A knowledge-level account of abduction
(preliminary version)

Hector J. Levesque*
Dept. of Computer Science
University of Toronto
Toronto, Canada M5S 1A4

Abstract

In this paper, we consider a new definition of abduction that makes it depend on an underlying formal model of belief. In particular, different models of belief will give rise to different forms of abductive reasoning. Based on this definition, we then prove three main theorems: first, that when belief is closed under logical implication, the corresponding form of abduction is precisely what is performed by the ATMS as characterized by Reiter and de Kleer; second, that with the more limited "explicit" belief defined by Levesque, the required abduction is computationally tractable in certain cases where the ATMS is not; and finally, that something is believed in the implicit sense iff repeatedly applying a limited abduction operator eventually yields something that is believed in the explicit sense. This last result relates deduction and abduction as well as limited and unlimited reasoning all within the context of a logic of belief.

1 Introduction

Using the terminology of C. S. Peirce, given sentences α, β, and $(\alpha \supset \beta)$, there are three operations one can consider: from α and $(\alpha \supset \beta)$, one might *deduce* β; from α and β, one might *induce* $(\alpha \supset^1 \beta)$; and from β and $(\alpha \supset \beta)$, one might *abduce* α. Of course, characterizing precisely what should be deduced, induced, or abduced in various circumstances is quite another matter, and the last of these is the subject of this paper.

Abduction can be thought of as a form of hypothetical reasoning. To ask what can be abduced from β is to ask for an α which, in conjunction with background knowledge,[2] is sufficient to account for β. When α and β are about the physical world, this normally involves finding a cause α for an observed effect β. For instance, β might say that a symptom of some sort is observed and α might say that a disease is present. We often say in this case, that α *explains* β. But not all abduction is concerned with cause and effect. If we happen to known that Marc is 3 or 4 years old, the fact that he is not yet 4 does not *explain* his being 3, although it does imply it, given what is known.[3] It would be more accurate to say that the α is sufficient to tell us that the β is true. But this is a bit cumbersome, so with this caveat in mind, we will often use the explanation terminology here.

When it comes to formally characterizing abduction, existing approaches fall into two broad camps: those, like [Reggia, 1983, Allernand *et al.*, 1987], that are *set-cover* based, and those, like [Poole, 1988, Eshghi and Kowalski, 1988], that are *logic* based. In the former case, abduction is defined over sets of observations and hypotheses, in terms of coverings, parsimony, plausibility, and the like. A disadvantage of this approach is that it is difficult to express how a small change in the background knowledge can contribute to changing what counts as an explanation. In the latter case, however, this knowledge is represented directly as a logical theory, and α is considered an explanation for β if (1) it is logically consistent with what is known, and (2) together with this knowledge, logically implies β. The disadvantage of defining abduction in this way is that it locks the specification of reasoning into global properties of the logic such as consistency and implication. Different reasoning abilities, deductive or abductive, will then require different notions of implication or consistency.

Here we take a different approach and characterize abduction in terms of a model of belief. When belief is closed under ordinary logical consequence, this account will coincide with the idealized logic-based version. However, we can look at different forms of abduction by varying the underlying notion of belief, without changing the meaning of implication. This *knowledge-level* approach [Newell, 1982, Levesque and Brachman, 1986] will also

*Fellow of The Canadian Institute for Advanced Research. This research was made possible in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

[1] More likely, one would want to induce $\forall x(\alpha \supset \beta)$ from instances of α and β.

[2] The distinction between knowledge and belief is not important here, and we will use the terms interchangeably.

[3] Another reason for distinguishing this from explanation is that we normally say that a explains β only when we believe β to be true, for example, when we have observed the symptoms in question. So a true account of explanation *per se* is complicated by the fact that it must consider what was known *prior* to believing β [Gardenfors, 1988], or else there will be nothing left to explain, given what is known.

force us to characterize the abduction task independently of how the knowledge is represented, and thus afford the greatest freedom in how to represent and manipulate this knowledge at the symbol level.

In the next section, we introduce notation, discuss the need for a simplicity measure, and define a new knowledge-level operator EXPLAIN. In Section 3, we discuss the concept of regular belief and its relation to EXPLAIN. In Section 4, we examine a form of belief that is closed under logical consequence, and the ATMS as an abductive reasoner. In Section 5, we consider the specification of a more limited abductive reasoner. Finally, we draw conclusions in Section 7.

