Determining Explanations using Transmutations*

Mary-Anne Williams
Information Systems Group
Department of Management
University of Newcastle, NSW 2308
Australia

Maurice Pagnucco
Knowledge Systems Group
Basser Department of Computer Science
University of Sydney, NSW 2006
Australia

Norman Foo
Knowledge Systems Group
Basser Department of Computer Science
University of Sydney, NSW 2006
Australia

Brailey Sims
Banach Astronautics Group
Department of Mathematics
University of Newcastle, NSW 2308
Australia

Abstract

Intelligent Information systems do not usually possess complete information about the world with which they interact. The AGM paradigm has become one of the standard frameworks for modeling changes to repositories of information. Its principal constructions for change operators rely on some form of underlying preference relation. The process of changing such a preference relation is known as a transmutation. Spohn's conditionalization can be interpreted as a transmutation that imposes a relative minimal change. A transmutation based on an absolute minimal change is an adjustment.

In this paper we develop a notion of explanation using transmutations of information systems. Following Gärdenfors lead we recast Spohn's notion of *reason for* within the general setting of transmutations and extend this to characterize most plausible explanations. We also investigate the relationship between explanation based on abduction and Spohnian reasons based on adjustments. Finally, and rather surprisingly, we identify explicit conditions that characterize the various forms of explanations identified by Boutilier and Becher using Spohnian reasons.

1 Introduction

Information systems use information theoretic descriptions to capture their view of the world. These descriptions are generally incomplete and evolve over time. An intelligent information system requires a mechanism for self-modification when new information about the world is acquired.

The AGM paradigm [Alchourrón *et al.*, 1985; Gärdenfors, 1988; Gärdenfors and Makinson, 1988] has become a standard framework for modeling information change. In particular, it provides a mechanism for the revision and contraction of information. Within the AGM paradigm, the family of revision functions and the family of contraction functions are constrained by rationality postulates. The logical properties of a body of information are not strong enough to uniquely determine a revision or contraction function. Therefore, the principal constructions for these functions rely on some form of underlying preference relation, such as a family of selection functions [Alchourrón *et al.*, 1985], a system of spheres [Grove, 1988], a nice preorder on models [Katsuno and Mendelzon, 1992; Peppas and Williams, 1995], or an epistemic entrenchment ordering [Gärdenfors and Makinson, 1988]. In this light we consider an information system to be composed of a set of information together with a preference relation.

We refer to the process of changing the underlying preference relation of an information system as a transmutation. Based on observations made by Gärdenfors [1988, p72], Williams [1994a] generalized the conditionalizations of Spohn [1988] to more general transmutations. Spohn represents an information system using ordinal conditional functions, which map possible worlds to ordinals, and argues that *conditionalization*, which is a constructive method for modifying an ordinal conditional function, is a reasonable procedure. Williams explored conditionalization further and examined an alternative construction; an *adjustment*. Both conditionalizations and adjustments are transmutations. Intuitively, conditionalization involves a *relative minimal change* necessary to effect the revision or contraction, whilst adjustment involves an *absolute minimal change*.

Spohn characterizes the notion of *reason for* in [1983] and shows how it can be captured within his framework in [1983; 1988].

The objective of this paper is, firstly, to recast Spohn's notion of *reason for* in a more general setting using transmutations rather than conditionalizations, and secondly, to extend this approach to determine most plausible explanations. Some preliminary work in this direction using a theory base representation can be found in [Williams, 1994b]. We also provide an intuitively appealing and transparent relationship between explanation based on abduction and Spohnian reasons based on adjustments. Finally and quite surprisingly, we find explicit conditions which capture the relationship between Spohnian reasons and the various forms of explanation identified by Boutilier and Becher [1995].

In section 2, we describe entrenchment rankings, while in section 3 we define transmutations of entrenchment rankings, and describe a particular type of transmutation, namely an adjustment. In section 4, we outline Spohn's notion of *reason for*, and provide conditions which allow this notion to be defined using both arbitrary transmutations and adjustments. In the spirit of Spohn we provide a notion of most plausible explanation. In section 5 we will illustrate some of our definitions and results with examples. In section 6 we compare our notion of explanation with one based upon the process of abductive inference. We discuss related work in section

* To appear in the Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, Canada. Morgan Kaufmann, 1995.

7, and summarise our position in section 8.

We begin with some technical preliminaries. Let $\mathcal{L}$ denote a language which contains a complete set of Boolean connectives. We will denote sentences in $\mathcal{L}$ by lower case Greek letters. We assume $\mathcal{L}$ is governed by a logic that is identified with its consequence relation $\vdash$ which is assumed to satisfy the following conditions [Gärdenfors, 1988]: (a) if α is a truth-functional tautology, then $\vdash \alpha$, (b) if $\vdash \alpha \rightarrow \beta$ and $\vdash \alpha$, then $\vdash \beta$ (*modus ponens*), (c) $\vdash$ is consistent, that is, $\nvdash \perp$, where $\perp$ denotes the inconsistent theory, (d) $\vdash$ satisfies the deduction theorem, and (e) $\vdash$ is compact.

