Rationality Postulates for Induction

Peter A. Flach†

Abstract. I study the process of inductive hypothesis formation from two perspectives: finding general rules that explain given specific evidence, and finding general rules that are confirmed by the evidence. Both forms of hypothesis formation are axiomatised on the metalevel of consequence relations, which provides us with a set of rationality postulates for various forms of induction.

1. Introduction

In this paper I develop a rationality theory of inductive reasoning, applying the methodology of Kraus, Lehmann & Magidor (1990) to views on philosophy of science developed by Peirce and Hempel. This introductory section sketches the motivation for this paper, defines the main concepts and notation, and provides an overview.

1.1 Motivation

Consider a reasoning agent ***R*** that, upon observing several black crows, conjectures that all crows are black. Such an agent is said to reason *inductively*: from specific observations about members of a class it tentatively draws general conclusions about the whole class. Clearly, not any general conclusion will do: for instance, ***R*** couldn't conclude that all crows are white. Thus, we may draw a distinction between rational and irrational behaviour of ***R***. This paper studies limits of rational behaviour of inductive reasoning agents.

In the ideal case, rationality simply prescribes what conclusions are allowed from what premisses, as in the case of formal deductive reasoning, where rationality is defined by a fixed notion of (deductive) validity. In the general case, however, rationality draws a distinction between reasoning behaviours that is more complex than distinguishing between valid and invalid arguments. For instance, inductive rationality as studied in this paper does not prescribe the number of black crows one must have observed before adopting the hypothesis 'all crows are black', but it does prescribe that once one adopts this hypothesis, observation of yet another black crow cannot be a reason to reject it. This establishes a rationality principle of the form 'if a reasoning agent accepts this inductive argument, it should also accept that inductive argument'. Most of the rationality principles we will consider in this paper are of this form. Other possible forms we will consider are 'if a reasoning agent accepts this inductive

†Author's address: INFOLAB, Tilburg University, PObox 90153, 5000 LE Tilburg, the Netherlands, tel. +31 13 466 3119, fax +31 13 466 3069, email Peter.Flach@kub.nl.
Part of this work was supported by Esprit Long Term Research Project 20237 (Inductive Logic Programming), and by PECO Pan-European Scientific Network ILPnet (no. CIPA3510OCT920044).

P.A. Flach (1996), 'Rationality postulates for induction'. In *Proc. Theoretical Aspects of Rationality and Knowledge*, Yoav Shoham (ed.), pp. 267-281, Morgan Kaufmann.

argument, it should not accept that argument' and 'if a reasoning agent does not accept this argument, then it should accept that argument'.

Related to the distinction between rationality and validity is the distinction between the processes of hypothesis formation and hypothesis selection. *Hypothesis formation* is the process that determines, given certain evidence, the set of possible hypotheses, i.e. conjectures that are not ruled out by the evidence. *Hypothesis selection* is the process of selecting, among the possible hypotheses, one or more that meet given criteria. For instance, the reasoning agent may select only those conclusions that are sufficiently justified by the premisses, typically by assessing the truth of the conclusion given the truth of the premisses. Such an assessment procedure, which can be said to generalise the two-valued function of deductive validity to a continuously-valued function of inductive validity or confirmation, has for instance been developed by Carnap (1953). However, in this paper I will concentrate on the purely logical process of hypothesis formation. That is, an inductive hypothesis is not necessarily the conclusion adopted by the inductive agent on the basis of the evidence but rather a possible conclusion.

1.2 Preliminaries

In formalising the logic of inductive hypothesis formation I assume a propositional language L, built up from a set of proposition symbols and the usual logical connectives, a set M of propositional models, and a satisfaction relation $\models \subseteq M \times L$ that is well-behaved with respect to the logical connectives (for all practical purposes M can be thought of as a set of truth assignments to propositional variables). $\models$ will be assumed compact. As usual, we write $m \models \alpha$ if $m \models \alpha$ for all $m \in M$. Note that M may be a proper subset of all possible propositional models, representing the set of models of an implicit background theory; thus, $\models \alpha$ can be interpreted as 'the background theory allows to conclude α'.

An *inductive consequence relation* $\mathrel{|\!\sim} \subseteq L \times L$ is a set of pairs of formulae; if $\alpha \mathrel{|\!\sim} \beta$ we say that β is a possible inductive hypothesis given evidence α. Inductive consequence relations model the behaviour of inductive agents — at this stage we do not fix a particular definition of $\mathrel{|\!\sim}$, but study rationality postulates limiting possible definitions. For instance, given background knowledge including

$$\texttt{crows\_are\_black} \rightarrow \texttt{chevy\_is\_black}$$

and the fact that the reasoner accepts the inductive argument

$$\texttt{chad\_is\_black} \mathrel{|\!\sim} \texttt{crows\_are\_black}$$

our rationality postulates should prescribe that the hypothesis 'all crows are black' can still be maintained after observation of Chevy being black, by stipulating that the inductive argument

$$\texttt{chad\_is\_black} \wedge \texttt{chevy\_is\_black} \mathrel{|\!\sim} \texttt{crows\_are\_black}$$

should likewise be accepted by t[illegible] inductive agent. This observation gives rise to the following postulate:

(1.1) If $\alpha \mid\!\sim \beta$ and [illegible]$\rightarrow\gamma$, then $\alpha\wedge\gamma \mid\!\sim \beta$.

This postulate expresses a principle of *verification*: if hypothesis β is tentatively concluded on the basis of evidence α, and prediction γ drawn from β is subsequently observed, then this counts as a verification of β. In this paper I will investigate a range of such rationality postulates, as well as some of the ways in which they interact.

