Synthese (2009) 170:131–146
DOI 10.1007/s11229-008-9361-3

Artificial explanations: the epistemological interpretation of explanation in AI

Andrés Páez

Received: 3 November 2007 / Accepted: 5 June 2008 / Published online: 1 July 2008

Abstract In this paper I critically examine the notion of explanation used in artificial intelligence in general, and in the theory of belief revision in particular. I focus on two of the best known accounts in the literature: Pagnucco's abductive expansion functions and Gärdenfors' counterfactual analysis. I argue that both accounts are at odds with the way in which this notion has historically been understood in philosophy. They are also at odds with the explanatory strategies used in actual scientific practice. At the end of the paper I outline a set of desiderata for an epistemologically motivated, scientifically informed belief revision model for explanation.

Keywords Explanation · Epistemology · Belief revision · Artificial intelligence

1 Introduction

Researchers in artificial intelligence often use epistemological notions in a fast and loose manner that wouldn't pass muster with philosophers. In the theory of knowledge representation, for example, there is no attempt to show that the information which is being represented is justified and true, two necessary conditions for knowledge. Given the purposes of the information sciences, in most cases ignoring basic philosophical distinctions does not have serious practical consequences. Nonetheless, this practice has created a gap between philosophy and AI. It is my view that some areas in artificial intelligence can benefit from the study of recent developments in epistemology and the philosophy of science, and that traditional epistemology can profit from using some of the formal methods used in AI.

A. Páez (✉)
Universidad de los Andes, Bogota, Colombia
e-mail: andrespaez@gmail.com

This essay focuses on the concept of *explanation* as it is used in AI in general, and in the theory of belief revision in particular. I will argue that the notion of explanation used in AI is at odds with the way in which this notion has historically been understood in philosophy. It is also at odds with the explanatory strategies used in actual scientific practice. At the end of the paper I outline the elements which are necessary to develop an epistemologically motivated, scientifically informed belief revision model for explanation.

2 Explanation in AI

The terms 'explanation' and 'abduction' are often used in AI in areas such as diagnostic reasoning, database updating, vision and text understanding, knowledge acquisition, belief revision, and natural language processing.[1] Approaches to explanation can be divided into two categories: set-cover based and logic based. The former approach is mostly used in solving diagnostic problems. In a diagnostic problem, we try to explain a set of *manifestations* (observations, symptoms, etc.) using our knowledge of the *disorders* (diseases, malfunctions, etc.) causally associated with them. An explanation consists in a set of causes whose related effects are a superset of all the effects observed. The best-known example of this approach is Peng and Reggia (1990) parsimonious covering theory.

Logic based approaches to explanation are used when the domain is best represented by a logical theory. This occurs, for example, when the domain involves disjunctive or negative information which cannot be expressed easily by simple mappings between causes and effects. Since the representation of an agent's beliefs requires the use of a logical theory, all belief revision theories of explanation are logic based.

Logic based approaches can be divided into two classes according to whether the inference relation involved is monotonic or nonmonotonic. The following is the most common general definition of abduction in the monotonic approaches. An abduction or explanation—the two terms are used interchangeably in AI—for a formula ϕ with respect to a background theory T is a formula ψ that satisfies the following conditions:

(i) $T \cup \{\psi\} \vdash \phi$
(ii) $T \cup \{\psi\}$ is consistent.[2]

In dealing with the question of how to compute updates of a logical database, for example, an explanation is used to support the addition of a piece of information ϕ into a database. An update request "insert (ϕ)" can be achieved by finding some formula consistent with the database such that the union of the set of ground facts in the database and the formula yields ϕ as a logical consequence (within the logic programming domain).

This definition is quite general and can generate many different explanations. Since normally we are interested in the best possible explanation, several restrictions are imposed on the possible candidates. The explanation must be minimal, simple, nontrivial, and it must fulfill certain syntactic conditions. These restrictions are characterized

[1] See Paul (1993) and Kakas et al. (1993) for a survey.

[2] It is generally assumed that neither $\psi \vdash \phi$ nor $T \vdash \phi$.

in different ways depending on the field of application. This initial filter still allows a great number of possible explanations. In order to select the best one, a large variety of methods, mostly quantitative, has been suggested. These include coherence (Thagard 1989; Ng and Mooney 1990), cost (Stickel 1991), and utility (Ram and Leake 1991), among others.

In the theory of belief revision, the background theory is interpreted as a belief set K, the linguistic representation of an agent's epistemic state at time t. Accepting a sentence in a belief set K entails full belief, in the sense that in K there is no doubt about the truth of that sentence for the time being. The formulas ϕ and ψ are interpreted as epistemic inputs. Instead of simply accepting an epistemic input ϕ, an abduction allows an agent to find some explanation or justification for it in the light of his currently held beliefs. The best-known approach is Pagnucco (1996) abductive expansion functions, which are based on the AGM model, the model for belief revision introduced by Alchourrón et al. (1985). I will examine Pagnucco's approach in detail in the next section. Wassermann and Dias (2001) have adapted Pagnucco's approach to model the expansion of belief bases.

