A study on the logic of abduction

Marta Cialdea Mayer[1] and **Fiora Pirri**[2]

Abstract.
In this paper we present a logic for abduction, namely a language, a semantics and a proof theory where the abductive relation can be expressed. The abductive relation is the relation between a preferred sentence explaining a given observation, in the context of a background theory, and such an observation itself. The proposed logical system, based on a modal approach, captures propositional abduction over a finite language.

1 INTRODUCTION

Abduction is a natural form of reasoning performed by an agent when new evidence informs her knowledge and it consists in inferring the truth of a sentence, from the fact that it would explain the new evidence. In general there may be different sentences explaining the same evidence, so alternative ones have to be rejected before the inference is performed. In other terms the inference is drawn when the agent is convinced that any other sentence is not plausible enough, compared with the chosen one, to be accepted as an explanation.

In the classical view [5, 6, 12, 14, 15], the notion of abductive explanation is characterized – modulo some generalization – as follows. Let K be a given knowledge base (also called background knowledge), $\models$ a consequence relation, E a statement expressing the new evidence to be explained and such that $K \not\models \neg E$. Assume moreover that a preference criterium is given, interpreting the requirement for an explanation to be considered "better" than another. Then a statement H is taken as a good explanation of E in K iff:

1. $K \cup \{H\} \models E$
2. H is "liked better" than any other statement in the set $\{H' \mid K \cup \{H'\} \models E\}$, according to the given preference criterium.

Usually, for H to be meaningful, the consistency of $K \cup \{H\}$ is also required; however consistency can be implicitly taken into account by the preference criteria: an explanation consistent with the background knowledge is just preferred to one not meeting this condition.

This paper offers an insight into the logical structure of *classical abduction*, as far as the characterization of a suitable notion of preference is concerned. In fact, it shows that the above notion can be expressed at the logical level by modeling, through nested kripke structures, both suitable preference relations on formulae and the choice of a best explanation according to the established preference criteria.

The basic intuition is the following. Assume that $<$ interprets a preference criterium, that is, if A and B are propositional sentences, then A is preferred to B iff $A < B$. Assume the background knowledge is finite, so that we can identify K and the conjunction of formulae occurring in it and write $H \models_{PC} K \rightarrow E$ for $K \cup \{H\} \models_{PC} E$. Here, $\models_{PC}$ is propositional logical consequence. We define a language, its semantics and proof theory, where the abductive relation can be expressed by a formula $H \rightsquigarrow (K \rightarrow E)$, that is logically implied by (thus derivable from) a set of formulae suitably characterizing the chosen preference criterium.

In order to express at the logical level that H is preferred, among all the sentences entailing E in K and that are comparable with H, we build a structure $\mathcal{M}$ in which the following conditions are met:

a The "points" k of $\mathcal{F}$ are structures that can be associated with propositional formulae, i.e. for each k there is a propositional formula A_k that is uniquely associated with k (up to logical equivalence). This association between a structure and a propositional formula is achieved by exploiting the notion of *only knowing* introduced at different levels and independently by Humberstone [11] and Levesque[13].

b There exists a modal formula $\alpha_{K\rightarrow E}$, representing $K \rightarrow E$, such that, for all points k:

$$k \models \alpha_{K\rightarrow E} \text{ iff } \models_{PC} A_k \rightarrow (K \rightarrow E)$$

Therefore, the fact that A_k is a possible explanation of $K \rightarrow E$ can be expressed in the logic.

c The preference relation $<$ is captured by the accessibility relation $\prec$ among the k's, i.e.

$$k \prec k' \text{ iff } A_k < A_{k'}$$

Therefore, A_k is an explanation of E in K iff:

1. $k \models \alpha_{K\rightarrow E}$;
2. for every point k' in $\mathcal{M}$, such that $k' \prec k$, $k' \not\models \alpha_{K\rightarrow E}$, i.e. any other sentence associated with a preferred point k' is not an explanation of $K \rightarrow E$.

The above logical structure is effectively captured by the multimodal system $AK4$ which is a combination of the classical modal system $K4$ with the system CKH^*, a modal logic defined on the basis of well known systems such as Levesque's [12], Humberstone's [10, 11] and Gargov, Passy and Tichev's one [8]. Our logic is a simplification of the above cited ones, in that in the modal language capturing propositional abduction no nesting of modal operators is needed. Furthermore, the basic propositional language is assumed to be finite.