1.3 Overview of the paper

As intuition constitutes the primary source of justification for such rationality postulates, we need to proceed with care — after all, intuitions sometimes turn out to be incompatible, leading to paradoxical situations. One such a paradox was noted by the philosopher Carl G. Hempel, when he investigated the logic of confirmation (Hempel, 1945). As I have demonstrated elsewhere this particular paradox can be dissolved by making a clear distinction between the statements 'this evidence *confirms* that hypothesis' and 'this evidence *is explained by* that hypothesis' (Flach, 1995). In general, the search for confirmed hypotheses may lead to different conjectures than the search for explanatory hypotheses, an observation that is backed up by recent experimental and theoretical work in inductive logic programming (De Raedt & Bruynooghe, 1993). In this paper I will therefore consider two different forms of induction, which I have termed *explanatory induction* (section 2.1) and *confirmatory induction* (section 2.2).

In addition to studying single rationality postulates, I will also study some of the ways in which they interact. Ultimately, this is done by constructing a semantics, and establishing the equivalence of [illegible]set of postulates and the semantics by means of a representation theorem. In sectio[illegible]3 two such representation theorems will be given, one for each form of induction[illegible]The significance, limitations and prospects of the approach will be discussed in sect[illegible]n 4.

2. Rationality postulates f[illegible] induction

We start with a number of gene[illegible]The first two postulates express the principles of verifica[illegible]n and falsification, well known from philosophy of science:

(I1) If $\alpha \mid\!\sim \beta$ and [illegible]$\alpha\wedge\beta\rightarrow\gamma$, then $\alpha\wedge\gamma \mid\!\sim \beta$.

(I2) If $\alpha \mid\!\sim \beta$ and [illegible]$\alpha\wedge\beta\rightarrow\gamma$, then $\alpha\wedge\neg\gamma \not\mid\!\sim \beta$.

(I1) has been considered above in a slightly less general form. The difference with (1.1) is that predictions may be deduced from a hypothesis together with the evidence on which the hypothesis is based — one might say that, in general, the 'epistemic outcome' of accepting an inductive argument $\alpha \mid\!\sim \beta$ is the belief that $\alpha\wedge\beta$ rather than

just β. This is relevant in cases [redacted] which hypotheses are not required to include all the information presented by the evidence. Of course, such a prediction may also be added as a further hypothesis:

(I3) If [redacted] then $\alpha \mathrel{|\!\sim} \beta \wedge \gamma$.

The postulate of [redacted]sification (I2) can be equivalently stated as follows (the proof requires the postulate of left logical equivalence stated below as (I7)):

(I2′) If [redacted] then $\alpha \not\mathrel{|\!\sim} \beta$.

Principle (I2′), obvious as it may be, has a few technical consequences. For instance, we have reflexivity only for consistent formulae. Clearly, in the light of (I2′) a formula is consistent if it occurs in an arbitrary inductive argument, so we have the following two weaker versions of reflexivity:

(I4) If $\alpha \mathrel{|\!\sim} \beta$, then [redacted] $\mathrel{|\!\sim} \alpha$.

(I5) If $\alpha \mathrel{|\!\sim} \beta$, then [redacted] $\mathrel{|\!\sim} \beta$.

Given an inductive consequence [redacted] α *admissible* iff $\alpha \mathrel{|\!\sim} \alpha$. Any admissible formula is consi[redacted] of postulates to be considered later the converse of this [redacted]tement is also true. Hence, whenever a postulate requires consistency of a formula th[redacted]ition of the form $\alpha \mathrel{|\!\sim} \alpha$.

The next two postulates e[redacted] depends on the logical contents of premiss and conclusio[redacted] rather than on their syntactic formulation:

(I6) If $\alpha \mathrel{|\!\sim} \beta$ and [redacted] then $\alpha \mathrel{|\!\sim} \gamma$.

(I7) If $\alpha \mathrel{|\!\sim} \gamma$ and [redacted] then $\beta \mathrel{|\!\sim} \gamma$.

One may argue that, in some cases, the syntactic form of a hypothesis is relevant: for instance, in inductive logic programming the hypotheses take the form of logic programs, and clearly an efficient or at least terminating program is desirable over an inefficient or non-terminating one. However, we deem such considerations to be extra-logical. Against (I7) one may raise the objection that many concept learning algorithms perform differently when the examples are re-ordered. However, such behaviour is peculiar to a particular algorithm: an algorithm that would be order-independent would not behave irrationally.

Clearly, postulate (I1–7) are too weak to distinguish induction from other forms of reasoning. In order to obtain stronger sets of postulates we need to consult our intuitions regarding explanatory and confirmed hypotheses, respectively.

2.1 Rationality postulates for explanatory induction

Throughout this section $\alpha \mathrel{|\!\sim} \beta$ is interpreted as 'β is a possible explanation of α', which is loosely taken to mean that α can be derived from hypothesis β by means of some *explanation mechanism*. For instance, in inductive concept learning β would be a concept definition, and α would represent the classification of an instance according to

that concept definitio[redacted] The[redacted]xplanation mechanism is assumed to be at least as powerful as deduction, [redacted] assu[redacted]ption that can be formulated as follo[redacted]:

(2.1) If [redacted]$\rightarrow\alpha$ [redacted]d $\gamma \mid\sim \gamma$, then $\alpha \mid\sim \gamma$.

As exp[redacted]ined above, the se[redacted]nd condition in (2.1) prevents the [redacted]onsideration of inconsi[redacted]ent hypotheses.

I[redacted] stronger postulate, which requi[redacted] *monotonicity* of the exp[redacted]nation mechanism:

(E1) If $\alpha \mid\sim \beta$, [redacted]$\rightarrow\beta$ and $\gamma \mid\sim \gamma$, then $\alpha \mid\sim \gamma$.