The set of all logical consequences of a set $T \subseteq \mathcal{L}$, that is $\{\alpha : T \vdash \alpha\}$, is denoted by $\mathrm{Cn}(T)$. A *theory* of $\mathcal{L}$ is any subset of $\mathcal{L}$ closed under Cn. We let $\mathcal{L}^{\bowtie}$ denote the set of all consistent nontautological sentences in $\mathcal{L}$. That is, the set of *contingent* sentences. Finally, a *well-ranked* preorder on a set Γ is a preorder such that every nonempty subset of Γ has a minimal member.

2 Entrenchment Rankings

Spohn [1988] introduced ordinal conditional functions, while Williams [1994a] introduced related structures, namely entrenchment rankings (defined below). Essentially, ordinal conditional functions take possible worlds to ordinals, whereas entrenchment rankings take sentences to ordinals. Both define a ranking of the respective domains which provides a response schema for all "possible consistent information" [Spohn, 1988].

The higher the ordinal assigned to a sentence by an entrenchment ranking the more firmly held it is. Throughout the remainder of this paper it will be understood that $\mathcal{O}$ is an ordinal chosen to be sufficiently large for the purpose of the discussion.

Definition: *An* entrenchment ranking *is a function* $\mathbf{E}$ *from the sentences in* $\mathcal{L}$ *into the class of ordinals such that the following conditions are satisfied.*

(ER1) *For all* $\alpha, \beta \in \mathcal{L}$, *if* $\alpha \vdash \beta$, *then* $\mathbf{E}(\alpha) \leq \mathbf{E}(\beta)$.
(ER2) *For all* $\alpha, \beta \in \mathcal{L}$, $\mathbf{E}(\alpha) \leq \mathbf{E}(\alpha \wedge \beta)$ *or* $\mathbf{E}(\beta) \leq \mathbf{E}(\alpha \wedge \beta)$.
(ER3) $\vdash \alpha$ *iff* $\mathbf{E}(\alpha) = \mathcal{O}$.
(ER4) *If* α *is inconsistent, then* $\mathbf{E}(\alpha) = 0$.

An entrenchment ranking represents a *well-ranked* epistemic entrenchment ordering [Gärdenfors, 1988; Gärdenfors and Makinson, 1988] where the higher the ordinal assigned to a sentence the more entrenched it is. The ordinal assignment can be used and viewed in two ways: (i) qualitatively, that is as a specification of a relative ordering of sentences, or (ii) quantitatively, that is as an assignment of intrinsic utility for each sentence. For a discussion of various assignment scales see [Gärdenfors and Makinson, 1994].

Entrenchment rankings are information systems; a set of information together with a (necessarily well-ordered) preference relation. We denote the family of all entrenchment rankings by $\mathcal{E}$. An information set represented by an entrenchment ranking is the set of sentences whose degree of acceptance is greater than zero. Formally we have the following definition [Williams, 1994a].

Definition: *We define the* information set *represented by* $\mathbf{E} \in \mathcal{E}$ *to be* $\{\alpha \in \mathcal{L} : \mathbf{E}(\alpha) > 0\}$, *and denote it by* $\mathrm{set}(\mathbf{E})$.

We refer to information whose degree of acceptance is zero as the non-beliefs, and information with a degree of acceptance greater than zero as the beliefs. We note that $\mathrm{set}(\mathbf{E})$ is a theory, and hence the set of beliefs are closed with respect to logical consequence.

3 Transmutations

Peppas [1993] introduced *well-behaved* revision functions and provided a construction using a well-ordered system of spheres [Grove, 1988]. Williams [1994a] gives representation theorems for entrenchment rankings and well-behaved change operators. In particular, she provided conditions which characterize both well-behaved contraction and well-behaved revision operators using an entrenchment ranking, and conversely. An extra postulate is required for the change operator to be constructed from a well-ranked preference relation.

The informational input for AGM contraction and revision functions for theories is a sentence alone. In contrast, when modifying an entrenchment ranking we require both a sentence and an ordinal. Consequently, the informational input for transmutations is a sentence α and an ordinal i. The interpretation being [Gärdenfors, 1988] that α represents the information to be accepted by the information system, and i is the degree of firmness with which this information is to be incorporated into the transmuted information system. We define a transmutation schema [Williams, 1994a] for entrenchment rankings below.

Definition: *We define a* transmutation schema *for entrenchment rankings,* *, *to be a function from* $\mathcal{E} \times \mathcal{L}^{\bowtie} \times \mathcal{O}$ *to* $\mathcal{E}$, *where* $\mathcal{O}$ *is an ordinal, such that* $(\mathbf{E}, \alpha, i) \mapsto \mathbf{E}^\star(\alpha, i)$ *satisfies:*

(i) $\mathbf{E}^\star(\alpha,i)(\alpha) = i$, *and (ii)* $\mathrm{set}(\mathbf{E}^\star(\alpha,i))$

$$= \begin{cases} \{\beta \in \mathcal{L} : \mathbf{E}(\neg\alpha) < \mathbf{E}(\neg\alpha \vee \beta)\} & \text{if } i > 0 \\ \{\beta \in \mathrm{set}(\mathbf{E}) : \mathbf{E}(\alpha) < \mathbf{E}(\alpha \vee \beta)\} & \text{otherwise.} \end{cases}$$

For a contingent sentence α and an ordinal i, we say $\mathbf{E}^\star(\alpha,i)$ is an (α,i)–transmutation of $\mathbf{E}$. It was shown in [Williams, 1994a], that according to the definition above, $\mathbf{E}^\star(\alpha,i)$ is an entrenchment ranking where α is assigned the degree of acceptance i, and if i is greater than zero then $\mathrm{set}(\mathbf{E}^\star(\alpha,i))$ is the well-behaved revision $(\mathrm{set}(\mathbf{E}))^*_\alpha$ constructed in the obvious way from the epistemic entrenchment ordering derived from the relative ordering given by $\mathbf{E}$. Similarly, if i is zero then $\mathrm{set}(\mathbf{E}^\star(\alpha,i))$ is the well-behaved contraction $(\mathrm{set}(\mathbf{E}))^-_\alpha$.