2 Abduction at the knowledge level

To define abduction, we start with $\mathcal{L}$, a standard propositional language (except that for convenience, we include a special constant $\Box$, for falsity). All beliefs will be expressed in $\mathcal{L}$. We use p, q, and r to range over the propositional letters of $\mathcal{L}$; α, β, and γ to range over the sentences of $\mathcal{L}$; m to range over the literals of $\mathcal{L}$; $\overline{m}$ to mean the complement of m; and, $\bigwedge\{\alpha_i\}$ to conjoin a set of sentences, and $\bigvee\{\alpha_i\}$ to disjoin them.

To talk about what is or is not believed, we use a logical language $\mathcal{L}^*$ that is structured like $\mathcal{L}$, except that all of its atomic sentences are of the form $\mathbf{B}\alpha$, where α is a sentence of $\mathcal{L}$.[4] For different kinds of belief, we use a subscript on the belief operator. So, $\mathbf{B}_\lambda\alpha$ says that α is a belief of type λ.

The languages $\mathcal{L}$ and $\mathcal{L}^*$ are both interpreted in the standard way in that the truth values of non-atomic sentences are the usual functions of the truth values of their components. For the atomic sentences of $\mathcal{L}$, an *assignment* is a total function from the propositional letters to $\{0,1\}$, and $w \models \alpha$ means that α is true with respect to assignment w according to the ordinary truth table (where $\Box$ always comes out false). For the atomic sentences of $\mathcal{L}^*$, we assume that an *epistemic state* of some sort determines which sentences of $\mathcal{L}$ are believed. The notation $e \models \mathbf{B}_\lambda\alpha$ says that $\mathbf{B}_\lambda\alpha$ is true at epistemic state e, which we will take as a primitive notion for now, until we look at specific types of belief.

For any sentence α of $\mathcal{L}$, it will be useful to talk about $\|\alpha\|$, the proposition expressed by α. Nothing hinges on how exactly propositions are defined, but for concreteness, they can be taken to be the set of all assignments where the sentence in question is true. Similarly, $\|\{\alpha_1,\ldots,\alpha_n\}\|$ is defined as $\{\|\alpha_1\|,\ldots,\|\alpha_n\|\}$.

2.1 Simplicity and uniqueness

Deductive and abductive reasoning appear to be duals, but one difference between the two is that in the case of deduction, we are usually interested in *testing* if some sentence is deducible, while in the case of abduction, we want to *produce* a sentence that is abducible.[5]

For example, consider a medical domain where sentences of $\mathcal{L}$ stand for properties that may or may not hold of a certain patient. Suppose we know that *male* and *(hepatitis ⊃ jaundice)* are both true. If we observe jaundice in the patient, we might be interested in determining what might explain it, based on what we know about the patient. In other words, we want to reason abductively from *jaundice,* to find something that accounts for it, given what is known. In this case, the answer is clearly *hepatitis,* but it is not obvious how to characterize in general the answers we are looking for.

First of all, we cannot expect a *single* explanation since, for example,

((¬¬*hepatitis* ∧ *migraines)* ∨ *(hepatitis* ∧ ¬*migraines))*

also accounts for *jaundice.* But even if we factor out logically equivalent sentences and think in terms of propositions, there will be propositions that are logically too strong, and others that are logically too weak. For instance, *(hepatitis* ∧ *migraines)* accounts for the jaundice in that it is consistent with what is known, and if it were true, then *jaundice* would be too. Similarly, *(hepatitis* ∨¬*male)* accounts for *jaundice* since it too is consistent with what is known, and if it were true, then *jaundice* would be also, since *male* is known to be true. Yet *(hepatitis/\ migraines)* implies *hepatitis* which implies *(hepatitis* ∨ -¬*malc).*

So what is it that distinguishes || *hepatitis*\\ from these other propositions? Is there a way to sort this out purely logically (in terms of sets of possible worlds and distance measures or whatever) and define an appropriate explanation? As it turns out, the answer is no. To see this, suppose to the contrary that there were a function *F* that given the proposition expressed by *( )* and the one expressed by B would always return the one expressed by *a.* That is, suppose that for every a and β, $F(\|(\alpha \supset \beta)\|,\|\beta\|)=\|\alpha\|$. Then, we would have $F(\|(q \supset q)\|,\|q\|)=\|q\|$ and $F(\|(\Box \supset q)\|,\|q\|)=\|\Box\|$. However, $\|(q \supset q)\|=\|(\Box \supset q)\|$ since the two sentences are logically equivalent.[6] But this implies that $\|\Box\|=\|q\|$, which is incorrect. So such a function *F* cannot exist, and we are forced to go beyond the logic of the sentences (that is, beyond the propositions expressed) to differentiate *hepatitis* from other potential explanations.