This p[redacted]stulate states that a possible explanatory hypothesis β [redacted]an be logically strengt[redacted]ned as long as it remains admissible (i.e. consistent). By p[redacted]ting $\beta=\alpha$ in (E1) we obt[redacted]n (2.1), if in addition we assume that $\alpha \mid\sim \alpha$ follows from [redacted]$\rightarrow\alpha$ and $\gamma \mid\sim \gamma$. Since [redacted]$\rightarrow\alpha$ implies $\neg\alpha \not\mid\sim \gamma$ by (I2′), this assumption is assured by the following postulate:

(E2) If $\gamma \mid\sim \gamma$ and $\neg\alpha \not\mid\sim \gamma$, then $\alpha \mid\sim \alpha$.

This postulate represents a third weakening (besides (I4) and (I5) discussed above) of reflexivity. The underlying intuition is best explained by considering its contrapositive: if $\alpha \not\mid\sim \alpha$, i.e. α is too strong with respect to the background theory, then its negation $\neg\alpha$ is so weak that it is explained by arbitrary admissible γ.

The next postulate expresses a principle well-known from algorithmic concept learning: if α represents the classification of an instance and β its description, then we may either induce a concept definition from examples of the form $\beta\rightarrow\alpha$, or we may add β to the background theory a[redacted] induce from α alone. Since in our framework background knowledge is include[redacted]implicitly, β is added to the hypothesis instead.

(E3) If $\alpha \mid\sim \beta\wedge\gamma$, t[redacted]n $\beta\rightarrow\alpha \mid\sim \gamma$.

The converse implication will be [redacted]system.

The last two postulates [redacted] explanatory induction establish a principle of compositionality, namely $\alpha\wedge\beta \mid\sim$ [redacted]ff $\alpha \mid\sim \gamma$ and $\beta \mid\sim \gamma$. The first postulate is simply the statement of one half of this pri[redacted]s a slightly more general form:

(E4) If $\alpha \mid\sim \gamma$ and [redacted]$\mid\sim \gamma$, then $\alpha\wedge\beta \mid\sim \gamma$.

(E5) If $\alpha \mid\sim \gamma$ and [redacted]$\rightarrow\beta$, then $\beta \mid\sim \gamma$.

Both postulates are of considerable importance for computational induction, since they allow for an incremental approach. (E4) states that pieces of evidence can be dealt with in isolation. Another way to say the same thing is that the set of evidence explained by a given hypothesis is conjunctively closed. By the consistency principle (I2′) this set is consistent, which yields the following principle:

(2.2) If $\alpha \mid\sim \beta$, then $\neg\alpha \not\mid\sim \beta$.

It is easy to show that, conversely, (2.2) implies (I2′) in the presence of (2.1) and (I5).

(E5) states a monotonicity property of induction, which can again best be understood by considering its contrapositive: a hypothesis that is rejected on the basis of evidence β cannot become feasible if stronger evidence α is available. In other words: the process of rejecting a hypothesis is not defeasible (i.e. based on assumptions), but based on the evidence only. This is the analogue of the monotonicity property of deduction (which is implication in the second condition). We may further note that the verification postulate (I1) and the monotonicity postulate (E5) can be jointly expressed as

(E5′) If $\alpha \mid\!\sim \beta$ and $\alpha\wedge\beta\rightarrow\gamma$, then $\gamma \mid\!\sim \beta$.

Again, this formulation stresses the fact that the epistemic outcome of accepting an inductive argument $\alpha \mid\!\sim \beta$ is $\alpha\wedge\beta$.

2.2 Rationality postulates for confirmatory induction

Throughout this section $\alpha \mid\!\sim \beta$ is interpreted as 'β is confirmed by α', which is loosely taken to mean that β expresses some regularity that is implicitly present in α. For instance, α could be a report of astronomical data, and β could be the regularity 'investments decrease when interests increase'. Note that this hypothesis is too weak to count as an *explanation* of the current of investment rate given the interest rate — it is merely being

The following principle expresses that deductive entailment is a special case of confirmation:

(2.3) If $\alpha\rightarrow\gamma$ and $\gamma \mid\!\sim \alpha$, then $\alpha \mid\!\sim \gamma$.

This principle may be generalised to the following postulate, which states that logical consequences of confirmed hypotheses are also confirmed:

(C1) If $\alpha \mid\!\sim \beta$ and $\beta\rightarrow\gamma$, then $\alpha \mid\!\sim \gamma$.

This principle of right weakening is distinguished from the right strengthening postulate (E1) obtained for explanatory induction. Upon noting that combining (C1) and (E1) leads to a situation in which arbitrary evidence confirms any hypothesis, Hempel dismissed (C1) as unacceptable and felt compelled to do the same, since (E1) and (C1) are intended to formalise quite different intuitive notions. Note that (C1) and (I3) can be jointly expressed as

(C1′) If $\alpha \mid\!\sim \beta$ and $\alpha\wedge\beta\rightarrow\gamma$, then $\alpha \mid\!\sim \gamma$.

The next postulate represents a confirmatory variant of reflexivity, analogous to (E2).

(C2) If $\alpha \mid\!\sim \alpha$ and $\alpha \not\mid\!\sim \neg\beta$, then $\beta \mid\!\sim \beta$.

---

[1] Stricly speaking, it would be better to say that the hypothesis is *not disconfirmed* by the evidence.

Taken contrapositively, (C2) states that if some β does not confirm itself, it must be too strong a statement — hence, its negation is so weak that it is confirmed by arbitrary admissible α.