Adjustments (defined below) were introduced in [Williams, 1994a]. They are transmutations which adopt an absolute measure under the the principle of minimal change. In particular, an entrenchment ranking is changed or disturbed, in an *absolute* sense, as little as necessary to give α the degree of acceptance i. That is, each sentence is reassigned an ordinal as close as possible to its previous ordinal assignment such that the resultant structure is an entrenchment ranking. An adjustment is based on intuition similar to that used by Boutilier [1993] to specify his *natural revision*; namely the principle of minimal change. According to Boutilier "when certain beliefs must be given up, it seems natural to try to keep not only important beliefs, but as much of the ordering as possible".

$$\mathbf{E}^\star(\alpha,i) = \begin{cases} (\mathbf{E}^-(\alpha,i)) & \text{if } i < \mathbf{E}(\alpha) \\ (\mathbf{E}^-(\neg\alpha,0)^+(\alpha,i) & \text{otherwise} \end{cases}$$

where

$$\mathbf{E}^-(\alpha,i)(\beta) = \begin{cases} i & \text{if } \mathbf{E}(\alpha) = \mathbf{E}(\alpha \vee \beta) \text{ and } \mathbf{E}(\beta) > i \\ \mathbf{E}(\beta) & \text{otherwise} \end{cases}$$

$$\mathbf{E}^+(\alpha, i)(\beta) = \begin{cases} \mathbf{E}(\beta) & \text{if } \mathbf{E}(\beta) > i \\ i & \text{if } \mathbf{E}(\beta) \le i < \mathbf{E}(\neg\alpha \vee \beta) \\ \mathbf{E}(\neg\alpha \vee \beta) & \text{otherwise} \end{cases}$$

4 Explanation

Spohn [1983] gave the following interpretation of α being a reason for β:

($\mathcal{R}1$) α is a *reason for* β iff raising the epistemic rank of α would raise the epistemic rank of β.

Gärdenfors [1990] translated Spohnian reasons into the AGM paradigm. However, in contrast to the translation we give later based on transmutations, his translation based on revision and contraction alone is somewhat artificial since the epistemic rank after a revision or contraction is undetermined. Essentially, his interpretation is that α is a reason for β if and only if β is not retained in the contraction of α. In the same paper Gärdenfors identifies two problems with this translation. Firstly he demonstrates that difficulties with the circularity of reasons arise, and secondly that multiple reasons for β cannot be modeled satisfactorily. These problems are discussed further by Olsson [1994] where an alternative solution to ours can be found, based on a comparative measure of epistemic entrenchment orderings for different theories. We note that the solution we present based on transmutations does not suffer from either of these problems.

Before recasting ($\mathcal{R}1$) using transmutations, some remarks are in order. Firstly, the determination of this condition is dependent on the type of transmutation employed since it will determine changes in epistemic rank for all sentences. We contend that an adjustment is appropriate for determining *reasons* on the basis that it is a transmutation which performs an *absolute minimal change*. In particular, every sentence is reassigned a new ordinal as close to its previous assignment as is consistent with the desired change. When the epistemic rank of α is increased it seems reasonable to require that this change should disturb the background information system as little as possible.

Secondly, we wish to distingish two types of reasons: (i) ordinary reasons, for which there *exists* an ordinal to which α can be raised, leading to an increase in the degree of β, and (ii) strong reasons, for which raising the degree of α by *any* amount leads to an increase in the degree of β.

Following Gärdenfors' lead, we can rewrite ($\mathcal{R}1$) using the notion of transmutation, and the degree of acceptance to capture both types of reasons:

($\mathcal{R}2$) Given an entrenchment ranking $\mathbf{E} \in \mathcal{E}$, α is a *reason for* β iff $\mathbf{E}^\star(\alpha, i)(\beta) > \mathbf{E}(\beta)$ for some $i > \mathbf{E}(\alpha)$.

($\mathcal{R}3$) Given an entrenchment ranking $\mathbf{E} \in \mathcal{E}$, α is a *strong reason for* β iff $\mathbf{E}^\star(\alpha, i)(\beta) > \mathbf{E}(\beta)$ for all $i > \mathbf{E}(\alpha)$.

Theorems 1 and 3, below, provide conditions that describe both reasons and strong reasons when the transmutation employed is an adjustment.