Nonmonotonic logic based approaches to explanation incorporate the fundamental idea that explanations are defeasible, that is, that the addition of new information may eliminate the explanatory value of a formula. Nonmonotonicity has been introduced in several ways in AI. The most straightforward manner to eliminate monotonicity is by replacing the classical consequence relation used in the previous models by a nonmonotonic inference relation. This approach is followed by Aliseda (2006), who sees abductive inference as a more structured form of classical consequence, and by Boutilier (1994), who uses conditional logics to define a predictive notion of explanation. Perhaps the best known approach is Mackinson & Gärdenfors' modeling of nonmonotonic reasoning on the revision of expectations (1991; Gärdenfors and Makinson 1994). Other accounts include Poole's (1989) assumption-based frameworks and Levesque's (1989) knowledge-level approach. Despite their technical differences, in all of these theories the function of an explanation is the same as in the monotonic ones: to support the addition to a belief set of an epistemic input by providing an explanation or justification for it.

3 Pagnucco on explanation

In this section I will discuss Pagnucco's (1996) approach to explanation, which is the most detailed theory of explanation in the belief revision literature. Pagnucco explains the guiding idea of his approach in the following passage:

> Many belief revision frameworks … aim to solely incorporate the epistemic input and any resulting consequences. However, it is our contention that a more natural and advantageous approach is for the agent[s] to first seek some explanation or justification for the epistemic input in light of their currently held beliefs and to incorporate this explanation together with the epistemic input into their new epistemic state (pp. 4–5).

Pagnucco defines an abduction or explanation—he uses the two terms interchangeably—for a sentence ϕ with respect to a domain theory K as a sentence ψ that satisfies the following two conditions (p. 79):

(i) $K \cup \{\psi\} \vdash \phi$
(ii) $K \cup \{\psi\}$ is consistent.

In general, there will be many sentences that qualify as abductions or explanations of ϕ. Pagnucco discusses several criteria that can be used to select the best abduction, although in most cases these criteria do not determine a unique choice. The first criterion is minimality: "assume as little as possible in proving a formula ϕ. This expresses the desire to avoid superfluous abductions" (p. 80). Using the consequence relation, Pagnucco introduces a partial order over the set of abductions of ϕ with respect to K. According to the weakness ordering, some abductions will be weaker than ϕ itself. In that case, the result of expanding K by ψ in the light of ϕ will be $Cn(K \cup \{\phi\})$. The differences among those abductions that are weaker than ϕ is effectively obliterated.

The second consideration is that an abduction should not be trivial. An abduction ψ of ϕ with respect to K is trivial iff $\psi \vdash \phi$. The idea is that the abduction should make use of K and not be able to prove the new information on its own.[3] Another way in which abductions may be seen as trivial is in those cases in which $\psi \rightarrow \phi$ is a theorem in K. Using the deduction theorem, $(\psi \rightarrow \phi) \in Cn(K)$ iff $\phi \in Cn(K \cup \{\psi\})$, we can prove that if $(\psi \rightarrow \phi) \in K$, ψ is an abduction of ϕ. "These types of abduction are inherent to the logic in a certain sense and may always be obtained regardless of the domain theory (up to inconsistency)" (pp. 82–83). Pagnucco claims that these trivial abductions can be weeded out using the other selection criteria described here, together with the selection mechanisms associated with the operation of abductive expansion described below.

Finally, Pagnucco argues that it would be desirable to have a criterion to determine degrees of specificity for abduction. Different explanations demand different degrees of specificity. In his view, "abduction can in a sense be viewed as an inference 'backwards' over an implication; from consequent to antecedent. One way to view specificity then, is to treat propositions further 'back' along an implication chain as more specific" (p. 83). He explores several possible ways in which levels of specificity can be formally determined, but none of them lead to a satisfactory definition.

As we saw above, Pagnucco's purpose in defining an abduction or explanation for a formula is to provide an inquiring agent with a reason to accept the new information she has gleaned. Pagnucco defines an operation called *abductive expansion* which captures this idea. An operation of abductive expansion of a belief set K by a formula ϕ is a function $\oplus$ that adds to K some formula ψ which explains ϕ, that is, a formula ψ which together with K implies ϕ without making the set inconsistent. Pagnucco offers the following definition:

[3] This condition is equivalent to the NES (No-Entailment-by-Singular-Sentence) requirement discussed by Achinstein (1983).

$K^{\oplus}_{\phi}$ is an abductive expansion of K with respect to ϕ iff
$K^{\oplus}_{\phi} = Cn(K \cup \{\psi\})$ for some $\psi \in L$ such that:
(i) $K \cup \{\psi\} \vdash \phi$ and
(ii) $K \cup \{\psi\} \nvdash \bot$
$K^{\oplus}_{\phi} = K$ if no such ψ exists.

He then introduces the following rationality postulates for the abductive expansion function $\oplus$:

(K$^{\oplus}$1) $K^{\oplus}_{\phi} = Cn(K^{\oplus}_{\phi})$. (closure)
(K$^{\oplus}$2) If $\sim\phi \notin K$, then $\phi \in K^{\oplus}_{\phi}$. (limited success)
(K$^{\oplus}$3) $K \subseteq K^{\oplus}_{\phi}$. (inclusion)
(K$^{\oplus}$4) If $\sim\phi \in K$, then $K^{\oplus}_{\phi} = K$. (vacuity)
(K$^{\oplus}$5) If $\sim\phi \notin K$, then $\sim\phi \notin K^{\oplus}_{\phi}$. (consistency)
(K$^{\oplus}$6) If $K \vdash \phi \leftrightarrow \gamma$, then $K^{\oplus}_{\phi} = K^{\oplus}_{\gamma}$. (preservation)
(K$^{\oplus}$7) $K^{\oplus}_{\phi} \subseteq Cn(K^{\oplus}_{\phi\vee\gamma} \cup \{\phi\})$. (supplementary 1)
(K$^{\oplus}$8) If $\sim\phi \notin K^{\oplus}_{\phi\vee\gamma}$, then $K^{\oplus}_{\phi\vee\gamma} \subseteq K^{\oplus}_{\phi}$. (supplementary 2)

These postulates only impose basic constraints on the operation of abductive expansion. In order to select the best abductive expansions, Pagnucco constructs three selection mechanisms: epistemic entrenchment, partial meet abductive expansion functions, and a construction based on Grove's system of spheres. I will only examine the first two.