The remaining postulates for confirmatory induction we will consider are considerably stronger. The following postulate states that the conjunction of confirmed hypotheses is itself confirmed:

(C3) If $\alpha \mid\!\sim \beta$ and $\alpha \mid\!\sim \gamma$, then $\alpha \mid\!\sim \beta \wedge \gamma$.

This postulate can be viewed as a completeness assumption with regard to the evidence, in the sense that α contains all the possible states of affairs that are *relevant* for the domain. For instance, when talking about ravens and their colours such complete evidence would contain some black ravens, but also a few non-ravens, some of which are black and some of which are not. Only under such a completeness assumption one can safely assume that 'what you are not told looks like what you are told' (Helft, 1989, p.149). Hempel considered (C3) a valid principle of confirmation. On the other hand, it is clear that (C3) does not necessarily follow from our intuitions about confirmation — indeed, Schlechta (1995) rejects it. Notice that in the presence of (C3) the consistency principle (I2′) is equivalent with

(2.4) If $\alpha \mid\!\sim \beta$, then $\alpha \not\mid\!\sim \neg\beta$.

Notice also that (C3) contradicts an incrementality principle like (E5): weakening the evidence typically disconfirms some hypotheses. A considerably weaker 'left weakening' principle that is compatible with (C3) is expressed by the following postulate:

(C4) If $\alpha \mid\!\sim \gamma$ and $\beta \mid\!\sim \gamma$, then $\alpha \vee \beta \mid\!\sim \gamma$.

This postulate states that if both α and β confirm γ, then the knowledge that at least one of them is true should not disconfirm γ.

The final postulate for confirmatory induction we will consider reinvokes the principle of verification (I1), but in a much stronger sense. Recall that verification expresses that if α confirms β, and γ is a prediction drawn from α and β, then subsequent observation that γ is indeed the case does not disconfirm β. The following postulates extends this to arbitrary γ that are also confirmed by α:

(C5) If $\alpha \mid\!\sim \beta$ and $\alpha \mid\!\sim \gamma$, then $\alpha \wedge \gamma \mid\!\sim \beta$.

In other words: the assumptions necessary to induce β from α do not contradict the assumptions needed to obtain γ from α — thus (C5) expresses another aspect of the 'completeness' of the evidence.

3. Representation theorems

In this section I present some initial representation results regarding the postulates discussed above.

3.1 Strong explanatory structures

The first result is that postulates (E1–5) are strong enough, together with (I1–7), to characterise explanatory induction as a v[illegible]. S[illegible] also envisage weaker (nonmonotonic) notions of explanation, this form of induction is called *strong* explanatory. The corresponding semantic structure is characterised by a subset of *M* (the set of models under consi[illegible]

DEFINITION 3.1. A *strong explanatory structure* is a set $W \subseteq$ [illegible]. The consequence relation it determines is denoted by $\mid\sim_W$ and is defined by: $\alpha \mid\sim \beta$ iff ([illegible] $m \in W$ such that $m_0 \models \alpha, \beta$, and) for every $m \in$ [illegible] $m \models \beta \rightarrow \alpha$. A c[illegible] stro[illegible]

LEM[illegible] *Any strong explanatory consequence relation satisfies* [illegible]

Pro[illegible] that [illegible] (I1–[illegible] i[illegible] $\alpha \wedge \beta \rightarrow \gamma$, then $m \models \beta \rightarrow \alpha \wedge \gamma$ and $m \models \beta \rightarrow \alpha \wedge \gamma$. (I3) [illegible] $\beta \rightarrow \alpha$ [illegible] $\alpha \wedge \beta \rightarrow \gamma$, the[illegible] $m \models \beta \wedge \gamma \rightarrow \alpha$. Furthermore, if [illegible], $m_0 \models \alpha$ and [illegible] γ, then $m_0 \models \beta \wedge \gamma$. (I4). If $m_0 \models \beta$ and $m_0 \models \beta \rightarrow \alpha$, then $m_0 \models \alpha$. (I5–7) [illegible]rivial.

(E1) If $\gamma \mid\sim_W \gamma$, then some model in *W* satisfies γ. Furthermore, if $m \models \beta \rightarrow \alpha$ and $m \models \gamma \rightarrow \beta$, then $m \models \gamma \rightarrow \alpha$. (E2): we have that some model in *W* satisfies γ, while not all models in *W* satisfy $\gamma \rightarrow \neg\alpha$, i.e. there is a model in *W* satisfying $\gamma \wedge \alpha$ and hence α. (E3–5): trivial.

In fact, (I3–7) can be derived from (I1–2, E1–5), but we won't prove that here. Below, we will also make use of the derived rules (I2′) and (2.1

In order to prove completeness, we need to build [illegible] *W* from a given consequence relation $\mid\sim$ satisfying (I1–7, E1–5), such that $\alpha \mid\sim \beta$ iff $\alpha \mid\sim_W \beta$. This structure is defined as follows:

$$W = \{m \in U \mid \text{for all } \alpha, \beta \text{ such that } \alpha \mid\sim \beta\text{: } m \models \beta \rightarrow \alpha\}$$

The follow[illegible] strong explanatory hypothesis is satisfiable i[illegible]

LEMMA 3.3. *Let $\mid\sim$ be a consequence relation satisfying (I1–7, E1–5), and let W be defined as above. If $\alpha \not\mid\sim \beta$ then there is a model $m \in W$ such that $m \models \alpha, \neg\beta$.*

Proof. Let $\alpha \not\mid\sim \beta$; we will prove that $\{\beta\} \cup \{\delta \rightarrow \gamma \mid \gamma \mid\sim \delta\}$ is satisfiable. Suppose not, then by compactness there is a finite $\Delta \subseteq \{\delta \rightarrow \gamma \mid \gamma \mid\sim \delta\}$ such that $\beta \rightarrow \neg\Delta$. Furthermore, since $\varphi \not\mid\sim \psi$ for any $\psi \rightarrow \varphi \in \Delta$, we have $\psi \rightarrow \varphi \not\mid\sim$ **true** for any $\psi \rightarrow \varphi \in \Delta$ by (E3), $\Delta \not\mid\sim$ **true** by (E4), and $\Delta \not\mid\sim \beta$ by (I5) and (E1). But then by (I2′) $\beta \rightarrow \neg\Delta$, a contradiction.