Theorem 1: *Let $\mathbf{E} \in \mathcal{E}$. Then α is a reason for β determined by an adjustment iff (i) $\mathbf{E}(\beta) < \mathbf{E}(\neg\alpha \vee \beta)$, and (ii) $\mathbf{E}(\neg\alpha) < \mathbf{E}(\neg\alpha \vee \beta)$.*

Corollary 2: *Let $\mathbf{E} \in \mathcal{E}$. Then α is a reason for β determined by an adjustment iff (i) $\neg\alpha \in (\mathrm{set}(\mathbf{E}))^*_{\neg\beta}$, and (ii) $\beta \in (\mathrm{set}(\mathbf{E}))^*_\alpha$ where * is the revision function uniquely determined by Gärdenfors and Makinson's [1988] construction based on the relative ordering of sentences given by $\mathbf{E}$.*

Theorem 3: *Let $\mathbf{E} \in \mathcal{E}$. If α is a reason for β determined by an adjustment, then $\mathbf{E}(\alpha) \le \mathbf{E}(\beta)$.*

Theorem 4: *Let $\mathbf{E} \in \mathcal{E}$. Then α is a strong reason for β determined by an adjustment iff (i) $\mathbf{E}(\beta) \le \mathbf{E}(\alpha) < \mathbf{E}(\neg\alpha \vee \beta)$, and (ii) $\mathbf{E}(\neg\alpha) < \mathbf{E}(\neg\alpha \vee \beta)$.*

Corollary 5: *Let $\mathbf{E} \in \mathcal{E}$. Let * be an adjustment. Then α is a strong reason for β iff α is a reason for β and $\mathbf{E}(\alpha) = \mathbf{E}(\beta)$.*

We say an information system given by $\mathbf{E}$ is in a state of *coherence* if whenever α is a *reason for* β then $\mathbf{E}(\alpha) \le \mathbf{E}(\beta)$. That is, the degree of acceptance of β is at least as high as the degree of acceptance of α. The justification for this can be made along similar lines to that of (ER1) [Gärdenfors 1988, p89]. Intuitively the explanandum should be at least as plausible as the explanation. Consequently, if β is an observation not yet accepted then the only reasons for it can be non-beliefs when the background information is represented by a coherent entrenchment ranking. In practical terms this suggests that in the case where $\mathbf{E}(\alpha) > \mathbf{E}(\beta)$ and α is supposed to be a reason for β then not only is the entrenchment ranking incoherent but the agent must *reorganise* [Gärdenfors 1988; Williams, 1994a] its information by transmuting it to a coherent state, and then use the resultant entrenchment ranking to determine reasons. Theorem 3 shows that adjustments respect this notion of *coherence*.

An appealing property of adjustments as the transmutation to specify reasons, apart from its adherence to the principle of minimal change, is that only the relative ordering given by the background information system itself is needed. In particular, only the relative ranking of *three* sentences need ever be compared!

We say an explanation is a nontrivial reason. That is, α is an explanation for β iff α is a reason for β and not logically equivalent to it.

In the spirit of Spohn we define a most plausible explanation to be an explanation which is capable of increasing the degree of acceptance of the explanandum as much as any other explanation during a transmutation.

Definition: *Let $\mathbf{E} \in \mathcal{E}$. Let A be a set of explanations for a sentence β. Define $\alpha \in A$ to be a most plausible (strong) explanation for β if $\mathbf{E}^\star(\gamma, i)(\beta) \le \mathbf{E}^\star(\alpha, i)(\beta)$ for all $\gamma \in A$, and for all ordinals $i > \mathbf{E}(\alpha)$.*

Theorem 6: *Let * be an adjustment and used for the determination of reasons. Let A be a set of (strong) explanations for a sentence β. Then α is a most plausible (strong) explanation in A for β iff $\mathbf{E}(\gamma \to \beta) \le \mathbf{E}(\alpha \to \beta)$ for all $\gamma \in A$.*

In section 7 we contrast this notion of plausibility with that of Boutilier and Becher [1995].

5 Examples

α: *wet grass*	δ: *it is raining*
β: *water main is broken*	η: *grass is under cover*
γ: *sprinkler is on*	ϕ: $E = mc^2$

The proposition α represents an observation while the other propositions constitute possible explanations for this observation.

`Example 1:` Let $\mathbf{E} \in \mathcal{E}$ be such that $0 = \mathbf{E}(\beta)$, $\mathbf{E}(\eta)$, $\mathbf{E}(\neg\alpha)$, $\mathbf{E}(\gamma), \mathbf{E}(\neg\delta) < \mathbf{E}(\neg\delta \vee \eta)$, $\mathbf{E}(\neg\eta), \mathbf{E}(\neg\gamma)$, $\mathbf{E}(\alpha), \mathbf{E}(\delta)$, $\mathbf{E}(\eta \to \alpha), \mathbf{E}(\phi \to \alpha)$, $\mathbf{E}(\delta \wedge \eta \to \alpha)$, $\mathbf{E}(\neg\beta)$ $< \mathbf{E}(\gamma \to \alpha) < \mathbf{E}(\beta \to \alpha) < \mathbf{E}(\delta \wedge \neg\eta \to \alpha) < \mathbf{E}(\phi)$

The grass is uncovered, the sprinkler is off, and it is raining.
η and ϕ are not reasons for α.
α, β, γ, $\delta \wedge \neg\eta$ are reasons for α.
α and $\delta \wedge \neg\eta$ are both strong reasons for α.
β is more plausible than γ as an explanation for α.
$\delta \wedge \neg\eta$ is the most plausible (strong) explanation for α.