Pagnucco's analysis of epistemic entrenchment is based on the notion of an expectations ordering proposed by Gärdenfors and Makinson (1994). An expectations ordering is an epistemic entrenchment ordering that only satisfies transitivity, dominance, and conjunctiveness. Maximality and minimality are dropped in order to apply the notion of epistemic entrenchment to nonmonotonic reasoning.

Pagnucco defines the notion of an abductive entrenchment ordering by adding a fourth condition to the notion of an expectations ordering:

(AE1) When $K \neq K_{\bot}$, $\phi \in K$ iff $\gamma \leq \phi$ for all $\gamma \in L$. (maximality)

An abductive entrenchment ordering is an expectations ordering in which all of the agent's beliefs are maximally entrenched. This makes it easy to extract the agent's current state of belief from the ordering and consider only the sentences that he does not currently accept. Pagnucco then shows that for any well-behaved abductive expansion function there exists an ordering of abductive epistemic entrenchment that generates the function.

Partial meet abductive expansion functions, on the other hand, are modeled after AGM's partial meet *contraction* functions. The dual of AGM's $K \perp \phi$, the set of all belief sets K' that are maximal subsets of K that fail to imply ϕ, is $K \top \phi$, the set of all belief sets K' that are maximally consistent supersets of K that imply ϕ. Let S be a function that picks out elements from $K \top \phi$. The intended interpretation is that S picks out the members of $K \top \phi$ that rank highest according to some index of explanatory merit. Using the function S, Pagnucco defines a partial meet abductive expansion function thus:

$$\text{(Def Part } \oplus\text{)} \quad K^{\oplus}_{\phi} = \bigcap S(K \top \phi) \text{ whenever } K \top \phi \text{ is nonempty;}$$
$$K^{\oplus}_{\phi} = K \text{ otherwise}$$

Levi (2004) has shown that this construction suffers from the same problem that has been identified in the case of AGM's partial meet contraction functions: partial meet abductive expansions are not closed under intersection. The intersection of two abductive expansions that rank best according to the selection function need not rank best according to the same function.

Pagnucco recognizes that further restrictions can be imposed on these selection mechanisms to make them more precise. For example, if the requirement of minimality is imposed on abductions, the result is a full meet abductive expansion function that turns out to be equivalent to AGM expansion. And if abductions are required to be maximally specific, the result is a maxichoice abductive expansion function. The choice of one restriction over another will depend on non-logical factors such as the intended use of the abductive operation.

Pagnucco does not offer an epistemological interpretation of the notions of abduction and abductive expansion, but he is aware that his definition of abduction might not capture what he calls "the intuitive notion of explanation." He refers the reader to Salmon's *Four Decades of Scientific Explanation* (1989) for details about this "intuitive notion," and he mentions that "parallels can be drawn from the discussion of work concerning Hempel and Oppenheim's deductive-nomological model of explanation" (p. 10, n. 8). In the rest of this section I examine whether the accounts of explanation and abductive expansion provided by Pagnucco have anything in common with the notion of explanation as it is understood in the philosophical literature.

The first obvious difference between the philosophical notion of explanation and the notion used in belief revision is that an explanation, in its epistemological sense, starts with a fact that has been previously accepted by an agent. One cannot try to explain something one does not believe is true. In contrast, the purpose of an abductive expansion is to provide grounds for believing that fact, since it has not been antecedently accepted in K. Some researchers in AI are aware of this difference. Boutilier & Becher, for example, state: "Our use of the term observation is somewhat non-traditional—it is a fact that has yet to be accepted (in some sense) as a belief" (1995, p. 59). But Pagnucco and others are oblivious to the difference. Aliseda, for example, states: "We [do] not consider accepted beliefs, since these do not call for explanation" (2006, p. 183).

Virtually every account of explanation in the philosophical literature also includes the requirement that the explanans statements of a bona fide explanation must be true, or must have been accepted as true for the time being. Pagnucco's model has the opposite requirement. In his model, and in most other models in AI, it is assumed that an explanation cannot be either directly observable nor already implied by the agent's current beliefs. In other words, an explanation of a fact must always be new to the agent's belief system. The main reason for this requirement is that belief systems are assumed to be sets closed under logical implication. If the explanation were already part of our belief system, the abduction would be superfluous. In consequence, an abduction always leads to an increment in our knowledge.

This requirement, however, is at odds with the development of a scientific theory. The explanation of a known fact has often been the result of developing the unforeseen consequences of the scientific theories accepted at a given time. Scriven (1959, p. 461), for example, uses cases taken from the history of science to illustrate how sometimes the derivation of an explanandum from an explanans can present considerable mathematical difficulties. Without such derivation, one cannot say that the phenomenon has been explained even though the explanation was already part of the agent's belief system.