Furthermore, we have that every inadmissible formula is unsatisfiable in *W*.

LEMMA 3.4. *Let $\mid\!\sim$ be a non-empty consequence relation satisfying (I1–7, E1–5), and let W be defined as above. If $\gamma \not\mid\!\sim \gamma$ then γ is unsatisfiable in W.*

Proof. Let $\alpha \mid\!\sim \beta$, then $\mathbf{true} \mid\!\sim \mathbf{true}$ by (E5) and (I4). Furthermore, if $\gamma \not\mid\!\sim \gamma$ then $\neg\gamma \mid\!\sim \mathbf{true}$ by (E2), hence $m \models \mathbf{true} \rightarrow \neg\gamma$ for every $m \in W$.

I will now show that W defines a consequence relation that is included in $\mid\!\sim$.

LEMMA 3.5. *Let $\mid\!\sim$ be a non-empty consequence relation satisfying (I1–7, E1–5), and let W be defined as above. If $\alpha \mid\!\sim_W \beta$ then $\alpha \mid\!\sim \beta$.*

Proof. Suppose that $\alpha \not\mid\!\sim \beta$, we will show that either no model in W satisfies α or there exists a model in W that does not satisfy $\beta \rightarrow \alpha$.
First of all, if $\beta \not\mid\!\sim \beta$ then β is unsatisfiable in W according to Lemma 3.4. For the remainder of the proof we will assume that $\beta \mid\!\sim \beta$. Define $\Gamma_0 = \{\neg\alpha\} \cup \{\delta \mid \delta \mid\!\sim \beta\}$; we will first show that Γ_0 is satisfiable. Suppose not, then by compactness there is a finite $\Delta \subseteq \{\delta \mid \delta \mid\!\sim \beta\}$ such that $\neg\alpha$ is inconsistent with $\beta \rightarrow (\Delta \rightarrow \alpha)$; by (2.1) $\Delta \mid\!\sim \neg\alpha \rightarrow \beta$ (recall that $\beta \mid\!\sim \beta$), and by (E4) $\mid\!\sim$ (E4) and (E5), we obtain $\mid\!\sim \beta$. Contradiction, so Γ_0 is satisfiable.
Let $m_0 \models \Gamma_0$; clearly $m_0 \models \neg\alpha$, and since $\beta \in \Gamma_0$, $m_0 \models \beta$. It remains to prove that m_0 is in W; i.e., that for all φ, ψ such that $\varphi \not\mid\!\sim \psi$ we have $m_0 \models \psi \rightarrow \varphi$. Let $\varphi \mid\!\sim \psi$; if $\neg\beta \not\mid\!\sim \psi$, then by (E1) $\varphi \not\mid\!\sim \psi \vee \beta$, and by (E3) $\varphi \not\mid\!\sim \varphi \vee \beta$, thus $\psi \rightarrow \varphi \in \Gamma_0$ and therefore $m_0 \models \psi \rightarrow \varphi$. On the other hand, if $\neg\beta \mid\!\sim \psi$ then by (E3) and (E5) $\beta \rightarrow \neg\psi \mid\!\sim \mathbf{true}$, by (E1) $\beta \rightarrow \neg\psi \mid\!\sim \beta$, and by (E4) and (E5) $\neg\psi \mid\!\sim \beta$; thus $\neg\psi \in \Gamma_0$ and therefore $m_0 \models \neg\psi$, hence $m_0 \models \psi \rightarrow \varphi$.

Armed with the previous three lemmas we can prove the correctness of (I1–7, E1–

THEOREM 3.6 (Representation theorem for strong explanatory relations). *A consequence relation is strong explanatory if it satisfies (I1–7, E1–5).*

Proof. The only-if part is Lemma 3.2. For the if part, let $\mid\!\sim$ be a non-empty consequence relation satisfying (I1–7, E1–5), and let

$$W = \{m \in U \mid \text{for all } \alpha, \beta \text{ such that } \alpha \mid\!\sim \beta\text{: } m \models \beta \rightarrow \alpha\}$$

Suppose $\alpha \mid\!\sim \beta$, then by the construction of W, $m \models \beta \rightarrow \alpha$ for all $m \in W$. Furthermore, by Lemma 3.3 there is a model in W satisfying β. We may conclude that $\alpha \mid\!\sim_W \beta$. Conversely, if $\alpha \mid\!\sim_W \beta$ then Lemma 3.5 proves that $\alpha \mid\!\sim \beta$. We conclude that W defines a consequence relation that is exactly $\mid\!\sim$. For an empty consequence relation put $W = \varnothing$.

3.2 Preferential confirmatory structures

The second result is the characterisation of (C1–5), which is an adaptation of a result by Kraus, Lehmann & Magidor (1990). The main idea is to construct, from the evidence, a set of *regular* interpretations, which exhibit the same regularities as the evidence, and to

define a hypothesis to be confirmed if it is true in every regular interpretation. Notice that such regular interpretations are not necessarily models of the evidence. For instance, in the ravens example we may interchange the names of ravens and non-ravens without invalidating the inductive conclusion.