`Example 2:` Let $\mathbf{E} \in \mathcal{E}$ be such that $0 = \mathbf{E}(\beta)$, $\mathbf{E}(\gamma)$, $\mathbf{E}(\neg\alpha)$, $\mathbf{E}(\neg\beta)$, $\mathbf{E}(\neg\gamma)$, $\mathbf{E}(\neg\delta)$, $\mathbf{E}(\delta)$, $\mathbf{E}(\neg\eta)$, $\mathbf{E}(\eta)$ $\mathbf{E}(\neg\delta \vee \eta) < \mathbf{E}(\alpha)$, $\mathbf{E}(\eta \rightarrow \alpha)$, $\mathbf{E}(\delta \wedge \eta \rightarrow \alpha) < \mathbf{E}(\gamma \rightarrow \alpha)$ $< \mathbf{E}(\beta \rightarrow \alpha) < \mathbf{E}(\delta \wedge \neg\eta \rightarrow \alpha) < \mathbf{E}(\phi)$

It is not known whether the grass is covered, whether the sprinkler is on, or whether it is raining.
η and ϕ are not reasons for α.
α, β, γ, $\delta \wedge \neg\eta$ are reasons for α.
α is the only strong reason for α.
$\delta \wedge \neg\eta$ is a more plausible explanation for α than β or γ.

6 Abductive Reasoning

One method currently gaining popularity for defining the notion of explanation is that of abductive reasoning [Paul, 1993; Stickel, 1991]. Abduction is a form of logical inference that aims to derive plausible explanations for information and is, in fact, often described as inference to the best explanation.

An abductive inference proceeds by proposing or generating hypotheses which would account for, or explain a group of facts. A common way of defining explanation as an abductive inference is the following (see [Paul, 1993], for example — however, note that we do not adopt his restriction to abducibles).

Definition: *An* abductive explanation *of a sentence β with respect to a set of sentences T is a sentence α such that (i) $T \cup \{\alpha\} \vdash \beta$, and (ii) $T \cup \{\alpha\} \not\vdash \bot$.*

Part (i) of the definition corresponds to having the sentence β accepted whenever the theory is expanded with α, whilst (ii) restricts α to being consistent with T. In other words, α is an abductive explanation for β if $\beta \in T^+_\alpha$, and $\neg\alpha \notin T^+_\alpha$.

Within the framework of transmutations, our domain theory is obtained from a given entrenchment ranking using proper cuts, defined below.

Definition: *Given an $\mathbf{E} \in \mathcal{E}$. We define a* proper cut *of $\mathbf{E}$ with respect to a sentence α, denoted by $\mathrm{cut}_<(\mathbf{E}, \alpha)$, to be the set of sentences $\{\beta \in \mathcal{L} : \ \mathbf{E}(\alpha) < \mathbf{E}(\beta)\}$.*

It is not hard to see that all proper cuts are theories, and that $\mathrm{set}(\mathbf{E}) = \mathrm{cut}_<(\mathbf{E}, \alpha \wedge \neg\alpha)$ that is, the largest proper cut is precisely the accepted information.

The idea of a proper cut has proven to be an expedient concept which has been used for several purposes elsewhere. For example, Rott [1991] uses them to describe theory contraction and revision, Williams [1994b] uses them to define change operators for theory bases, and Gärdenfors and Makinson [1994] use them to specify their nonmonotonic inference relation based on expectation orderings. More particularly, if we allow relative orderings of sentences given by $\mathbf{E}$ to determine the epistemic entrenchment ordering then, using the construction of Gärdenfors and Makinson [1988], we have $(\mathrm{set}(\mathbf{E}))^*_\alpha = (\mathrm{cut}_<(\mathbf{E}, \neg\alpha))^+_\alpha$ and $(\mathrm{set}(\mathbf{E}))^-_\alpha = (\mathrm{cut}_<(\mathbf{E}, \alpha))^+_{\neg\alpha} \cap \mathrm{set}(\mathbf{E})$.

The following theorem shows that abductive explanation is closely related to Spohnian reasons.

Theorem 7: *Let $\mathbf{E}$ be an entrenchment ranking and α, $\beta \in \mathcal{L}^{\bowtie}$. Let * be an adjustment and used for the determination of reasons. Then α is a reason for β iff α is an abductive explanation of β with respect to the theory $\mathrm{cut}_<(\mathbf{E}, \beta) \cap \mathrm{cut}_<(\mathbf{E}, \neg\alpha)$.*

Theorem 7 says, α is a reason for β if and only if α is an abductive explanation of β with respect to the smaller of the two proper cuts, $\mathrm{cut}_<(\mathbf{E}, \beta)$ or $\mathrm{cut}_<(\mathbf{E}, \neg\alpha)$. The former being the set of sentences strictly greater than β itself, and the latter being the largest proper cut which is consistent with α.

Corollary 8: *Let $\mathbf{E}$ be an entrenchment ranking and α, $\beta \in \mathcal{L}^{\bowtie}$. Let * be an adjustment and used for the determination of reasons. Then α is a reason for β iff $\beta \in (\mathrm{cut}_<(\mathbf{E}, \beta))^*_\alpha$.*

Moreover, depending on the relative status of the sentences determined by $\mathbf{E}$ the reason may be a strong one. In particular, if $\mathrm{cut}_<(\mathbf{E}, \alpha) = \mathrm{cut}_<(\mathbf{E}, \beta)$ then the abductive explanation described in Theorem 7 is a strong reason since this condition will hold iff $\mathbf{E}(\beta) = \mathbf{E}(\alpha)$.