Neither can Pagnucco's explanations be interpreted as *potential* explanations of ϕ in K, that is, as sentences that would form part of the deductive explanation of ϕ if they were accepted in K. Potential explanations have important uses in science. Hempel mentions, for example, that a potential explanation can be used to examine "whether a novel and as yet untested law or theory would provide an explanation for some empirical phenomenon" (1965, p. 338). Harman (1965), on the other hand, argues that the explanatory power of a new theory is part of the evidence that leads us to accept it. But Pagnucco's explanations can be put to no such use. As Hempel's comment makes clear, in order to test the explanatory potential of a theory, one must have antecedently accepted the facts that would be explained by the theory. Pagnucco's abductions work the other way around: one must first accept one of the potential explanations of ϕ in order to accept ϕ. There is no independent way of assessing the explanatory value of ψ because its status as a potential explanation is tied to a sentence ϕ about whose truth the agent is still undecided.

Suppose that the agent expands her belief set by ϕ using the operation of abductive expansion. It seems to me that the explanation used to support the acceptance of ϕ cannot itself be accepted without defeating the whole purpose of Pagnucco's approach. The type of expansion operation that he defines is intended to provide support for every new sentence that is added to K. The problem is that, although ϕ in $K_{\phi}^{\oplus}$ is supported by ψ, the latter sentence is added to K in an abductive expansion without itself having any support other than the fact that it is part of the best explanation of ϕ. In other words, for every justified formula ϕ that the agent accepts, she must accept an unjustified formula ψ whose only credentials are that it is an element of the intersection of the highest ranking maximally consistent supersets of K that imply ϕ. In fact, the agent might not even know which of the potential explanations of ϕ is the one that ends up in her belief set, as Pagnucco himself acknowledges:

> In abductive expansion, as we have seen, the concern is to determine which beliefs should be incorporated into the current epistemic state using an abductive strategy to identify the appropriate expansion given new information. In so doing however, the process of abduction has become "internalized" in the belief expansion process and therefore, the actual abduction(s) made to effect a change in epistemic state is, in a sense, lost. That is, it may not be possible to determine the abduction selected for a belief set K and epistemic input ϕ in the sense of [the definition of abduction provided above] (i.e., it may not be possible to identify ψ) (p. 135).

Pagnucco argues that once ϕ has been accepted, it will be possible to find an abduction "capable of doing the job." The strategy is to examine the set $K_{\phi}^{\oplus}\backslash K_{\phi}^{+}$

and determine a finite axiomatization of it. "The conjunction of the elements of this finite axiomatization will suffice as an appropriate abduction" (p. 135). There might be many ways to finitely axiomatize the set, but Pagnucco argues that restrictions such as minimality and the like can be used to select the best abduction.

Unfortunately, this strategy does not answer my initial objection. Not only is the agent accepting an unexplained sentence ψ every time she accepts a sentence ϕ via an abductive expansion. If the agent has to examine $K^{\oplus}_{\phi}$ in order to find out which of the potential explanations of ϕ was used in the abductive expansion, it is absurd to say that the agent used that potential explanation to *justify* or support the expansion of his belief set into $K^{\oplus}_{\phi}$. Although Pagnucco wants to offer an account of expansion in which the agent "first seek[s] some explanation or justification for the epistemic input," the abductive expansion function that he defines does not reflect his declared intention.

Finally, the claim that the conjunction of the elements of the finite axiomatization of $K^{\oplus}_{\phi}\backslash K^{+}_{\phi}$ suffices as an appropriate explanation of ϕ can also be challenged. Pagnucco's account is vulnerable to the same counterexamples that have been raised against approaches, such as Hempel's, in which explanation is understood as an inferential relation. Consider Bromberger's well-known flagpole example. Let ϕ = 'The flagpole in front of the building is 10 m tall' and ψ = 'The length of the flagpole's shadow is 10 m and the elevation of the sun in the sky is 45°'. Suppose that the agent wants to add ϕ to his belief state. Using ψ and his mathematical knowledge, the agent can conclude that the flagpole is 10 m tall. Thus if we examine the conjunction of the elements of the finite axiomatization of $K^{\oplus}_{\phi}\backslash K^{+}_{\phi}$, we will find ψ. So ψ is an appropriate abduction of ϕ according to Pagnucco, but clearly it does not explain why the flagpole is 10 m tall. It does, however, justify the addition of ϕ to the agent's belief set.

The result of Pagnucco's abductive expansion functions are thus better understood as internalized justifications and not as explanations in the philosophical sense. Although his approach to explanation is logically impeccable, it does not provide an adequate framework for the formulation of an epistemologically motivated notion of explanation.

4 Gärdenfors on explanation

In *Knowledge in Flux* (1988), Gärdenfors offers a very different account of explanation using the AGM framework for belief revision. The result is a highly complex formal theory based on a very simple idea. Most of my objections will be directed against the latter, so we will not need to worry too much about the technical details of his approach.

Unlike Pagnucco, Gärdenfors takes it for granted that if an agent wants to explain the fact stated by ϕ, the agent already believes that ϕ is true. An explanation does not determine whether a person accepts ϕ or not. Nonetheless, Gärdenfors argues that an explanation does have an effect on the way we believe that ϕ: "it is quite clear that the fact that E may be more or less *surprising* or *unexpected*, and the principal effect of a successful explanans is that the surprise at E is *decreased*" (p. 167).

To make these notions more precise, Gärdenfors introduces a probabilistic model of epistemic states using a first-order language, instead of the propositional language used in AGM. An epistemic state K suitable for the formulation of Gärdenfors's theory consists of (i) a set W of possible worlds, (ii) for each world w a probability measure P_W defined over sets of individuals in w, and (iii) a belief function B that measures the probability of sets of possible worlds. The measure P_W together with B yield a second-order probability distribution of properties, and using this distribution Gärdenfors defines a first-order expected probability measure.