The paper is organized as follows. In the following section the basic structures constituting the points of the main structure $\mathcal{M}$ are introduced. In Section 3 the modal logic CKH^* is presented and its soundness and strong completeness are established. In Section 4 the

[1] Dipartimento di Informatica, Terza Università di Roma, via della Vasca Navale 84, 00146 Roma, Italia. e-mail: cialdea@inf.uniroma3.it
[2] Dipartimento di Informatica, Università di Roma "La Sapienza", via Salaria 113, 00198 Roma, Italia. e-mail: pirri@dis.uniroma1.it

ECAI 96. *12th European Conference on Artificial Intelligence*
Edited by W. Wahlster
Published in 1996 by John Wiley & Sons, Ltd.

complete structure, i.e. the preference structure, is introduced with its axiomatization; soundness and strong completeness are established as well. In Section 5 the abductive relation is defined and in Section 6 two examples of axiomatization of a preference relation are given together with some samples of derivations in the logic. Finally, we discuss the relation of our approach with some others, relevant in the literature on abduction. For more details and complete proofs of the results established in this paper, see [4].

2 THE "POINTS" OF THE PREFERENCE STRUCTURES

In this section we introduce the basic structures on which the logical framework for abduction is built upon.

The modal language $\mathcal{L}$ is built by use of a finite set of propositional letters $\mathbf{P} = \{p_1, \ldots, p_n\}$, the propositional connectives and two modal operators, $\Box$ and $\overleftarrow{\Box}$. The well-formed formulae of the language, called *k-formulae*, are defined as follows:

(i) every propositional letter is a k-formula;
(ii) if A is a propostional formula, then $\Box A$ and $\overleftarrow{\Box} A$ are k-formulae;
(iii) if α and β are k-formulae, then $\neg\alpha$ and $\alpha \wedge \beta$ are k-formulae.

We use lowercase greek letters for modal formulae and reserve latin letters for propositional formulae.

The dual operators $\Diamond$ and $\overleftarrow{\Diamond}$ are defined by $\Diamond A = \neg\Box\neg A$ and $\overleftarrow{\Diamond} A = \neg\overleftarrow{\Box}\neg A$. Disjunction and implication are defined as usual. Furthermore, the following abbreviations will be used:

$$\overleftrightarrow{\Box} A \equiv_{def} \Box A \wedge \overleftarrow{\Box} A$$

$$\overleftrightarrow{\Diamond} A \equiv_{def} \Diamond A \vee \overleftarrow{\Diamond} A$$

$$\mathbf{O}A \equiv_{def} \Box A \wedge \overleftarrow{\Box}\neg A$$

$\overleftrightarrow{\Box}$ is the universal modality, $\mathbf{O}$ the operator for "only knowing".

In the next section, we refer to particular formulae of the propositional calculus called *names*: a name C is a conjunction of literals and such that for every $p \in \mathbf{P}$ exactly one between p and $\neg p$ occurs in C.

The basic modal structures for $\mathcal{L}$ are called *k-structures*; they constitute the elementary points of the construction. A k-structure k is a tuple $\langle w_0, W, R, V\rangle$ where W is a set, $w_0 \in W$ is a distinguished element (the "actual world"), $R \subseteq W^2$, $V : W \mapsto \mathbf{S}$ maps worlds into propositional interpretations of the language $\mathbf{P}$. V satisfies the following conditions:

i. $\{V(w) \mid w \in W\} = \mathbf{S}$ (*totality*);
ii. for all $w, w' \in W$, if $\langle w_0, w\rangle \in R$ and $\langle w_0, w'\rangle \notin R$ then $V(w) \neq V(w')$ (*canonicity*).

Since there is no risk of confusion, the notation $w \in k$ is sometimes used instead of $w \in W$.

A propositional letter p is true in a world $w \in k$ ($k \models_w p$) just in case $V(w)(p) = true$ and the truth of a k-formula α in w, is defined by adding the following clauses to the usual inductive definition:

- $k \models_w \Box A$ iff for all $w' \in k$ such that $\langle w, w'\rangle \in R$, $k \models_{w'} A$
- $k \models_w \overleftarrow{\Box} A$ iff for all $w' \in k$ such that $\langle w, w'\rangle \notin R$, $k \models_{w'} A$.