However, in my first investigations I have followed the more traditional approach of assuming a preference ordering on interpretations, and to define the regular interpretations as the most preferred models of the evidence. This idea is motivated by the approaches of Helft (1989) and De Raedt & Bruynooghe (1993), who implement the completeness assumption regarding the evidence by the Closed World Assumption. As a result, the characterisation of confirmatory induction stated below as Theorem 3.14 is a close variant of Kraus *et al.*'s characterisation of preferential reasoning. readers familiar with the latter paper may want to skip the present section, which is in completeness' sake. The implications of this correspondence between form of induction and nonmonotonic reasoning will be discussed in section 4 below.

> DEFINITION 3.7. A *preferential confirmatory structure* is a triple $W = \langle S,l,< \rangle$, where S is a set of *states*, l: $S \to M$ is a function that labels every state with a model, and $<$ is a strict partial order on S, called the *preference ordering*, that is smooth[2]. A state $s \in S$ satisfies a formula $\alpha \in L$ iff $l(s) \models \alpha$; the set of states satisfying α is denoted by $[\alpha]$, and a minimal element of $[\alpha]$ (wrt. $<$) will be called a *regular* state for α. The consequence relation defined by W is denoted by $\mid\!\sim_W$ and is defined by: $\alpha \mid\!\sim_W \beta$ iff (*i*) there is a state $s \in S$ satisfying α, and (*ii*) every regular state for α satisfies β. A consequence relation is called preferential confirmatory iff it is defined by a preferential confirmatory structure.

In comparison with the preferential models of (Kraus *et al.*, 1990), the only difference is that in a confirmatory argument the evidence is required to be satisfiable, in order to guarantee the validity of (I2). The intermediate states, mainly needed for technical reasons, and can be interpreted as the models the reasoning agent considers possible in that epistemic state.

> LEMMA 3.8 (Soundness). *Any preferential confirmatory consequence relation satisfies (I1–7, C1–5).*
> *Proof.* The proof is easy and will only be carried out for (C1–5).
> (C1): if all regular states for α satisfy β and $\beta \to \gamma$, then all regular states for α satisfy γ. (C2): suppose $[\alpha]$ is non-empty, and not all regular states for α satisfy $\neg\beta$; it follows that some state in S satisfies β. (C3): if all regular states for α satisfy β and γ, then clearly they satisfy $\beta \wedge \gamma$.
> (C4): note that $[\alpha \vee \beta] = [\alpha] \cup [\beta]$; thus, if $[\alpha]$ and $[\beta]$ are non-empty then so is $[\alpha \vee \beta]$. Furthermore, the set of regular states for $\alpha \vee \beta$ is a subset of the

[2] I.e. for any $S' \subseteq S$ and for any $s \in S'$, either s is minimal in S', or there is a $t \in S'$ such that $t < s$ and t is minimal in S'. This condition is satisfied if $<$ does not allow infinite descending chains.

union of the regular states for α and β, since a state cannot be minimal in $[\alpha\vee\beta]$ without being minimal in at least one of $[\alpha]$ and $[\beta]$.

(C5): we need the fact that $<$ is a smooth partial order. Suppose that $[\alpha]$ is non-empty, and all regular states for α satisfy β and γ — clearly, $[\alpha\wedge\beta]$ is non-empty. Now, let s be regular for $\alpha\wedge\beta$, then $s\in[\alpha]$; I will prove that s is regular for α. Suppose not, then there is a $t\in[\alpha]$ such that $t<s$ and t is regular for α. Now, every state regular for α satisfies β, hence $t\in[\alpha\wedge\beta]$. But this contradicts the minimality of s in $[\alpha\wedge\beta]$, hence s is regular for α and thus satisfies γ.

As a matter of fact, (I1) and (I4–6) can be derived from (I2–3, I7, C[illegible]-5). Another derived rule that will be used below is (2.4).

In order to prove completeness, we need to build a preferential [illegible] structure W from a given consequence relation $\mid\!\sim$ satisfying (I1–7, C1–5), such that $\alpha\mid\!\sim\beta$ iff $\alpha\mid\!\sim_W\beta$. As in the case of explanatory structures, such a confirmatory structure is built from a specific set of models. These models are selected relative to a given formula, as follows.

DEFINITION 3.9. Let $\mid\!\sim$ be a conjectural consequence relation. The model $m\in U$ is said to be *normal for* α iff for all β in L such that $\alpha\mid\!\sim\beta$, $m\Vdash\beta$.

So, a model is normal for a formula if it satisfies every confirmed hypothesis. Thus, given certain evidence the set of normal models decreases when the set of confirmed hypotheses increases. Notice that every model in U is normal for an inadmissible formula, which is therefore not satisfied by some of its normal models. An admissible formula is satisfied by every normal model, however. Notice also that if α is admissible and γ is inadmissible, then by (C2) $\alpha\mid\!\sim\neg\gamma$, hence no normal model for α satisfies γ.

The set of models normal for admissible formulas will be used to build a preferential confirmatory structure. The following lemma states the key result about normal models.

LEMMA 3.10. *Suppose a consequence relation $\mid\!\sim$ satisfies (C1) and (C3), and let α be an admissible formula. All normal models for α satisfy β iff $\alpha\mid\!\sim\beta$.*

Proof. The if part follows from Definition 3.9.