Gärdenfors's strategy is to take as a basis for the analysis, not the agent's present belief state K, but the state K_ϕ^- in which the explanandum sentence has been contracted from K. Once the explanandum sentence is deleted, it is possible to measure how surprising it would be to find out that it is true. "The surprise value of ϕ is inversely related to the degree of belief associated with ϕ in K_ϕ^-. The central criterion on explanations is that the explanans in a nontrivial way should decrease the surprise value of the explanandum" (p. 168). In other words, the explanans should make ϕ more believable in K_ϕ^-. These considerations lead to the following necessary conditions for explanation (p. 178):

(EXP) An explanation of a singular sentence ϕ relative to a state of belief K (where $\phi \in K$) consists of (i) a conjunction T of a finite set of probability sentences and (ii) a conjunction C of a finite set of singular sentences that satisfy the requirements that (iii) $B_\phi^-(\phi/T\&C) > B_\phi^-(\phi)$, where B_ϕ^- is the belief function in the state K_ϕ^-, and (iv) $B(T\&C) < 1$ (that is, $T\&C \notin K$).

The definition amounts to a counterfactual analysis of an epistemic context in which the agent accepts neither the explanans nor the explanandum. The purpose of introducing the additional context is to determine what information should be accepted in K in order to make the acceptance of ϕ in K less surprising than it would be in the absence of the additional information.

Gärdenfors argues that a consequence of his definition is that there will be degrees of explanation. "The more an explanation increases the belief value of the explanandum, the better it is" (p. 185). From condition (iii), Gärdenfors obtains a measure of the explanatory power of an explanans. The greater the difference between $B_\phi^-(\phi/T\&C)$ and $B_\phi^-(\phi)$, the greater the explanatory power of $T\&C$ relative to ϕ. Since (EXP) allows many different explanations of ϕ, Gärdenfors argues that the explanatory power of $T\&C$ can be used to determine which of all the possible explanations is the best one. Deductive explanations turn out to be the best because they increase the belief value of the explanandum to the maximum value.

Gärdenfors contrasts his approach to Hempel's view according to which the explanans of an inductive-statistical (I-S) explanation shows that the phenomenon described by the explanandum sentence was to be *expected*. The demand that the explanans should make the explanandum *less surprising* is an analogous but weaker claim which Gärdenfors believes is immune to the objections that have been raised against the I-S model. An examination of these objections will allow us to test the adequacy of Gärdenfors's definition.

In *Aspects of Scientific Explanation*, Hempel characterized statistical explanations as inductive inferences or arguments in which "the explanans confers upon the explanandum a more or less high degree of inductive support or of logical (inductive) probability" (1965, p. 385). The probability associated with the explanation determines the strength of our expectations. Since Hempel required that the probability associated with an I-S explanation be fairly close to 1, the explanandum of an inductive-statistical explanation will always be expected "with 'practical' certainty, or with very high likelihood" (p. 389).

In response to criticism by Richard Jeffrey, in the first German edition of *Aspects* Hempel (2001) gave up the high probability requirement, together with the claim that the explanans of an I-S explanation should show that the phenomenon described by the explanandum sentence was to be expected.

Jeffrey (1971) presented two objections to the requirement that the probability associated with an I-S explanation be fairly close to 1. In the first place, Jeffrey argued that it is erroneous to set limits to the concept of explanation. We cannot exclude the possibility of explaining any phenomenon, regardless of the probability associated with it. Peter Railton makes the same point: "Virtually impossible events may occur, and they deserve and can receive the same explanation as the merely improbable or the virtually certain" (1978, p. 213).[4]

Secondly, Jeffrey argued that when we try to explain why the explanandum of an I-S explanation did not obtain, we use the same statistical laws and the same initial conditions that we would use in explaining its occurrence. In Jeffrey's view, an explanation of either the occurrence or the nonoccurrence of a statistical phenomenon "consists of a statement that the process is a stochastic one, following such-and-such a law. . . . The knowledge that the process was random answers the question, 'Why?'—the answer is, 'By chance' " (p. 24). Understanding the outcomes of a stochastic process does not involve a justificatory argument as to why a given outcome obtained; it requires an adequate description of the process involved.

Despite Gärdenfors's allegations to the contrary, Jeffrey's objections can be reformulated in the context of his own theory. I will argue that Gärdenfors's account erroneously limits the scope of the notion of explanation, and that it leads to an excessive relativization of the concept. I will begin with the latter objection.

In his discussion of Scriven's famous paresis example, Gärdenfors argues that if an agent wants to know why, of all people, Nietzsche developed paresis, an acceptable explanation is that he suffered from syphilis and that there is a low but nonvanishing probability that a syphilitic patient will develop paresis. So far, Gärdenfors seems to agree with Jeffrey and Railton. However, if the agent wants to know why, of all syphilitics, Nietzsche developed paresis, Gärdenfors argues that there is no explanation. There is no further factor that would make the explanandum less surprising. A more accurate way of describing the situation would be to say that the agent already knows the explanation, namely, that there is a low but nonvanishing probability that a syphilitic patient will develop paresis, not that there is no explanation. Gärdenfors's account does not allow this formulation because as soon as the explanans becomes part of the agent's state of belief, it loses all its explanatory power. In Gärdenfors's

[4] This comment was actually a criticism of Jeffrey, who had argued that the I-S model was unobjectionable only in the "beautiful cases" in which the probability is so high "as to make no odds in any gamble or deliberation" (p. 27). Railton argued that the "beautiful cases" were logically identical to the rest.