Truth in a k-structure is identified with truth in the actual world w_0:

$$k \models \alpha \text{ iff } w_0 \models \alpha$$

The notation $\Gamma \models_* \alpha$ is used to say that α is true in any k-structure satisfying every formula of Γ.

The following properties hold:

1. $\mathbf{O}A$ is always satisfiable;
2. for every k-structure k, there exists A such that $k \models \mathbf{O}A$;
3. if $k \models \mathbf{O}A$ and $k \models \mathbf{O}B$, then $\models_{PC} A \equiv B$;
4. for every k-structure k, $k \models \overleftrightarrow{\Box} A$ iff $\models_{PC} A$;
5. if $k \models \mathbf{O}A$ then: $k \models \Box B$ iff $\models_{PC} A \to B$;
   i.e. $\models \mathbf{O}A \to \Box B$ iff $\models_{PC} A \to B$.

3 THE SYSTEM CKH^*

The modal system CKH^* is obtained by addition of the following axioms and inference rules to PC:

K	$\Box(A \to B) \to (\Box A \to \Box B)$	
K'	$\overleftarrow{\Box}(A \to B) \to (\overleftarrow{\Box} A \to \overleftarrow{\Box} B)$	
H	$\Box A \wedge \overleftarrow{\Box} B \to A \vee B$	
L	$\overleftrightarrow{\Diamond} C$	for any name C
FIN	$\Box\neg C \vee \overleftarrow{\Box}\neg C$	for any name C
Nec	$\dfrac{A}{\Box A}$	
Nec'	$\dfrac{A}{\overleftarrow{\Box} A}$	

Axiom schema H (from Humberstone [10]) ensures that k-structures are complementary; axiom schema L (Levesque's axiom restricted to names [13]) ensures *totality*; finally, axiom schema (FIN), introduced on a different setting in [8], ensures *canonicity* of k-structures. The notation $\Gamma \vdash_{CKH^*} \alpha$ is used to say that α is provable in CKH^* from Γ.

Theorem 1 *The system CKH^* is sound and complete, i.e. for any set of k-formulae $\Gamma \cup \{\alpha\}$, $\Gamma \vdash_{CKH^*} \alpha$ iff $\Gamma \models_* \alpha$.*

We sketch here the problems met when considering an infinite language $\mathbf{P}$. Clearly, in that case, condition (ii) in the definition of k-structures is not enough to ensure property 2 above. Moreover, the notion of name makes no sense. The set of k-structures enjoying property 2 is however captured by the system *KH**, that contains all the axioms and rules of *CKH** but (FIN), and where (L) applies to every satisfiable propostional formula. KH^* is in fact *weakly* complete with respect to k-structures satisfying 2. But strong completeness does not hold.

4 PREFERENCE STRUCTURES

A preference structure is essentially a modal **K4** structure, where the "points" are k-structures:

Definition 1 *A* preference structure $\mathcal{M}$ *(P-structure for short) is a structure $\langle k_0, K, \succ, \varphi\rangle$, where K is a set, $k_0 \in K$, $\succ$ is a transitive relation on K and φ is a function that maps every element of K to a k-structure.*

We sometimes write $k \in \mathcal{M}$ instead of $k \in K$. Moreover, we shall identify the elements k with the corresponding $\varphi(k)$, treating in fact K as a multiset of k-structures. Clearly, every $k \in K$ is equipped with its own accessibility relation R_k, so that a P-structure is in fact a multimodal structure. However, since each R_k does not interact with the relations in the other k-structures, this distinction can be ignored.

The language that is interpreted in a P-structure is $\mathcal{L}_{[\star]}$, i.e. $\mathcal{L}$ augmented with the modal operator $[\star]$. Formulae of $\mathcal{L}_{[\star]}$ are called P-formulae and inductively defined as follows:

(i) if α is a k-formula, then α is a P-formula.
(ii) if α and β are P-formulae, then $\neg\alpha$, $\alpha\wedge\beta$, and $[\star]\alpha$ are P-formulae.