For the only-if part, suppose $\alpha\mid\!\sim\alpha$ and [illegible] $\mid\!\sim\beta$, I will now that there is a normal model for α that does not satisfy β. Let $\Gamma_0=\{\alpha\wedge\neg\beta\}\cup\{\delta\mid\alpha\mid\!\sim\delta\}$; it suffices to show that Γ_0 is satisfiable. Suppose not, then by compactness there is a finite $\Delta\subseteq\{\delta\mid\alpha\mid\!\sim\delta\}$ such that $\Delta\rightarrow\beta$, i.e. $\alpha\rightarrow(\Delta\rightarrow\beta)$; by (C1) $\alpha\mid\!\sim\Delta\rightarrow\beta$. But by (C3) $\alpha\mid\!\sim\Delta$; using (C3) and (C1) we obtain $\alpha\mid\!\sim\beta$, a contradiction.

Notice from the proof of Lemma 3.10 that normal models exist for any admissible α.

Given a consequence relation $\mid\!\sim$ satisfying (I1–7, C1–5), the completeness proof is based on a preferential confirmatory structure $W = \langle$ [illegible] follows:

(1) $S = \{\langle m,\alpha\rangle \mid \alpha$ is an admissible formula, and m is a normal model for $\alpha\}$;
(2) $l(\langle m,\alpha\rangle) = m$;
(3) $\langle m,\alpha\rangle < \langle n,\beta\rangle$ iff $\alpha\vee\beta \mid\!\sim \alpha$ and $m \not\models \beta$.

Thus, states are pairs of admissible formulas and normal models. The labeling function simply maps a state to the model it contains. Condition (3) defines the preference ordering between states: note that $\beta \mid\!\sim \alpha$ is a special case of $\alpha\vee\beta \mid\!\sim \alpha$ by means of (C4), and the fact that α is admissible. The condition $m \not\models \beta$ is added to make the ordering irreflexive; note that as a consequence any $\langle m,\alpha\rangle \in S$ is minimal in $[\alpha]$.

The main difference between the preferential consequence relations of Kraus *et al.* and my preferential confirmatory consequence relations is the way unsatisfiable formulas are treated. In Kraus *et al.*'s framework unsatisfiable formulas are characterised by the fact that they have every formula in L as a plausible consequence, which means that they don't have normal models. In my framework, unsatisfiable formulas confirm no hypotheses, and have all models in U as normal models. In both cases, the structure W that is used to prove completeness contains only satisfiable formulas in its states. This means that we can replicate most of Kraus *et al.*'s results about the structure W (KLM 5.13 refers to (Kraus *et al.*

PROPOSITION 3.11. *(1) (KLM 5.13) The relation $<$ is a strict partial order.*
(2) (KLM 5.15) The relation $<$ is smooth: for any $s \in [\alpha]$, either s is minimal in $[\alpha]$ or there exists a state $t<s$ minimal in $[\alpha]$.
(3) (KLM 5.11) If $\alpha\vee\beta \mid\!\sim \alpha$ and m is a normal model for α that satisfies β, then m is a normal model for β.
(4) (KLM 5.14) $\langle m,\alpha\rangle$ is minimal in $[\beta]$ iff $m \not\models \beta$ and $\alpha\vee\beta \mid\!\sim \alpha$.

The first two propositions express that W is a preferential confirmatory structure. The remaining two are used in the proof of the following lemma.

LEMMA 3.12. *Let $\mid\!\sim$ be a consequence relation satisfying (I1–7, C1–5), and let W be defined as above. If $\alpha \mid\!\sim \beta$ then $\alpha \mid\!\sim_W \beta$.*

Proof. Suppose that $\alpha \mid\!\sim \beta$; we will show that (*i*) there is a state in S satisfying α, and (*ii*) every minimal state in $[\alpha]$ satisfies β.

(*i*) By (I4) α is admissible; furthermore, by (2.4) $\alpha \not\mid\!\sim \neg\beta$, so by Lemma 3.10 there exists a model m normal for α. We conclude that $\langle m,\alpha\rangle \in [\alpha]$.

(*ii*) Suppose $\langle n,\gamma\rangle$ is minimal in $[\alpha]$, then γ is an admissible formula, n is a normal model for γ that satisfies α, and $\gamma\vee\alpha \mid\!\sim \gamma$ by Proposition 3.11 (4). By Proposition 3.11 (3) n is a normal model for α, hence $n \models \beta$.

The following lemma proves the converse of Lemma 3.12, and completes the proof of the representation theorem.

LEMMA 3.13. *Let $\mid\!\sim$ be a consequence relation satisfying (I1–7, C1–5), and let W be defined as above. If $\alpha \mid\!\sim_W \beta$ then $\alpha \mid\!\sim \beta$.*
Proof. Suppose $\alpha \mid\!\sim_W \beta$, then α must be admissible (since no state in *S* satisfies an inadmissible formula). Furthermore, given any model *m* normal for α, $\langle m,\alpha\rangle$ is minimal in $[\alpha]$, hence *m* satisfies β, and the conclusion follows by Lemma 3.10.

We may now summarise.

THEOREM 3.14 (Representation theorem for preferential confirmatory consequence relations). *A consequence relation is preferential confirmatory iff it satisfies (I1–7, C1–5).*
Proof. The only-if part is Lemma 3.8. For the if part, let $\mid\!\sim$ be a consequence relation satisfying (I1–7, C1–5) and let *W* be defined as above. Lemmas 3.12 and 3.13 prove that $\alpha \mid\!\sim \beta$ iff $\alpha \mid\!\sim_W \beta$, i.e. $\mid\!\sim$ is preferential confirmatory.

4. Discussion

In the last part of this paper I will discuss the significance of the approach and results presented above and make a comparison with related work.

Although the proof-theoretic approach of section 2 is nicely balanced by a semantic analysis in section 3, I think the former — analysing the logical characteristics of induction on the level of consequence relations — constitutes the main contribution of the paper. This approach has of course been heavily influenced by work of Gabbay (1985), Makinson (1989) and Kraus *et al.* (1990) on nonmonotonic consequence relations and operations. If anything, my approach demonstrates that their method is in fact a methodology. In this respect I should also mention Gärdenfors' work on rationality postulates for belief revision operators (1988).