One obvious approach is to maintain a list of sentences that are marked as possible hypotheses as is done in [Poole, 1988, Reiter, 1987] and to only consider sentences appearing in this list. But this fails to account for why we find *hepatitis* so compelling as the unique explanation for *jaundice* in the above. Perhaps it is because *hepatitis* does not deal with any other conditions, either to insist on (conjoin) irrelevant restrictions like *migraines,* or to allow for (disjoin) possibilities known to be false like *-imalc.* This suggests that we should be looking for sentences that are as *simple* as possible in their subject matter. With this notion of simplicity in mind, we are ready to provide a formal definition of abduction in terms of belief.

[4] For the purpose of this paper, therefore, we will not consider beliefs about other beliefs.

[5] However, see Section 7 where the symmetry between deduction and abduction is reconsidered.

[6] A stronger argument would be needed for a notion of proposition that was finer-grained than logical equivalence.

2.2 A general definition

First, we define explanation wrt an epistemic state e for a type of belief λ:

Definition 1 α **expl**$_\lambda$ β **wrt** e **iff** $e \models [\mathbf{B}_\lambda(\alpha \supset \beta) \wedge \neg \mathbf{B}_\lambda \neg \alpha]$.[7]

This definition does not distinguish between trivial and non-trivial occurrences of $B(\alpha \supset \beta)$. For example, assuming that *(jaundice* $\supset$ *jaundice)* is believed but that $\neg$*jaundice* is not, *jaundice expl*$_\lambda$ *jaundice* wrt *e* holds, that is, having jaundice is clearly (and trivially) sufficient to account for having jaundice. More generally, if nothing is believed about β other than logical truths, then there will only be trivial explanations. In addition, for many types of belief, we have that if β is believed, then there will be no explanations at all, whereas ifβ itself is believed, then $\neg\Box$ will be-the unique explanation.

As discussed above, the definition of explanation must depend on some syntactic criterion of simplicity. Perhaps the easiest one is the following:

Definition 2 The literals of α, LITS(α), is defined by:
LITS($\Box$)=$\emptyset$; LITS(p)=$\{p\}$;
LITS($\neg\alpha$)=$\{\overline{m} \mid m \in \text{LITS}(\alpha)\}$;
LITS($\alpha \wedge \beta$)=LITS($\alpha \vee \beta$)=LITS(α) $\cup$ LITS(β).

Definition 3 α is *simpler* than β (written $\alpha \prec \beta$) iff LITS(α) $\subset$ LITS(β); also $\alpha \preceq \beta$ iff LITS(α) $\subseteq$ LITS(β).[8]

So "simpler" means "containing fewer propositional letters," but keeping track of their polarity. For example $(p \wedge \neg q) \prec (q \supset (p \vee r))$, but $(p \wedge \neg q) \not\prec (\neg q \supset (p \vee r))$. We define a simplest explanation in the obvious way:

Definition 4
α **min-expl**$_\lambda$ β **wrt** e iff α **expl**$_\lambda$ β **wrt** e and
for no $\alpha^* \prec \alpha$ is it the case that α^* **expl**$_\lambda$ β **wrt** e.

Finally, since there may be more than one simplest- explanation, and since we do not really care at this level how each simplest explanation is expressed, the task of abduction will be to return the set of propositions of all simplest explanations:

Definition 5
$\text{EXPLAIN}_\lambda[\![e, \beta]\!] = ||\{\alpha \mid \alpha \text{ min-expl}_\lambda \beta \text{ wrt } e\}||$.

These simplest explanations should be understood *disjunctively.* For example, if we know that *(p*$_1$ $\supset$ *q\)* and *(P'2* $\supset$ q_2), then *pi* is a simplest explanation of *(q*$_1$ Vq_2) and so is P2. However, it is the disjunction (p1 VP2) that fully and non-trivially accounts for *(q* V q2).

[7]In the final paper, various other options for these two conjuncts will be examined. Instead of the first one, we might want to say that if we were *told at,* then we would believe *β,* which need not be the same as believing $(\alpha \supset \beta)$ in the presence of defaults; instead of the second one, we might prefer saying for a given 7 that we do *not* believe $(\alpha \supset \gamma)$ (to handle negative evidence), which for regular belief (see below) coincides with the above when 7 is $\Box$

[8]For some applications, we might wish to use a superset of this relation. For example, we might want to say that $p \prec q$ even though both are atomic, if we consider *p* to be much more likely than *q.* But we should never have to consider a *subset* of the relation.