view, a set of sentences is an explanation in some epistemic contexts, but not in others. In that regard, his account closely resembles van Fraassen's theory of explanation, as Gärdenfors himself acknowledges.[5] I have argued elsewhere that this extreme form of epistemic relativity is untenable.[6]

We now turn to Jeffrey's first objection. Gärdenfors argues that characterizing a successful explanans as one that increases our degree of belief in a proposition does not preclude the explanation of facts that are familiar or not surprising at all because when we ask for an explanation of a well-known fact, "we in a sense pretend that ϕ is surprising" (p. 167). How does one pretend that something is surprising? Contracting one's belief state by the explanandum sentence will not do. If the belief value of ϕ in K_{ϕ}^{-} is very close to 1, it is difficult to see why an agent would want an explanation of ϕ if an explanation is conceived as information that raises the credal probability of ϕ. What additional information could possibly make ϕ less surprising?

Perhaps Gärdenfors could argue that there is a certain threshold beyond which explanations are no longer required. But the following example should dispel that idea. Consider the case of an old sailor who wants to know why the tide rose today at noon. According to Gärdenfors, the sailor should engage in an exercise of hypothetical belief revision and contract his epistemic state by the belief that the tide rose today at noon. Because of his many years of experience observing the correlation between the regular ebb and flow of the tides and the position and phase of the moon, the credal probability that he assigns to the belief that the tide rose today at noon in the contracted state is very close to 1. That fact, however, cannot preclude him from wondering why the tide rose today at noon. Furthermore, an explanation in terms of Newton's law of gravitation will not change the sailor's near certainty that the tide rose today at noon. It will, however, help the sailor *understand* why the tide rose today at noon.

The problem is not limited to cases in which the initial degree of belief is very close to 1. Suppose a tourist visiting the Sahara is quite surprised to see a storm approaching since it is her belief that it never rains in the Sahara. If the tourist had been told beforehand that the barometer was falling, and that whenever the barometer is falling a storm is approaching, the tourist's credal probability that there would be a storm would have been very close to 1. And yet, the falling barometer does not explain why a storm is approaching. In fact, we can say that even though the tourist would no longer be surprised by the occurrence of the storm, she would still be *puzzled* in a sense that has nothing to do with the credal probability that she assigns to the occurrence of the storm. Her intellectual curiosity would not be satisfied, and she would demand an explanation. Her epistemic situation would be similar to that of the old sailor who wanted an explanation of the tides.

The source of the problem has long been identified by philosophers of science: "some regularities have explanatory power, while others constitute precisely the kinds of natural phenomena that demand explanation" (Salmon 1984, p. 121). The regularities captured by the probability sentences in Gärdenfors's definition will often have no explanatory value even if they make the explanandum completely unsurprising.

[5] "The theory of explanation that comes closest to the present one is van Fraassen's" (p. 170).

[6] Páez (2006), ch. 1.

Without further restrictions on the explanans, the account will always be vulnerable to such counterexamples.

Gärdenfors argues that one should choose the explanans $T\&C$ that has the highest degree of explanatory power, but this demand does not solve the problem. It is difficult to imagine a set of sentences that would raise the tourist's credal probability more than the information regarding the falling barometer. Sets of sentences that are maximally explanatory in Gärdenfors's sense will often be useful for prediction and practical deliberation, but they will fail to provide understanding.

A deeper problem with Gärdenfors's account has to do with the notion of surprise. If we assume that the prior degree of belief in a sentence can be used to measure how surprising it would be to find out that it is true, Gärdenfors's account will often be inapplicable because the initial credal probability will be indeterminate. What is my degree of belief that it rained today in the Azores? Or that the economy of Bolivia grew more than 1% in 1987? It would be a mistake to assign any value to my degree of belief because I have no elements whatsoever to judge the issue one way or the other. The rational attitude, it seems to me, is to suspend judgment.

Levi (1988) proposes a much more fruitful way to think of the notion of surprise: "The truth of h is not surprising relative to a body of information if and only if the acceptance of $\sim h$ via inductive inference from that information is not legitimate. The truth of h is to be expected relative to that body of information if and only if its inductive acceptance is legitimate" (p. 207). An inductive inference in Levi's sense is a deliberate expansion of the agent's state of belief that seeks the best trade off between the informational value obtained and the error incurred.

According to these definitions, it would not be surprising for me to find out that it rained today in the Azores, or that the economy of Bolivia grew more than 1% in 1987, because it would have been illegitimate for me to infer the opposite. But the truth of these sentences would not be expected either. If my discovery that a sentence is true is neither surprising nor to be expected, it is because I had not judged the issue one way or the other. But if no credal probability can be assigned to the explanandum in B_{ϕ}^{-}, then Gärdenfors's account of explanation will be inapplicable in the vast majority of cases. In general, if the agent has no prior information about the explanandum, it will be meaningless to say that the potential explanans should make the explanandum less surprising.[7]

Like Pagnucco's analysis of abductive expansions, Gärdenfors's theory of explanation is better understood as an account of justification. The similarities between the two accounts are not difficult to see. Both accounts begin with a belief state in which the explanandum is absent, either because it has not yet been accepted, or because it has been counterfactually deleted. In both cases, an explanation lends support to the acceptance of ϕ in K (or in K_{ϕ}^{-}). In Pagnucco's case, an explanation, together with K, entails ϕ, thereby leading to a new state $K_{\phi}^{\oplus}$; in Gärdenfors's case, an explanation raises the credal probability of ϕ in K_{ϕ}^{-}, thus making it more believable in K. Both

[7] This objection is not explicitly stated in Levi (1988), but it follows from the definitions quoted above. Levi (personal communication) agrees with the conclusion of my analysis.