Note that a $[\star]$ operator is not allowed to occur inside the scope of a $\Box$ or $\overleftarrow{\Box}$. In fact, k-formulae are the "atoms" over which P-formulae are built in the standard modal way. The dual operator $\langle\star\rangle\alpha$ is defined as $\neg[\star]\neg\alpha$.

The truth of a formula in $k \in \mathcal{M}$ is defined as follows:

1. if α is a k-formula then $\mathcal{M} \models_k \alpha$ iff $k \models \alpha$
2. $\mathcal{M} \models_k \neg\alpha$ iff $\mathcal{M} \not\models_k \alpha$
3. $\mathcal{M} \models_k \alpha \wedge \beta$ iff $\mathcal{M} \models_k \alpha$ and $\mathcal{M} \models_k \beta$
4. $\mathcal{M} \models_k [\star]\alpha$ iff for all k', if $k \succ k'$, $\mathcal{M} \models_{k'} \alpha$

Truth in $\mathcal{M}$ is defined by:

$$\mathcal{M} \models \alpha \text{ iff } \mathcal{M} \models_{k_0} \alpha$$

The notation $\Gamma \models \alpha$ is used to say that α is true in any k-structure satisfying every formula of Γ.

The system **AK4** is obtained by adding the following axioms and inference rule to the system **CKH***:

$$(K_P) \qquad [\star](\alpha \to \beta) \to ([\star]\alpha \to [\star]\beta)$$

$$(4) \qquad [\star]\alpha \to [\star][\star]\alpha$$

$$(Nec_P) \qquad \frac{\alpha}{[\star]\alpha}$$

As the modal operator $[\star]$ does not interact with $\Box$ and $\overleftarrow{\Box}$, **AK4** contains no axioms ruling the connection between the two levels.

Theorem 2 **AK4** *is sound and complete with respect to P-structures, i.e. for any set of P-formulae* $\Gamma \cup \{\alpha\}$, $\Gamma \models \alpha$ *iff* $\Gamma \vdash_{AK4} \alpha$.

5 THE ABDUCTIVE RELATION

In this section, we assume that a preference relation $<$ on formulae is given. We first introduce the notion of P-models of $<$ and then we show how the abductive relation, based on $<$, can be expressed by means of a P-formula. If the abductive relation holds between two PC-formulae A and B, the corresponding formula is true in every P-model of $<$, and vice-versa. This implies that, if Ω is a set of formulae axiomatizing $<$, i.e. such that for any P-structure $\mathcal{M}$, $\mathcal{M}$ is a model of Ω iff $\mathcal{M}$ is a model of $<$, then the abductive relation holds between two PC-formulae A and B iff the P-formula expressing this relation is derivable from Ω in **AK4**. This is why strong completeness of the system is needed. In the next section examples of some suitable Ω's will be given.

The next definition introduces the basic semantical relation between preference structures and ordering relations on propositional formulae.

Definition 2 *Let $<$ be an ordering relation on PC-formulae. A preference structure $\mathcal{M}$ is a P-model of $<$ iff for all propositional formulae A and B, if $\mathcal{M} \models \langle\star\rangle(\mathbf{O}A \wedge \langle\star\rangle\mathbf{O}B)$, then $B < A$.*

In order for this definition to be meaningful, the existence of P-models of preference relations has to be ensured. This can be done only for orderings that satisfy the following property.

Definition 3 *An ordering relation $<$ on PC-formulae is* compatible with logical equivalence *iff for all PC-formulae A, B, C, D if $\models_{PC} A \equiv B$ and $\models_{PC} C \equiv D$, then $A < C$ iff $B < D$.*

Theorem 3 *For any ordering relation $<$ on PC-formulae, if $<$ is compatible with logical equivalence, then there is a P-model $\mathcal{M}$ of $<$.*

We can finally turn to the charaterization of the abductive relation. Let $K \cup \{E\}$ be a set of PC-formulae with K the background knowledge and E the new evidence to be explained. We assume that K is a finite set of formulae, so, without loss of generality, we take K to be a single formula and we simply speak of explanations of formulae $K \to E$, meaning that an explanation of E, in K, is a PC-formula A such that $\models_{PC} A \to (K \to E)$. In the following any formula C such that $\models_{PC} C \to (K \to E)$ is called an *admissible* explanation. Assuming that $<$ is fixed, the notion of preferred explanation is stated as follows.