This completes the knowledge-level characterization of abduction. The theorems to follow below (especially the relationship to the ATMS) are the best evidence that the definition is apt. But it is worth noting here how simple and general the account is. It is the first (to my knowledge) that not only works for sentences *β* of arbitrary syntactic form, but is also sensitive to what is known without requiring an explicit list of the known sentences. In other words, it does *not* depend in any way on how the epistemic state e is represented (and so is truly at the knowledge level). Computations at the symbol level, of course, will need to operate on finite symbolic representations of that state. Typically, for each type of belief λ, there will be a function $\Re_\lambda$ that maps (finite) sets of sentences into epistemic states. At the symbol level, there will be a procedure of some sort that takes a representation of knowledge KB and a sentence β as arguments, and produces a set of sentences by abductive reasoning. For an abductive procedure to be *correct,* the sentences it returns must express all and only the simplest explanations of *(3* wrt the epistemic state represented by KB. Thus, what we will want to establish for various types of belief and associated computational procedures explain[KB,β] is the following:[9]

$$\text{EXPLAIN}_\lambda[\![\Re_\lambda(\text{KB}), \beta]\!] = ||\,\textbf{explain}[\text{KB}, \beta]\,||.$$

Note that for this general account, correctness does not require the sentences returned by the symbol-level procedure to be in a certain syntactic form, provided that they express the right propositions.

3 A generic abduction operation

Before looking at two specific types of belief, we define what it means for belief to be *regular.* In what follows, we use the following notation: *x, y,* and *z* stand for *clauses,* that is, finite sets of literals always understood disjunctively; the empty clause is $\Box$; *(x — y)* is the clause whose literals are those in the set difference *of x* and *y; x* is the set of complements of the literals in *x,* now understood conjunctively; E and F stand for sets of clauses; for any Σ, $\mu\Sigma$ is the set of smallest (in the sense of subset) elements of Σ; and finally, CNF*(cx)* is the set of smallest clauses that result from converting a to conjunctive normal form, and analogously for DNF(a).

Definition G A type of belief A is *regular* iff for every epistemic state, the following sentences of £* are true:

1. $\mathbf{B}_\lambda \neg \Box$;
2. $(\mathbf{B}_\lambda \alpha \vee \mathbf{B}_\lambda \beta) \supset \mathbf{B}_\lambda(\alpha \vee \beta)$;
3. $\mathbf{B}_\lambda(\alpha \wedge \beta) \supset (\mathbf{B}_\lambda \alpha \wedge \mathbf{B}_\lambda \beta)$;
4. $(\mathbf{B}_\lambda \alpha \wedge \mathbf{B}_\lambda \beta) \supset \mathbf{B}_\lambda(\alpha \wedge \beta)$;
5. $\mathbf{B}_\lambda \alpha \equiv \mathbf{B}_\lambda \alpha^*$, if α^* is α in CNF or DNF, or is the result of replacing any subformula β in α by β^*, where (recursively) $\mathbf{B}_\lambda \beta \equiv \mathbf{B}_\lambda \beta^*$ is always true.[10]

We now define a very general operation on two sets of clauses (which we will eventually use for both types of belief below) as follows:

[9]We use this font to indicate a symbol level procedure.
[10]Note that this does *not* sanction replacing *β* by everything logically equivalent to it.

Definition 7 $\nabla(\Sigma,\Gamma) =$

$$\mu\left\{\overline{z} \mid \begin{array}{l}\forall y \in \Sigma,\ y \not\subseteq z \quad \text{and} \\ \forall x \in \Gamma,\ \exists y \in \Sigma,\ x \cap y \neq \emptyset \text{ and } (y - x) \subseteq z\end{array}\right\}$$

This generalizes the MIN-SUPPORTS operation of [Reiter and de Kleer, 1987]: MIN-SUPPORTS(x,Σ) = $\nabla(\text{IMPS}(\Sigma), \{x\})$, where IMPS($\Sigma$) is defined below. Informally, the elements of Σ should be thought of as the clauses that are believed, the elements of Γ as the clauses to be explained, and $\nabla(\Sigma,\Gamma)$ as the minimal explanations. For instance, if

$\Sigma = \{(p_1 \vee \overline{p}_4), (p_1 \vee \overline{p}_5 \vee p_7), (p_2 \vee \overline{p}_6 \vee p_7), (\overline{p}_3 \vee \overline{p}_8), (\overline{p}_4 \vee \overline{p}_8)\}$
and
$\Gamma = \{p_1, (p_2 \vee \overline{p}_3)\}$,

then $\nabla(\Sigma,\Gamma) = \{(p_4 \wedge p_6 \wedge \overline{p}_7), (p_5 \wedge p_6 \wedge \overline{p}_7), (p_5 \wedge \overline{p}_7 \wedge p_8)\}$. That is, if we take one of these explanations, $(p_4 \wedge p_6 \wedge \overline{p}_7)$, and assume that it and the elements of Σ are true, we get that $(p_1 \wedge p_2)$ must be true, which implies that the clauses of T must all be true. The other two explanations work analogously. Note that $(p_4 \wedge p_8)$ is *not* returned as an explanation since it is believed to be false, that is, its negation is an element of Σ.