The definition of explanation, through the process of abductive inference, will in general classify many sentences as explanations. Often we are interested in determining a single *best* explanation or, at the very least, reducing the number of possible explanations. This is usually done by adopting some form of preference criteria (e.g., cost-based measures, probabilities, etc. — see [Paul, 1993] for an overview).

In the framework discussed here we are equipped with a preference ordering over the sentences; namely, an entrenchment ranking which assigns a degree of acceptance to each sentence. Given this interpretation of entrenchment rankings it would seem natural to adopt this ordering to characterize best abductive explanations. Formally we have the following theorem.

Theorem 9: *Let $\mathbf{E}$ be an entrenchment ranking and α, $\beta \in \mathcal{L}^{\bowtie}$. Let * be an adjustment and used for the determination of reasons. Let A be a set of (strong) explanations for a sentence β. Then α is a most plausible (strong) explanation in A for β iff α is an abductive explanation of β with respect to $\mathrm{cut}_<(\mathbf{E}, \beta) \cap \mathrm{cut}_<(\mathbf{E}, \neg\alpha)$, and there is no abductive explanation γ in A of β with respect to $\mathrm{cut}_<(\mathbf{E}, \beta) \cap \mathrm{cut}_<(\mathbf{E}, \neg\gamma)$ for any $\gamma \in A$ such that $\mathbf{E}(\alpha \rightarrow \beta) < \mathbf{E}(\gamma \rightarrow \beta)$.*

Another important consideration when determining abductive explanations is the *specificity* of the explanation. The notion of most (and least) specific abduction was discussed by Stickel [1991]. It is described within a resolution-based system for performing abduction. A most specific abduction only allows pure literals (literals that resolve with no other clause in the information set) to be used in an abduction. On the other hand, a least specific abduction for a sentence is just the sentence itself. Stickel suggests that most specific abductions are suited to diagnostic tasks while least specific abductions are more appropriate for natural language interpretation. The specificity of an abductive explanation is given below:

Definition: *An abductive explanation α of β with respect to a theory T is more* specific *than an abductive explanation γ of β with respect to T iff $T \cup \{\alpha\} \vdash \gamma$.*

Equating the theory T of the definition above with $\mathrm{cut}_<(\mathbf{E}, \beta) \cap \mathrm{cut}_<(\mathbf{E}, \neg\alpha)$ motivates the following definition of specificity for reasons based on adjustments.

Definition: *Let $\mathbf{E} \in \mathcal{E}$. Let * be an adjustment and used for the determination of reasons. Let A be a set of (strong) explanations for a sentence β. Then $\alpha \in A$ is a* more specific (strong) explanation *for β than $\gamma \in A$ iff α is a (strong) reason for γ.*

Theorems 6 and 9 allow us to say that if α and γ are both explanations for β and α is also an explanation for γ then the more specific explanation α is more plausible if $\mathbf{E}(\gamma \rightarrow \beta) \leq \mathbf{E}(\alpha \rightarrow \beta)$. This can be generalised for longer *chains* of explanation to yield a method for determining the preferred level of specificity.

7 Related Work

Recently Gardenfors [1988] as well as Boutilier and Becher [1995] have given related accounts of explanation. Gärdenfors uses changes in probability, whilst we use changes in plausibility. The relationship between Boutilier and Becher's work and that presented in this paper is somewhat unexpected because both approaches draw their intuitions from seemingly different directions, indeed opposite directions. In particular, our analysis focuses on changes at the preference relation level whilst Boutilier and Becher's analysis considers the behaviour at the informational content level. However their underlying similarity appears to stem from the committment to the principle of minimal change.

Boutilier and Becher use an explicit assumption that explanations for beliefs are beliefs, and explanations for non-beliefs are non-beliefs. By way of justification they say that observations in their framework are not accepted into an information set until some explanation is found and accepted. Our idea of coherence is related to this assumption. For instance, according to it the only explanations for non-beliefs are non-beliefs. However, it does allow beliefs to have non-beliefs as explanations, a trait which is more in keeping with the usual interpretation of abductive explanations. Using Corollary 2 and the connections below one can show that Boutilier and Becher's explanations respect our idea of coherence, and therefore their assumption is stronger than that of coherence due to the restriction that only beliefs can be explanations for beliefs.

Boutilier and Becher's definition of α *explains* β, although restricted to the propositional case and based on the assumption above, is closely related to α is a *reason for* β based on *adjustments*. In particular, if their explanations are based on a well-ranked CO revision model structure [Boutilier, 1993] then we have:

α is a *predictive explanation* for β iff
 (i) α is a reason for β, (ii) $\alpha \in \text{set}(\mathbf{E})$ iff $\beta \in \text{set}(\mathbf{E})$, and (iii) $\neg\alpha \in \text{set}(\mathbf{E})$ iff $\neg\beta \in \text{set}(\mathbf{E})$.