Pagnucco and Gärdenfors thus identify justificatory or evidential information with explanatory information.

In order to offer an adequate account of explanation in AI, we must abandon the idea that *any* information that justifies or serves as evidence for the explanandum also explains it. This idea was introduced by Hempel and Oppenheim (1948) as part of their thesis that there is a logical symmetry between prediction and explanation, and although Hempel never explicitly rejected that part of the symmetry thesis, he later declared it "open to question" (1965, p. 367). Numerous counterexamples, together with the rejection of the high probability requirement for inductive-statistical explanations, suggest that it is simply false.

5 Desiderata for an AI theory of explanation

In this final section I will state some of the desiderata for a scientifically accurate and epistemologically informed theory of explanation in AI. Some of them stem from the problems discussed above, and others from an analysis of certain more fundamental characteristics common to most approaches in the AI literature.

The first challenge in the construction of an epistemologically motivated model for explanation is to restore the factivity condition, which states that both the explanans and the explanandum of a bona fide explanation must be true.[8] In an adequate model, the fact for which the agent wants to find an explanation should already be part of his belief system. On the other hand, the model should be able to accommodate cases in which the information sought as an explanation for an observation is already part of the agent's belief system, but he is not aware of it due his limitations in time, memory, or computational ability. This could be achieved, for example, using Levesque (1989) distinction between implicit and explicit belief, Wassermann (2000) distinction between active and passive beliefs, or any other formal distinction that leads to a more realistic notion of a belief set, one that takes into account the limitations of the actual agents that are being modelled.

Belief revision theories of explanation always assume that K is a logically consistent theory and forbid deliberate expansion into an inconsistent state. Gärdenfors, for example, states that an inconsistent theory is "epistemic hell" (1988, p. 51): given the explosive nature of standard logics, an inconsistent theory would entail any belief whatsoever. The problem is that current scientific theories do not always correspond to logically consistent theories. Anomalies and conflicting observations are a common trait of scientific practice, and despite them it is still possible to reason within the inconsistent theories that contain them. Furthermore, progress in science often involves the introduction of conceptual novelties that generate partial contradictions with our current scientific knowledge. The solution is to find a way to reason within inconsistency. Hans Rott (2001) and Hansson and Wassermann (2002) have independently developed the suggestion made by David Lewis (1982) that the best way to deal

[8] More precisely, the explanans-sentences and the explanandum-sentence must be true. If one holds, following Lewis (1986) and Woodward (2003), that the relata of the explanation relation are particulars, i.e., things or events, the claim amounts to saying that the agent believes that the things or events occurring in both the explanans and the explanandum position exist or occur.

with inconsistencies in a belief state is to quarantine the inconsistency from the rest of the beliefs in the belief state, and try to solve the problem locally. Parikh (1999) follows a different approach in dealing with the same problem. The model for representing belief structures that he proposes relies on a notion of partial language splitting that tolerates some amount of inconsistency while retaining classical logic. An optimal model for explanation could use some of these technical tools in order to allow the agent to implement strategies for abduction even when there are inconsistencies within his current knowledge.[9]

Finally, it is very important to recognize that we cannot talk of only one type of abductive expansion of a belief set. As Schurz (2007) argues convincingly, the kind of hypotheses or explanations that are sought in scientific research determine different abductive strategies. In consequence, an optimal model should be flexible enough to accommodate these different types of abductive strategies. In particular, it should not be tied to a single notion of explanation since different explanatory contexts require different types of explanations. The language used in the model should therefore have enough formal resources to express causal, functional, intentional, and probabilistic statements, and it should be possible to establish a distinction between individual events and general laws.

In brief, an epistemologically motivated, scientifically informed belief revision model for abduction must fulfill the following desiderata:

1. The model should be based on a more realistic notion of a belief state. It must be based on finite agents with limited time, memory, and deductive ability, and it must distinguish between their implicit and explicit beliefs.[10]
2. The model must allow cases in which both the explanans and the explanandum have been previously accepted in the belief system of the agent.
3. The model must allow the possibility of inconsistent theories, and it must find ways to isolate these inconsistencies within the belief system.
4. The model must have enough flexibility to accommodate different types of abductive strategies and it should not be tied to a single theory of explanation.
5. The model must have enough formal resources to express causal, functional, intentional, and probabilistic statements, and to distinguish between individual facts and general laws.

It is improbable that any model will be able to fulfill the five desiderata at the same time, but we can still hope for partial success.

Acknowledgements For useful comments and suggestions I would like to thank Arnold Koslow, Isaac Levi, Michael Levin, Rohit Parikh, Markus Werning, and especially Gerhard Schurz. Research for this paper was financed by the Colombian Institute for the Advancement of Science and Technology (COLCIENCIAS) and by the School of Social Sciences at Universidad de los Andes.

[9] To be sure, the ability to deal with inconsistency is a desirable feature not only for a theory of explanation in AI, but generally for any theory of belief revision.

[10] The topic of bounded reasoning has been extensively explored in philosophy by Cherniak (1986) and Harman (1986).

References

Achinstein, P. (1983). *The nature of explanation*. New York: Oxford University Press.

Alchourrón, C., Gärdenfors, P., & Mackinson, D. (1985). On the logic of theory change: Partial meet contraction functions and their associated revision functions. *Journal of Symbolic Logic, 50*, 510–530. doi:10.2307/2274239.

Aliseda, A. (2006). *Abductive reasoning*. Dordrecht: Springer.