The important property of ∇ is that although it only deals with clauses, it can be used to provide correct abductive reasoning for regular belief:

Theorem 1 *For regular belief,*
$\text{EXPLAIN}_\lambda[\![e,\beta]\!] = \|\nabla(\{y \mid e \models \mathbf{B}_\lambda y\}, \text{CNF}(\beta))\|$.

Proof: The proof depends on two key lemmas:

Lemma 1.1 *If* α expl_λ β **wrt** e *then*
$\exists \overline{z} \in \text{DNF}(\alpha)$, $\overline{z}$ expl_λ β **wrt** e.

Lemma 1.2 $\overline{x}$ expl_λ β **wrt** e *iff*
$\exists z \subseteq x, \overline{z} \in \nabla(\{y \mid e \models \mathbf{B}_\lambda y\}, \text{CNF}(\beta))$.

The final paper proves these and the theorem. ■

What this theorem establishes at a very abstract level is that for regular belief, it is sufficient to work with the set of clauses believed and the CNF of the sentence to be explained. This will immediately lead to two abductive procedures below.

4 Case 1: Implicit belief

The first notion of belief we consider is the "classical" one where beliefs are closed under logical consequence. Following [Levesque, 1984], we call this *implicit belief* and use B_1 as the belief operator. An epistemic state for implicit belief can be modeled by any set of assignments, where we have the following:

$$e \models \mathbf{B}_1\alpha \quad \text{iff} \quad \text{for every } w \in e,\ w \models \alpha.$$

If KB is a set of sentences, then R_1(KB), the epistemic state represented by KB, is modeled by the set of all assignments that satisfy every element of KB. What is believed in this state is precisely what follows from KB, that is, if $e = \Re_1(\text{KB})$, we have that $e \models \mathbf{B}_1\alpha$ iff KB $\models \alpha$. From this it follows that

α expl_1 β **wrt** e iff
KB $\cup \{\alpha\} \models \beta$ and KB U {a} is consistent,

which is precisely the account of explanation given by (among others) Poole in [Poole, 1988].

4.1 The ATMS

One abductive procedure that is receiving considerable attention is the ATMS [de Kleer, 1986]. Unfortunately, descriptions of the overall function computed by the ATMS have been largely in terms of *how* it goes about computing it. The first account that attempted to provide a logical reconstruction was that of Reiter and de Kleer in [Reiter and de Kleer, 1987]. Although idiosyncratic terms like labels, nodes, and nogoods are no longer part of the formulation, their definition is in terms of clause intersections and differences, notions that are (arguably) still best understood as symbol level manipulations of sentences in a certain form. However, given their characterization, they are able to show the following:

Definition 8
The *implicants* of Σ, $\text{IMPS}(\Sigma) = \{y \mid \Sigma \models y\}$.

Theorem 2 (Reiter and de Kleer) *Given a set of Horn clauses* Σ *and a letter* p, *the ATMS procedure is defined by* $\mathbf{atms}[\Sigma,p] = \{(q_1 \wedge \ldots \wedge q_k) \mid k \geq 0$ and $\{\overline{q}_1, \ldots, \overline{q}_k, p\} \in \mu\text{IMPS}(\Sigma)\}$.

In fact, Reiter and de Kleer generalize the account of the ATMS to where the first argument is not necessarily Horn and the second argument is any clause. However, we can go even further by noting that

$$\begin{aligned}\mathbf{atms}[\Sigma,p] &= \{(\overline{y} - \{\overline{p}\}) \mid p \in y \text{ and } y \in \mu\text{IMPS}(\Sigma)\} \\ &= \mu\{\overline{z} \mid \forall y \in \text{IMPS}(\Sigma),\ y \not\subseteq z \quad \text{and} \\ &\qquad \exists y \in \text{IMPS}(\Sigma),\ p \in y \text{ and } (y - \{p\}) = z\} \\ &= \nabla(\text{IMPS}(\Sigma), \{\{p\}\}).\end{aligned}$$

Using this as a pattern, we can define a generalized ATMS as follows:[11]

Definition 9 $\mathbf{gatms}[\Sigma,\beta] = \nabla(\text{IMPS}(\Sigma), \text{CNF}(\beta))$.

Clearly this coincides with the ATMS specification when Σ is a set of Horn clauses and β is a propositional letter. But what do these operations *mean,* and why should anyone care about them? The answer, we claim, is that the ATMS procedure correctly performs abduction for implicit belief:

Theorem 3 $\text{EXPLAIN}_1[\![\Re_1(\Sigma),\beta]\!] = \|\mathbf{gatms}[\Sigma,\beta]\|$.