Definition 4 *If A and B are PC-formulae, then A pref-expl B iff $\models_{PC} A \to B$ and there is no $C < A$ such that $\models_{PC} C \to B$.*

The P-formula that we claim expresses the abductive relation is:

$$A \leadsto B \equiv_{def} [\star](\mathbf{O}A \to (\Box B \wedge [\star]\neg\Box B))$$

Intuitively, the formula states that:

1. $[\star](\mathbf{O}A \to \Box B)$, that is, by property 4 of k-structures, that $\models_{PC} A \to B$; therefore, if $B = K \to E$, A is an admissible explanation;
2. $[\star](\mathbf{O}A \to [\star]\neg\Box B)$: for every k-structure k that may correspond to a formula C (i.e. $\mathcal{M} \models_k \mathbf{O}C$) that is better than A (i.e. $C < A$), $\mathcal{M} \models_k \mathbf{O}C \wedge \neg\Box B$ (i.e. $\not\models_{PC} C \to B$); therefore, C is not an admissible explanation.

The intuition is confirmed by the theorem that follows.

Theorem 4 *If A and B are PC-formulae, then A pref-expl B iff for every P-model $\mathcal{M}$ of $<$, $\mathcal{M} \models A \leadsto B$.*

By means of the semantics and proof theory of the logic, structural properties of the abductive relation can be proved, that are independent from any particular preference relation. For example, the fact that $A \leadsto B, A \leadsto C \vdash_{AK4} A \leadsto (B \wedge C)$ shows that the abductive implication enjoys the "and-right" property:

1.	$\mathbf{O}A \to (\Box B \wedge [\star]\neg\Box B)$	Assumption
2.	$\mathbf{O}A \to (\Box C \wedge [\star]\neg\Box C)$	Assumption
3.	$\mathbf{O}A \to (\Box B \wedge \Box C)$	From 1,2, by PC
4.	$(\Box B \wedge \Box C) \to \Box(B \wedge C)$	Theorem of CKH^*
5.	$\mathbf{O}A \to \Box(B \wedge C)$	From 3,4, by PC
6.	$\neg\Box B \to \neg\Box(B \wedge C)$	Theorem of CKH^*
7.	$[\star](\neg\Box B \to \neg\Box(B \wedge C))$	From 6, by Nec_P
8.	$[\star](\neg\Box B) \to [\star](\neg\Box(B \wedge C))$	From 7
9.	$\mathbf{O}A \to [\star]\neg\Box(B \wedge C))$	From 1 and 8, by PC
10.	$\mathbf{O}A \to (\Box(B \wedge C) \wedge [\star]\neg\Box(B \wedge C))$	From 5 and 9, by PC

Conversely, the fact that some properties do not hold in general can be proved semantically, by exhibiting a countermodel. For example, monotonicity and right weakening do not hold: in fact, $A \rightsquigarrow B \not\vdash_{AK4} A \wedge C \rightsquigarrow B$ and $A \rightsquigarrow B \not\vdash_{AK4} A \rightsquigarrow B \vee C$.

However, the study of the structural properties of abductive inference is, contrary to the case of nonmonotonic reasoning, not very interesting. In fact, being the abductive relation weaker than classical deductive inference, it enjoys very few properties. In contrast, what is of interest in the case of abduction is the possibility of expressing properties of the underlying preference relation that allow one to *introduce* $\rightsquigarrow$, i.e. to infer a sentence of the form $A \rightsquigarrow B$ from premisses that do not have the same form. The next section faces this issue.

6 AXIOMATIZING PREFERENCE RELATIONS

Let us assume that Ω is a set of P-formulae axiomatizing a given preference relation $<$ among PC-formulae, i.e. for any P-structure $\mathcal{M}$, $\mathcal{M}$ is a P-model of Ω iff $\mathcal{M}$ is a model of $<$. Hence, by Theorem 4, if A and B are PC-formulae, then A pref-expl B iff $\Omega \vdash_{AK4} A \rightsquigarrow B$. In this section we give two examples of Ω's axiomatizing commonly used preference relations.

