. *Argumentation* **14**, 361–380.

Modgil, S.: 2006, 'Hierarchical Argumentation'. In: M. Fisher, W. van der Hoek, B. Konev, and A. Lisitsa (eds.): *JELIA*, Vol. 4160 of *Lecture Notes in Computer Science*. pp. 319–332, Springer.

Modgil, S.: 2009, 'Reasoning about preference in argumentation frameworks'. *Artificial Intelligence* **173**(9-10), 901–934.

Moulin, B., H. Irandoust, M. Bélanger, and G. Desbordes: 2002, 'Explanation and Argumentation Capabilities:Towards the Creation of More Persuasive Agents'. *Artificial Intelligence Review* **17**(3), 169–222.

Nielsen, S. H. and S. Parsons: 2006, 'A Generalization of Dung's Abstract Framework for Argumentation: Arguing with Sets of Attacking Arguments'. In: N. Maudet, S. Parsons, and I. Rahwan (eds.): *Argumentation in Multi-Agent Systems*, Vol. 4766 of *Lecture Notes in Computer Science*. pp. 54–73, Springer.

Oren, N. and T. J. Norman: 2008, 'Semantics for Evidence-Based Argumentation'. In: *Proceeding of the 2008 conference on Computational Models of Argument*. Amsterdam, The Netherlands, The Netherlands, pp. 276–284, IOS Press.

Oreskes, N.: 1999, *The Rejection of Continental Drift: Theory and Method in American Earth Science*. New York, Oxford: Oxford University Press.

Pera, M.: 1994, *The Discourses of Science*. Chicago, London: The University of Chicago Press.

Pera, M.: 2000, 'Rhetoric and Scientific Controversies'. In: M. P. Peter Machamer and A. Baltas (eds.): *Scientific Controversies: Philosophical and Historical Perspectives*. New York, Oxford: Oxford University Press, pp. 50–66.

Schurz, G.: 1991, 'Erklärungsmodelle in der Wissenschaftstheorie und in der Künstlichen Intelligenz'. In: H. Stoyan (ed.): *Proceedings of Erklärung im Gespräch – Erklärung im Mensch-Maschine-Dialog*. pp. 1–42, Springer.

Stewart, J. A.: 1990, *Drifting Continents & Colliding Paradigms: Perspectives on the Geoscience Revolution*. Indiana University Press, Bloomington.

Thagard, P.: 2007, 'Coherence, truth, and the development of scientific knowledge'. *Philosophy of Science* **74**(1), 28–47.

Weber, E.: 1999, 'Unification: What is it, how do we reach and why do we want it?'. *Synthese* **118**(3), 479–499.