Proof: It is not hard to show that implicit belief is regular, and we have that

$$\Re_1(\Sigma) \models \mathbf{B}_1 x \quad \text{iff} \quad x \in \text{IMPS}(\Sigma).$$

The theorem then follows from Theorem 1. ■

However else it has been characterized in the past, this theorem establishes that an ATMS can be understood as computing all simplest explanations with respect to this type of implicit belief. Among other things, this guarantees that Poole's account of abduction (with the addition of the notion of simplicity defined here) also specifies the task performed by an ATMS.

5 Case 2: Explicit belief

The second notion of belief we consider is a variant of the one introduced in [Levesque, 1984] called *explicit belief.* We use B_E as the belief operator for beliefs of this type.

In the final paper, we will consider a very different way of generalizing the ATMS to handle arbitrary sentences.

The motivation behind explicit belief was to study a form of belief that was more computationally tractable than implicit belief, but remained defined in terms of truth conditions on the sentences believed. Since a sentence is implicitly believed if it comes out true at each element of a set of assignments (or alternatively, accessible possible worlds), it follows that implicit belief is closed under logical consequence. For explicit belief, instead of using assignments, we use *situations,* which can be taken to be total functions from the *literals* to {0,1}, such that for every *p,* at least one of *p* or *p* is assigned to 1.[12] We can think of assignments as those situations where *s(p)* = 1 — *s(p)* for every letter *p.* But because not every situation is an assignment, we must define truth support recursively over sentences *and* their negations:

$$s \models p \text{ iff } s(p) = 1; \qquad s \models \neg p \text{ iff } s(\overline{p}) = 1;$$
$$s \models (\alpha \wedge \beta) \text{ iff } s \models \alpha \text{ and } s \models \beta;$$
$$s \models \neg(\alpha \wedge \beta) \text{ iff } s \models \neg\alpha \text{ or } s \models \neg\beta;$$
$$s \models \neg\neg\alpha \text{ iff } s \models \alpha.$$

An epistemic state for explicit belief is modeled by a set of situations where we have the following:

$$e \models \mathbf{B_E}\alpha \text{ iff for every } s \in e,\ s \models \alpha.$$

As in [Levesque, 1984], it is also useful to talk about the implicit beliefs of e:

$$e \models \mathbf{B_I}\alpha \text{ iff for every assignment } s \in e,\ s \models \alpha.$$

As before, $R_E(KB)$ is modeled by the set of all situations that satisfy every element of KB. What is explictly believed in such a state is not what logically follows from KB, but rather what is *tautologically entailed* by the KB (once tautologies are taken into account) in the sense of Relevance Logic [Anderson and Belnap, 1975, Dunn, 1976]. More precisely, if $e = \mathfrak{R}_{\mathbf{E}}(KB)$, we have that $e \models \mathbf{B_E}\alpha$ iff KB U T tautologically entails a, where T is the set of all clauses of the form {p,p}.

5.1 Limited abductivc reasoning

To establish what form of abductive reasoning is appropriate for explicit belief, we need something that will play the role that IMPS(E) played for implicit belief:

Definition 10
EXPS(E) = $\{y \mid y \text{ is tautologous or } \exists y^* \in \Sigma,\ y^* \subseteq y\}$.

The abductive reasoning we will use for explicit belief is the same as that performed by the ATMS, but using EXPS(E) instead of IMPS(E):

Definition 11 $\mathbf{abd}[\Sigma, \beta] = \nabla(\mathbf{EXPS}(\Sigma), \mathbf{CNF}(\beta))$.

To see the difference between this procedure and the ATMS, suppose that KB1 = $\{\{q\}, \{\overline{s} \vee p\}, \{\overline{p} \vee \overline{q} \vee r\}\}$. In this case, atms[KBi,r]={r,s,{pA q}}, so there are three simplest explanations for r wrt implicit belief; but abd[KB_1, r]={r, *{p* A *q}},* so *s* is *not* a simplest explanation for r wrt to explicit belief. The difference is that whereas *(s* V r) is implicitly believed (since it follows from KB_1), it is not explicitly believed. In other words, unlike the ATMS, abd[Σ,β] will not chain backwards to see what might explain β, and this is exactly what is required for explicit belief:

Theorem 4 $\text{EXPLAIN}_E[\mathfrak{R}_{\mathbf{E}}(\Sigma), \beta] = \|\mathbf{abd}[\Sigma, \beta]\|$.