Let us first consider the following preference relation:

$$A < B \quad \text{iff} \quad \models_{PC} B \to A \text{ and } \not\models_{PC} A \to B$$

and let Ω be the set of all the instances of the following schemata:

(IRR) $[\star](\mathbf{O}A \to [\star]\neg\mathbf{O}A)$
(IMP) $[\star](\overleftarrow{\Box}A \to [\star]\overleftarrow{\Box}A)$

The following result establishes that Ω is in fact an axiomatization of $<$.

Theorem 5 *For every P-structure $\mathcal{M}$, $\mathcal{M} \models \Omega$ iff $\mathcal{M}$ is a P-model of $<$.*

Abduction based on the preference relation $<$ amounts to a quite trivial task: the abductive relation is in fact the equivalence relation, i.e. A pref-expl B iff $\models_{PC} A \equiv B$. For example, if $B = (p \to q) \to (p \wedge q)$, B is logically equivalent to p, so p is the only abductive explanation of B. As a sample derivation, we cannot thus show but a derivation in **AK4** of $A \rightsquigarrow A$ from Ω. Some trivial passages are omitted.

1	$[\star](\mathbf{O}A \to \overleftarrow{\Box}\neg A)$	Definition of $\mathbf{O}A$
2	$[\star](\overleftarrow{\Box}\neg A \to [\star]\overleftarrow{\Box}\neg A)$	IMP
3	$[\star](\mathbf{O}A \to [\star]\overleftarrow{\Box}\neg A)$	from 1 and 2
4	$[\star](\mathbf{O}A \to [\star]\neg\mathbf{O}A)$	IRR
5	$[\star](\mathbf{O}A \to [\star](\neg\Box A \vee \overleftarrow{\Diamond}A))$	from 4
6	$[\star](\mathbf{O}A \to [\star]\neg\Box A)$	from 3 and 5
7	$[\star](\mathbf{O}A \to \Box A)$	Definition of $\mathbf{O}A$
8	$[\star](\mathbf{O}A \to (\Box A \wedge [\star]\neg\Box A))$	from 6 and 7

In order to rule out those trivial explanations for a given abductive problem that are equivalent to the problem itself, syntactical restriction are usually considered. A common one is that an abductive explanation has to be (equivalent to) a conjunction of literals. Let us call *C-formulae* those PC-formulae that are equivalent to conjunctions of literals.

Correspondingly, let us consider the following restriction of $<$:

$$A \ll B \quad \text{iff} \quad A < B \text{ and for no formula } C \text{ it is the case that } \models_{PC} A \equiv B \vee C$$

Thus, in particular, if $A \ll B$ and B is a C-formula, then A is a C-formula, too.

Let Ω^* consist of all the formulae in Ω plus the instances of

(CF) $[\star]\neg(\mathbf{O}A \wedge \langle\star\rangle\mathbf{O}(A \vee B))$

Ω^* is an axiomatization of $\ll$:

Theorem 6 *For every P-structure $\mathcal{M}$, $\mathcal{M} \models \Omega^*$ iff $\mathcal{M}$ is a P-model of $\ll$.*

The main difference between abduction based on $<$ and $\ll$ is that, when $\ll$ is considered, $A \vee B$ is not better than A, so that A may be an abductive explanation of $A \vee B$. The following sample derivation shows that $\Omega^* \vdash_{AK4} p \rightsquigarrow (p \vee q)$. Here, **P**=$\{p, q\}$.