α is a *hypothetical explanation* for β iff
 (i) α is a reason for β, (ii) $\alpha, \beta \notin \text{set}(\mathbf{E})$, and
 (iii) $\neg\alpha \in \text{set}(\mathbf{E})$ iff $\neg\beta \in \text{set}(\mathbf{E})$.

α is a *factual explanation* for β iff
 (i) α is a reason for β, (ii) $\alpha, \beta \in \text{set}(\mathbf{E})$, and
 (iii) $\neg\alpha \in \text{set}(\mathbf{E})$ iff $\neg\beta \in \text{set}(\mathbf{E})$.

α is a *counterfactual explanation* for β iff
 (i) α is a reason for β, and
 (ii) if $\beta \in \text{set}(\mathbf{E})$ then $\alpha \in \text{set}(\mathbf{E})$.

We note, hypothetical explanations are strong reasons since $\mathbf{E}(\alpha) = \mathbf{E}(\beta) = 0$.

In contradistinction to reasons based on adjustments, Boutilier and Becher's explanations preclude non-beliefs being explanations for accepted beliefs. In particular, according to Boutilier and Becher if $\alpha \notin \text{set}(\mathbf{E})$ and $\beta \in \text{set}(\mathbf{E})$ then α cannot explain β, but on our account α could be a reason for β. Consequently, the notion of reason for based on adjustment is more general.

Clearly, the connections described above may not hold if transmutations other than adjustments are used to determine reasons. In this sense, Spohnian reasons specified by ($\mathcal{R}2$) are far more general than the explanations of Boutilier and Becher, because their notion of explanation essentially corresponds to reasons based on a particular transmutation; namely, an adjustment.

Boutilier and Becher argue that *preferred explanations* are those that are most plausible; they require the *least change* in the information set in order to be accepted. In particular, they maintain as much information as possible, including as much default information as possible.

Their notion of plausibility is very different to the notion of plausibility advocated in this paper, where we claim that the most plausible explanation is one that is capable of increasing the degree of acceptance of the explanandum the most. For them, if α and γ are both beliefs and explanations for another belief β then they are equally preferred. In other words, within Boutilier and Becher's framework factual explanations can not be ranked on the basis of plausibility. Consequently, our notion of plausibility is substantially more discerning.

Boutilier and Becher highlight the desirable defeasible character of their explanations. Spohnian reasons based on adjustments also exhibit this feature, and moreover they give rise to consistency preserving rational consequence relations. In particular, if we define $\alpha \mid\!\sim \beta$ to be α is a reason for β based on an adjustment, then $\mid\!\sim$ is consistency preserving and rational. Consequently, there exists a nice preference structure [Gärdenfors and Makinson, 1994] and an expectation ordering which induces reasons based on adjustments.

Given the transparent connection between entrenchment rankings and expectation orderings [Williams, 1995b] we can give the following intuitive interpretation of *reason for* in terms of expectations. We can say that α is a reason for β if $\neg\alpha \vee \beta$ is strictly more expected than either $\neg\alpha$ or β, and that α is a strong reason for β if in addition α and β are equally expected. Similar readings can be given for plausibility and specificity.

Williams [1995a] gives a *computational model* for adjustments which can be used to implement Spohnian reasons, abductive, predictive, factual, hypothetical and counterfactual explanations.

Spohn [1988] revised his notion of reason for, and gave a condition which is equivalent [Williams, 1994a] to the condition below (it is related to conditionalization).

α is a *reason for* β iff
 (i) $-\mathbf{E}(\neg\alpha) + \mathbf{E}(\neg\alpha \vee \beta) > -\mathbf{E}(\alpha) + \mathbf{E}(\alpha \vee \beta)$, or
 (ii) $-\mathbf{E}(\alpha) + \mathbf{E}(\alpha \vee \neg\beta) > -\mathbf{E}(\neg\alpha) + \mathbf{E}(\neg\alpha \vee \neg\beta)$.

Intuitively, this condition says that α is a reason for β iff β has a higher degree of acceptance, or degree of plausibility given α rather than given $\neg\alpha$ [Spohn, 1988]. However, this condition leads to counterintuitive results; sentences turn out to be reasons simply by virtue of the fact they are strictly more plausible than the explanandum. Therefore we can say that reasons based on the condition above do not respect the notion of coherence introduced in section 4.

8 Discussion

Transmutations of an information system describe not only how a set of information is revised, but how the underlying preference structure is revised. Consequently, it is possible to capture Spohn's notion of reason using transmutations. Indeed we would obtain a different characterization of a Spohnian reason for every type of transmutation.

An adjustment is a transmutation that involves an *absolute minimal change*. That is, the underlying preference relation is changed as little as necessary to effect the required change in epistemic rank. For this reason we have argued that an adjustment is an appropriate transmutation for determining explanations because in the determination of an explanation we would normally require that such a change disturb the background information system as little as necessary. The suitability of using an adjustment for Spohnian reasons is further supported by its perspicuous connection with the comprehensive work of Boutilier and Becher.

We showed how an adjustment can be used to determine a most plausible explanation, by selecting an explanation that is capable of raising the epistemic rank of the explanandum the most. This notion is more discerning than that given by Boutilier and Becher whose method suffers from the problem of not being able to rank the plausibility of factual explanations.