Proof: Like implicit belief, explicit belief is regular. Also we have that

$$\mathfrak{R}_{\mathbf{E}}(\Sigma) \models \mathbf{B_E}x \quad \text{iff} \quad x \in \mathbf{EXPS}(\Sigma).$$

The theorem then follows from Theorem 1. ■

This theorem establishes that abd[E, β] correctly calculates all simplest explanations with respect to this type of explicit belief.

But why should we care about a procedure that cannot find some perfectly reasonable explanations that can be found by an ATMS? The problem is that we may have to wait too long for an ATMS to find them. This has caused researchers to look for parallel realizations of the procedure [Dixon and de Kleer, 1988]. But this is not just an ATMS *implementation* problem; the *task* it performs is inherently difficult: in general, there will be an exponential number of clauses to find,[13] and just deciding if {p,p} has *any* explanations at all is equivalent to determining whether or not the set of clauses E is satisfiable. So although (a parallel version of) the ATMS may work fine in many application areas, as a general-purpose mechanism for abductive reasoning, it has serious computational drawbacks.

On the other hand, just as explicit belief is easier than implicit belief when it comes to deductive reasoning, a similar result carries over to abductive reasoning:

Theorem 5 *If* KB *is in CNF, there is an* O(|KB|-|X|) *algorithm for calculating* abd[KB, x].

Proof: We use the fact that

$$\text{abd[KB,z]} = \{(\overline{y} - \overline{x}) \mid y \in \mu\mathbf{EXPS}(\mathbf{KB}),\ x \cap y \neq \emptyset\}.$$

We construct the answer as follows: cycle through the elements of μKB, and for each *y* that is not tautologous and that has an intersection with x, put *(y — x)* into a set *T.* Then, for each *m* € x, put *m* into T, unless *{m}* ∈ KB. Finally, return z for each $z \in \mu$T. ■

So for single clauses anyway, abductive reasoning for explicit belief is considerably easier than abductive reasoning for implicit belief.

For arbitrary sentences , the case is not so clear even if *(3* is in CNF. Although we can quickly calculate abd[E, *x]* for each clause *x* in β, putting the answers together involves converting a sentence into DNF:

Theorem 6 *Suppose* $\text{CNF}(\beta) = \{x_1, \ldots x_n\}$.

$$\text{abd}[\Sigma, \beta] = \mu\{\overline{z} \mid \forall y \in \text{EXPS}(\Sigma),\ y \not\subseteq z \text{ and } \overline{z} \in \text{DNF}(\textstyle\bigwedge_{i=1}^{n} \bigvee \mathbf{abd}[\Sigma, x_i])\}.$$

Proof: The proof will appear in the final paper. ■

To see how this works, let KB_2 be $KB_1 \cup \{\{\overline{a} \vee b\}, \{\overline{c} \vee b\}\}$, where KB_1 is defined above. Then $\mathbf{abd}[KB_2, (r \wedge b)]$ can be computed by calculating $\bigvee\mathbf{abd}[KB_2, r]$ (as above), which gives $(r \vee (p \wedge q))$, then $\bigvee\mathbf{abd}[KB_2, b]$, which gives $(a \vee b \vee c)$,

[12] This restriction on situations was not present in [Levesque, 1984]. It has the effect of making explicit belief similar to the knowledge retrieval of [Frisch, 1988] in that tautologies are always believed. This does not adversely affect the desirable computational properties of explicit belief, since for (non-quantificational) CNF, tautologies can be detected in linear time.

[13] In an unpublished note, David McAllester shows that this remains true even when Σ is a set of Horn clauses.

and then conjoining and putting the result into DNF, which gives

$$\{\{p \wedge q \wedge a\}, \{p \wedge q \wedge b\}, \{p \wedge q \wedge c\}, \{r \wedge a\}, \{r \wedge b\}, \{r \wedge c\}\}.$$

The only potential difficulty here is calculating the DNF. When β has very few clauses, or when almost all of the abd[Σ,x] return fewer than 2 simplest explanations, the entire operation will be fast. But to *guarantee* that it will work well in all cases appears to require an even more restricted form of belief.[14]

6 From explicit to implicit belief

One of the reasons for introducing explicit belief in [Levesque, 1984] was to specify a tractable deductive service for Knowledge Representation in terms of a set of beliefs which, unlike the implicit ones, could always be reliably computed. However, one difficulty with this whole approach is how exactly to go beyond what is explicitly believed. When deliberately trying to solve a problem (in what is called *puzzle mode* in [Levesque, 1988]), it is necessary to combine beliefs and follow through on their consequences in a controlled and systematic way. If all that is available at the knowledge level is a way of finding out if something is explicitly believed and a way of finding out if something is implicitly believed (in one very large unsupervised step), there is nothing the agent can do to begin exploring in a controlled way the implications of what is explicitly believed. For instance, the agent cannot simply perform theorem proving over what is known without access to the sentences at the symbol level used to represent that knowledge.