1	$[\star](\mathbf{O}p \to \overleftarrow{\Box}\neg p)$	Definition of $\mathbf{O}p$
2	$[\star](\mathbf{O}p \to \Box p)$	Definition of $\mathbf{O}p$
3	$[\star](\mathbf{O}p \to \Box(p \vee q))$	from 2
4	$[\star]((\overleftarrow{\Box}\neg p \wedge \overleftarrow{\Diamond}q) \to \overleftarrow{\Diamond}(\neg p \wedge q))$	by K' and Nec'
5	$[\star](\Box\neg(\neg p \wedge q) \vee \overleftarrow{\Box}\neg(\neg p \wedge q))$	FIN
6	$[\star](\overleftarrow{\Diamond}(\neg p \wedge q) \to \Box\neg(\neg p \wedge q))$	from 5
7	$[\star]((\overleftarrow{\Box}\neg p \wedge \overleftarrow{\Diamond}q) \to \Box\neg(\neg p \wedge q))$	from 4 and 6
8	$[\star]((\overleftarrow{\Box}\neg p \wedge \overleftarrow{\Diamond}q \wedge \Box(p \vee q)) \to$ $\to (\Box(p \vee \neg q) \wedge \Box(p \vee q)))$	from 7
9	$[\star]((\overleftarrow{\Box}\neg p \wedge \overleftarrow{\Diamond}q \wedge \Box(p \vee q)) \to \Box p)$	from 8
10	$[\star]((\overleftarrow{\Box}\neg p \wedge \overleftarrow{\Diamond}q \wedge \Box(p \vee q)) \to \mathbf{O}p)$	from 9
11	$[\star](\mathbf{O}p \to [\star]\neg\mathbf{O}p)$	IRR
12	$[\star](\overleftarrow{\Box}\neg q \wedge \overleftarrow{\Box}\neg p \to \overleftarrow{\Box}\neg(p \vee q))$	by K' and Nec'
13	$[\star](\overleftarrow{\Box}\neg q \wedge \overleftarrow{\Box}\neg p \wedge \Box(p \vee q)$ $\to \mathbf{O}(p \vee q))$	from 12
14	$[\star]\neg(\mathbf{O}p \wedge \langle\star\rangle\mathbf{O}(p \vee q))$	CF
15	$[\star](\mathbf{O}p \to [\star]\neg\mathbf{O}(p \vee q))$	from 14
16	$[\star](\mathbf{O}p \to$ $[\star](\overleftarrow{\Box}\neg p \wedge \neg\mathbf{O}p \wedge \neg\mathbf{O}(p \vee q)))$	from 1, 11 and 15
17	$[\star](\mathbf{O}p \to [\star](\neg(\overleftarrow{\Diamond}q \wedge \Box(p \vee q))\wedge$ $\neg(\overleftarrow{\Box}\neg q \wedge \Box(p \vee q))))$	from 10, 13 and 16
18	$[\star](\mathbf{O}p \to [\star]\neg\Box(p \vee q))$	from 17 by PC
19	$[\star](\mathbf{O}p \to (\Box(p \vee q)\wedge$ $[\star]\neg\Box(p \vee q)))$	from 3 and 18

7 DISCUSSION

In the Artificial Intelligence tradition, abductive reasoning has mostly been seen as the generative problem of computing *explanantes* and the research has been focused mainly on the study of proof-theoretical methods for the generation of possible *explanantes*, that are subsequently compared and filtered to obtain the "good" ones (see [5, 6]). The problem of precisely defining the underlying notion of abductive relation has been faced only recently and existing formulations all rely on meta-logical notions.

The only logical approach to abduction, to our knowledge, is by Boutilier and Becher [1] who, along the lines of [9] and [7], give a formal logical account of explanation in terms of belief revision. The notion of explanation is captured through a conditional connective $A \Rightarrow E$, and abduction is dealt with as a revision process of the state of belief of an agent. This approach accounts for the authors' intention to formalize a notion where it is not required that an *explanans* deductively entail the *explanandum*, nor is it assumed that the observation is consistent with the background knowledge (belief set). On the contrary, abductive explanations are those sentences that "normally" entail the observation – i.e. the underlying logic is nonmonotonic – and preferred *explanantes* are those that require the least change in the belief set in order to be accepted. Therefore:

- the notion of preference between sentences relies on a notion of preferred models in nonmonotonic logic;
- the notion of preference is *context dependent*, i.e. it depends on the background theory.

In other words, "commonsense explanantes" are considered preferable without taking into account any other plausible criteria an agent may want to adopt, so, in particular, if the observation is consistent with the background knowledge, it is a good explanation of itself.

Much research in abductive reasoning relies on a different notion of abduction (see for instance [12, 14, 15]). These works make some assumptions that simplify the generative task: it is assumed that the new evidence E to be explained is consistent with the background theory K and that any *explanans* of E deductively entails E in K. The preference relation between formulae and the criteria for the best choice of the *explanans* are given a logical formulation that is independent of the background theory K. Often, external (meta-logical) conditions are imposed on *explanantes*, with the aim at selecting "good reasons" for an *explanandum* E among all the formulae that logically imply E in the background theory K. Preference criteria have been stated in many different ways, from absolutely general to very empirical ones. In [12] preference is defined as "being simpler", where a (propositional) formula A is simpler than C whenever the set of literals occurring in A is a subset of those occurring in C. A similar notion is used in [14, 15]. A straightforward generalization to logics different from propositional logic is obtained by establishing that A is preferred to C whenever $C \models A$: weaker *explanantes* are preferred[2, 3].