Boutilier, C. (1994). Unifying default reasoning and belief revision in a modal framework. *Artificial Intelligence, 68*, 33–85. doi:10.1016/0004--3702(94)90095-7.

Boutilier, C., & Becher, V. (1995). Abduction as belief revision. *Artificial Intelligence, 77*, 43–94. doi:10.1016/0004-3702(94)00025-V.

Cherniak, C. (1986). *Minimal rationality*. Cambridge: MIT Press.

Gärdenfors, P. (1988). *Knowledge in flux. Modeling the dynamics of epistemic states*. Cambridge: MIT Press.

Gärdenfors, P., & Makinson, D. (1994). Nonmonotonic inference based on expectations. *Artificial Intelligence, 65*, 197–245. doi:10.1016/0004-3702(94)90017-5.

Hansson, S. O., & Wassermann, R. (2002). Local change. *Studia Logica, 70*, 49–76. doi:10.1023/A:1014654208944.

Harman, G. (1965). The inference to the best explanation. *The Philosophical Review, 74*, 88–95. doi:10.2307/2183532.

Harman, G. (1986). *Change in view*. Cambridge: MIT Press.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: The Free Press.

Hempel, C. G. (2001). Postscript 1976: More recent ideas on the problem of statistical explanation. In J. H. Fetzer (Ed.), *The philosophy of Carl G. Hempel. Studies in science, explanation, and rationality*. New York: Oxford University Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*, 135–175.

Jeffrey, R. (1971). Statistical explanation vs. statistical inference. In W. C. Salmon (Ed.), *Statistical explanation and statistical relevance*. Pittsburgh: Pittsburgh University Press.

Kakas, A. C., Kowalski, R. A., & Toni, F. (1993). Abductive logic programming. *Journal of Logic and Computation, 2*, 719–770. doi:10.1093/logcom/2.6.719.

Levesque, H. J. (1989). A knowledge-level account of abduction. In *Proceedings of the Eleventh International Joint Conference in Artificial Intelligence*. Los Altos: Morgan Kaufman.

Levi, I. (1988). Four themes in statistical explanation. In W. L. Harper & B. Skyrms (Eds.), *Causation in decision, belief change, and statistics*. Boston: Kluwer Academic Publishers.

Levi, I. (2004). *Mild contraction. Evaluating loss of information due to loss of belief*. New York: Oxford University Press.

Lewis, D. (1982). Logic for equivocators. *Nous, 16*, 431–441. doi:10.2307/2216219.

Lewis, D. (1986). Causal explanation. In *Philosophical papers* (Vol. II). New York: Oxford University Press.

Mackinson, D., & Gärdenfors, P. (1991). Relations between the logic of theory change and nonmonotonic Logic. In A. Fuhrmann & M. Morreau (Eds.), *The logic of theory change*. Berlin: Springer.

Ng, H. T., & Mooney, R. J. (1990). On the role of coherence in abductive explanation. In *Proceedings of the Eighth National Conference on Artificial Intelligence*. Los Altos: Morgan Kaufmann.

Páez, A. (2006). *Explanations in* K. *An analysis of explanation as a belief revision operation*. Oberhausen: Athena Verlag.

Pagnucco, M. (1996). The role of abductive reasoning within the process of belief revision. Dissertation, University of Sydney.

Parikh, R. (1999). Beliefs, belief revision, and splitting languages. In L. Moss, J. Ginzburg, & M. de Rijke (Eds.), *Logic, language, and computation*. Amsterdam: CSLI Publications.

Paul, G. (1993). Approaches to abductive reasoning: An overview. *Artificial Intelligence Review, 7*, 109–152. doi:10.1007/BF00849080.

Peng, Y., & Reggia, J. A. (1990). *Abductive inference models for diagnostic problem-solving*. Berlin: Springer.

Poole, D. (1989). Explanation and prediction: An architecture for default and abductive reasoning. *Computational Intelligence, 5*, 97–110. doi:10.1111/j.1467-8640.1989.tb00319.x.

Railton, P. (1978). A deductive-nomological model of probabilistic explanation. *Philosophy of Science, 45*, 206–226. doi:10.1086/288797.

Ram, A., & Leake, D. (1991). Evaluation of explanatory hypotheses. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*. Hillsdale: Erlbaum.

Rott, H. (2001). *Change, choice, and inference. A study of belief revision and nonmonotonic reasoning*. Oxford: Oxford University Press.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. C. (1989). *Four decades of scientific explanation*. Minneapolis: University of Minnesota Press.

Schurz, G. (2007). Patterns of abduction. *Synthese* (in press). doi:10.1007/s11229-007-9223-4.

Scriven, M. (1959). Truisms as the grounds for historical explanation. In P. Gardiner (Ed.), *Theories of history*. New York: The Free Press.

Stickel, M. E. (1991). A Prolog-like inference system for computing minimal-cost abductive explanation in natural language interpretation. *Annals of Mathematics and Artificial Intelligence, 4*, 89–106. doi:10.1007/BF01531174.

Thagard, P. R. (1989). Explanatory coherence. *The Behavioral and Brain Sciences, 12*, 435–502.

Wassermann, R. (2000). *Resource-bounded belief revision*. Amsterdam: Institute for Logic, Language and Computation Dissertation Series.

Wassermann, R., & Dias, W. (2001). Abductive expansion of belief bases. In *Proceedings of the IJCAI Workshop on Abductive Reasoning*, Seattle.

Woodward, J. (2003). *Making things happen. A theory of causal explanation*. New York: Oxford University Press.