The representation theorems have been stated mainly as a starting point for further research in this area. In Section 2 I interpreted explanatory induction to reversed deduction. An equivalent but more interesting perspective is obtained by defining

$$\alpha \mid\!\sim \beta \;\Leftrightarrow\; Cn(\alpha) \subseteq Cn(\beta) \subset L$$

where $Cn(\alpha) = \{\gamma \mid \alpha \vdash \gamma\}$ denotes the set of consequence or the *explanatory power* of α. In words: given evidence α, β is an explanatory hypothesis of the explanatory power of β exceeds that of α (without reaching inconsistency). This view of induction as *explanation-preserving* reasoning can be subsequently generalised by considering weaker (nonmonotonic) explanation mechanisms $\mid\!\sim$.[3]

[3]As Daniel Lehmann observed, any inductive consequeynce relation thus defined will satisfy transitivity (Lehmann, personal communication). We are currently studying cumulative, preferential and rational explanation mechanisms (Kraus *et al.*, 1990; Lehmann & Magidor, 1992).

The preferential semantics for modelling the form of confirmatory induction considered by Hempel, Helft and De Raedt has of course been borrowed from work in nonmonotonic reasoning. Readers familiar with that field will have recognised some of the postulates — for instance, our stronger verification principle (C5) is known as cautious monotonicity. On one hand, the correspondence is not so surprising: both forms of reasoning rely on some form of completeness of the premisses, be it 'what you don't know is false' or 'what you don't know resembles what you know'. A pragmatic difference is that in nonmonotonic reasoning conclusions are usually specific ('it flies'), while in induction conclusions are typically general ('all crows are black'). On the other hand, induction often proceeds from incomplete evidence. A formalisation of weaker forms of confirmatory induction requires to relax the completeness assumption regarding the evidence, and thus to drop (C3) — but (C3) will be sound for any semantics that selects a subset of the models of the evidence, all of which should satisfy the hypothesis. A semantics that constructs a set of regular interpretations from the evidence, some of which do **not** satisfy the evidence (for instance because names have been permuted) seems a promising research direction.

Philosophically speaking, the view of induction as inference of explanations owes much to the ideas of Peirce. In fact, Peirce identified the formation of explanatory hypotheses with reversed deduction (Hartshorne *et al.*, 1931–58, Vol.1, p.117). The framework of rationality postulates provides a finer-grained perspective on the matter. Hempel was the first to analyse the logical relation of confirmation, and even proposed a number of rationality postulates or adequacy conditions, among which (variants of) I2′, I6, I7, C1, C3, 2.3 and 2.4 (see Hempel, 1943). He then proceeded to develop a definition of confirmation, which comes surprisingly close to minimal Herbrand model semantics. However, Hempel never gave a complete axiomatisation of his definition.

In summary, we have proposed a new framework for thinking about the logic of induction, a problem that has been bothering philosophers for over twenty centuries, by bringing together old and recent work in philosophy, logic, and artificial intelligence.

References

R. CARNAP (1950), *Logical Foundations of Probability*, Routledge & Kegan Paul, London.

L. DE RAEDT & M. BRUYNOOGHE (1993), 'A theory of clausal discovery', in *Proc. 13th International Conference on Artificial Intelligence IJCAI'93*, Morgan Kaufmann, San Mateo: 1058–1063.

P.A. FLACH (1995), *Conjectures — an inquiry concerning the logic of induction*. PhD thesis, Tilburg University.

D. GABBAY (1985), 'Theoretical foundations for non-monotonic reasoning in expert systems', in *Logics and Models of Concurrent Systems*, K.R. Apt (ed.), Springer-Verlag, Berlin: 439–457.

P. GÄRDENFORS (1988), *Knowledge in Flux — modeling the dynamics of epistemic states*, MIT Press, Cambridge, MA.

C. HARTSHORNE, P. WEISS & A. BURKS, EDS (1931-58), *Collected Papers of Charles Sanders Peirce*, Harvard University, Cambridge.

N. HELFT (1989), ‘Induction as nonmonotonic inference’, in *Proc. First International Conference on Knowledge Representation and Reasoning KR’89*, Morgan Kaufmann, San Mateo: 149–156.

C.G. HEMPEL (1943), ‘A purely syntactical definition of confirmation’, *Journal of Symbolic Logic* **6**(4): 122–143.

C.G. HEMPEL (1945), ‘Studies in the logic of confirmation’, *Mind* **54**(213): 1–26 (Part I); **54**(214): 97–121 (Part II).

S. KRAUS, D. LEHMANN & M. MAGIDOR (1990), ‘Nonmonotonic reasoning, preferential models and cumulative logics’, *Artificial Intelligence* **44**: 167-207.

D. LEHMANN & M. MAGIDOR (1992), ‘What does a Conditional Knowledge Base Entail?’, *Artificial Intelligence* **55**(1): 1-60.

D. MAKINSON (1989), ‘General theory of cumulative inference’, in *Proc. 2nd International Workshop on Non-Monotonic Reasoning*, M. Reinfrank, J. de Kleer, M.L. Ginsberg & E. Sandewall (eds.), Lecture Notes in Artificial Intelligence 346, Springer-Verlag, Berlin: 1–18.

K. SCHLECHTA (1995), ‘A reduction of the theory of confirmation to the notions of distance and measure’, *Proc. European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty ECSQARU ’95*, Lecture Notes in Artificial Intelligence 946, Springer-Verlag, Berlin: 387–394.