We contrasted Spohnian reasons based on adjustment with a common definition of explanation in terms of abductive reasoning. We noted that the determination of reasons based on adjustments are carried out qualitatively using the relative ordering of the background information system itself, and the associated nonmonotonic consequence relation is consistency preserving and rational.

Future work will explore the relevance of competing explanations [Gärdenfors, 1993] and the strengths of competing explanations using the quantitative information encapsulated in an entrenchment ranking. Preliminary results suggest these are promising directions to pursue.

Acknowledgements

The authors have benefited from the helpful suggestions of Peter Gärdenfors, Abhaya Nayak, Eric Olsson, Judea Pearl, Pavlos Peppas and Hans Rott.

References

[Alchourrón *et al.*, 1985] Alchourrón, C., Gärdenfors, P., Makinson, D., *On the Logic of Theory Change: Partial Meet Functions for Contraction and Revision*, Journal of Symbolic Logic, 50: 510-530, 1985.

[Boutilier, 1993] Boutilier, C., *Revision Sequences and Nested Conditionals*, In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, 1993.

[Boutilier, and Becher, 1995] Boutilier, C. and Becher V., *Abduction as Belief Revision*, Artificial Intelligence Journal (to appear).

[Gärdenfors, 1988] Gärdenfors, P., *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, The MIT Press, Cambridge, Massachusetts, 1988.

[Gärdenfors, 1990] Gärdenfors, P., *The Dynamics of Belief Systems: Foundations vs. Coherence Theories*, Revue Internationale de Philosophie, 44:24–46, 1990.

[Gärdenfors, 1993] Gärdenfors, P., *On the Logic of Relevance*, J. P., Dubucs (ed), Philosophy of Probability, pp. 35–54, Kluwer Academic, 1993.

[Gärdenfors and Makinson, 1988] Gärdenfors, P., and Makinson, D., *Revisions of Knowledge Systems using Epistemic Entrenchment*, Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge, pp. 83–96, 1988.

[Gärdenfors and Makinson, 1994] Gärdenfors, P., and Makinson, D., *Nonmonotonic Inference Based on Expectations*, Artificial Intelligence Journal 65, pp. 197–245, 1994.

[Grove, 1988] Grove, A., *Two Modellings for Theory Change*, Journal of Philosophical Logic, 17:157–170, 1988.

[Katsuno and Mendelzon, 1992] Katsuno, H. and Mendelzon, A.O., *On the Difference between Updating a Knowledge Database and Revising it*, in Belief Revision, Gärdenfors, P. (ed), Cambridge University Press, Cambridge, 1992.

[Olsson, 1994] Olsson, E., *Representing Reasons Within a Coherentist Framework*, Unpublished Manuscript, Uppsala University, Sweden, 1994.

[Paul, 1993] Paul, G., *Approaches to Abductive Reasoning: An Overview*, Artificial Intelligence Review, 7:109–152, 1993.

[Peppas and Williams, 1995] Peppas, P., and Williams, M.A., *A Unified View of the Constructive Modellings for Revision*, Notre Dame Journal of Formal Logic, (to appear).

[Peppas, 1993] Peppas, P., *Belief Change and Reasoning about Action* , PhD Thesis, The University of Sydney, Australia, 1993.

[Rott, 1991] Rott, H., *Two Methods of Constructing Contractions and Revisions of Knowledge Systems*, Journal of Philosophical Logic, 20:149–173, 1991.

[Rott, 1992] Rott, H., *On the Logic of Theory Change: Partial Meet Contraction and Prioritized Base Contraction*, Report 27, Zentrum Philosophie and Wissenschaftheorie, Universität Konstanz, 1992.

[Spohn, 1983] Spohn, W., *Deterministic and Probabilistic Reasons and Causes*, Erkenntnis 19:371–396, 1983.

[Spohn, 1988] Spohn, W., *Ordinal Conditional Functions: A Dynamic Theory of Epistemic States*, In Harper, W.L., and Skyrms, B. (eds), Causation in Decision, Belief Change, and Statistics, II, Kluwer Academic, pp. 105–134, 1988.

[Stickel, 1991] Stickel, M.E., *A Prolog-Like Inference System for Computing Minimal-Cost Abductive Explanations in Natural-Language Interpretation*, Annals of Mathematics and Artificial Intelligence, 4:89–106, 1991.

[Williams, 1994a] Williams, M.A., *Transmutations of Knowledge Systems*, in J. Doyle, E. Sandewall, and P. Torasso (eds), Proceedings of the Fourth International Conference on the Principles of Knowledge Representation and Reasoning, Morgan Kaufmann, San Mateo, CA, pp. 619–629, 1994.

[Williams, 1994b] Williams, M.A., *Explanation and Theory Base Transmutations*, in the Proceedings of the European Conference on Artificial Intelligence, pp. 341–346, 1994.

[Williams, 1994c] Williams, M.A., *On the Logic of Theory Base Change*, in the Proceedings of the Fifth European Workshop on Logics in Artificial Intelligence, LNCS No. 835, 86 – 105, 1994.

[Williams, 1995a] Williams, M.A., *Iterated Theory Base Change: A Computational Model*, in the Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995 (this volume).

[Williams, 1995b] Williams, M.A., *Changing Nonmonotonic Inference Relations*, in the Proceedings of the World Conference on the Foundations of Artificial Intelligence, Paris, (to appear).