This work approaches the notion of preference in abductive reasoning in a way that is close to the above sketched ones: the notion introduced in Section 1 and that we called "classical" is not committed to any particular preference relation and, therefore, subsumes most notions of abductive relation proposed in the AI literature. A logic for such notion of abductive relation has been defined, whose semantics relies on the models of a preference relation $<$. The formulations of abduction that are based upon the classical approach can be embedded in this semantic framework, by means of a suitable characterization of the underlying preference relation.

The fact that this logic turns out to be quite cumbersome should not surprise, due to the strong interplay of different concepts in the meta-level definition of abduction, such as logical validity and minimality in a set of formulae.

The study is still preliminary and leaves a number of open problems, such as the extensions to an infinite language and/or infinite background theory, the study of automated deduction methods for the logic, the axiomatization of other interesting preference relations, the extension to different logics, such as first order logic and non monotonic logics.

ACKNOWLEDGEMENTS

This work has been partially supported by the Esprit BRA project MEDLAR II, the project MURST 40% "Rappresentazione della conoscenza e meccanismi di ragionamento" and the project TK932 of Agenzia Spaziale Italiana.

REFERENCES

[1] C. Boutillier and V. Becher, 'Abduction as belief revision', *Artificial Intelligence Journal*, **77**, 43–94, (1995).

[2] M. Cialdea Mayer and F. Pirri, 'First order abduction via tableau and sequent calculi', *Bulletin of the IGPL*, **1**, 99–117, (1993).

[3] M. Cialdea Mayer and F. Pirri, 'Modal propositional abduction', *Bulletin of the IGPL*, **6**(3), 99–117, (1995).

[4] M. Cialdea Mayer and F. Pirri, 'A study on the relation between *explanans* and *explanandum*: a modal approach', Technical report, Dipartimento di Informatica e Sistemistica. Università di Roma "La Sapienza", (1995). Available by ftp at ftp.dis.uniroma1.it as pub/cialpir/tr95.ps.gz.

[5] P. T. Cox and T. Pietrzykowski, 'Causes for events: their computation and applications', in *Proceedings of the Eighth Conference on Automated Deduction (CADE-86)*, pp. 608–621, (1986).

[6] R. Demolombe and L. Fariñas del Cerro, 'An inference rule for hypothesis generation', in *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)*, pp. 152–157, (1991).

[7] P. Gärdenfors, 'An epistemic analysis of Explanations and Causal beliefs', *Topoi*, **9**, 109–124, (1990).

[8] Gargov, Passy, and Tinchev, 'Modal enviroment for boolean speculations', in *Proceedings of the 1986 Goedel Conference in Bulgaria*, pp. 253–263, (1986).

[9] C.G. Hempel, 'Deductive-nomological vs. statistical explanation', *Minnesota Studies in the Philosophy of Science*, **3**, 98–169, (1962).

[10] I.L. Humberstone, 'Inaccessible worlds', *Notre Dame Journal of Formal Logic*, **24**, 346–352, (1983).

[11] I.L. Humberstone, 'The modal logic of 'All and Only'', *Notre Dame Journal of Formal Logic*, **28**, 177–188, (1987).

[12] H. J. Levesque, 'A knowledge-level account of abduction', in *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, pp. 1061–1067, (1989).

[13] H. J. Levesque, 'All I know: a study in autoepistemic logic', *Artificial Intelligence Journal*, **42**, 381–386, (1990).

[14] D. Poole, 'Representing knowledge for logic-based diagnosis', in *Proceedings of the International Conference on Fifth Generation Computer Systems*, pp. 1282–1289, (1988).

[15] R. Reiter and J. de Kleer, 'Foundations of assumption-based truth maintenance systems: Preliminary report', in *Proceedings of the National Conference on Artificial Intelligence*, pp. 183–188, (1987).