With a limited abduction operation, on the other hand, there is a way of moving under the control of the agent from the explicit beliefs towards the implicit beliefs. To find out if a sentence is implicitly believed, the procedure (roughly) is this: first find out ifβ is explicitly believed; if it is, then exit with success; otherwise, calculate the full (explicit) explanation for /?; if there is none or it is trivial, then exit with failure; otherwise replace *β* by the explanation, and repeat. In other words, the procedure deals with the following questions, starting with some ft^0*:* according to what is believed,

is ft^0 true? what would it take for ft^0 to be true? (call that β^1)
is ft^1 true? what would it take for β^1 to be true? (call that ft^2)
is ft^2 true? *etc.*

This "backward-chaining" procedure terminates when it either finds something that is believed or fails to find a non-trivial explanation. Each step in this procedure is tractable,[15] and the agent can exit the loop if it seems to be taking too long relative to the importance of the original question. More formally, we have the following:

Definition 12 For any epistemic state e and any β from $\mathcal{L}$, a sequence of sentences β_e^k, $k = 0, 1, 2, \ldots$ is defined by $\beta_e^0 = \beta$ and $\beta_e^{k+1} = \bigvee \text{EXPLAIN}_{\mathbf{E}}[\![e, \beta^k]\!]$.[16]

Theorem 7 $e \models \mathbf{B}_1\beta$ *iff for some* k, $e \models \mathbf{B}_{\mathbf{E}}\beta_e^k$.

Proof: The proof is based on the following:

> **Lemma 7.1** *If* Σ *is satisfiable,* $\Sigma \cap \Gamma = \emptyset$, *and there exists a linear set-of-support resolution refutation of* $\Sigma \cup \Gamma$, *with* Γ *as the set of support, then* $e \models \mathbf{B}_{\mathbf{E}}\alpha^k$ *where* $e = \Re_{\mathbf{E}}(\Sigma)$, $\alpha = \bigvee\{\overline{x} \mid x \in \Gamma\}$, *and* k *is the depth of the refutation tree.*
>
> In the final paper, we prove this lemma and the theorem. ■

So a sentence is implicitly believed iff it is accounted for *ultimately* by something that is explicitly believed.

To see this in action, let $KB_3 = KB2 \cup \{\{s\}, \{a \vee c\}\}$, where KB2 is defined above, and let $e = \Re_{\mathbf{E}}(\mathbf{KB_3})$, the epistemic state represented by KB3. Although $(r \wedge b)$ is not explicitly believed in state e, it is implicitly believed and so should be derivable. First we set *(3°* to $(r \wedge b)$, and compute $\beta^1 = \text{EXPLAIN}_{\mathbf{E}}[\![e, \beta^0]\!] = \mathbf{abd}[KB_3, \beta^0]$, which is

$$\{\{p \wedge q \wedge a\}, \{p \wedge q \wedge b\}, \{p \wedge q \wedge c\}, \{r \wedge a\}, \{r \wedge b\}, \{r \wedge c\}\}.$$

as presented above for KB_2. Putting this into CNF, we get $\{\{a \vee b \vee c\}, \{q \vee r\}, \{p \vee r\}\}$.[17] Notice that the first two clauses of β^1 are already explicitly believed by KB_3. Now calculate $\beta^2 = \text{EXPLAIN}_{\mathbf{E}}[\![e, \beta^1]\!] = \mathbf{abd}[KB_3, (p \vee r)] = \{r, p, s\}$. But then β^2 is explicitly believed (since s is), and so we are done. Notice how the iterative procedure works its way hack to s the way an ATMS would, but now in bite-sized pieces under the control of the agent.[18]

This theorem thus has the following perhaps surprising conclusion: we can determine if something is a logical consequence of what is (explicitly) believed without ever getting access to the set of sentences that are believed. We need only be able to ask for any specific sentence two questions: is it believed? and if not, what would be sufficient to account for it, according to what is believed?

The theorem also provides for the first time a knowledge-level account, that is, an account that is independent of how knowledge is symbolically represented, of how a limited notion of belief can be extended systematically to include all of its logical consequences. It also suggests a knowledge-level account of how an agent's beliefs could be made to evolve deductively over time: starting with some beliefs in some state e0, the agent would believe a in state efc+i iff he believed an explanation of *a*

It appears that a type of belief that is regular except for condition 4, closure under conjunction, does the trick here, but this needs further investigation.

Strictly speaking this is not true because of the DNF problem noted in the previous section. I suspect, however, that the procedure will also work for the more restricted notion of belief, but this has yet to be established.

We are abusing notation here in treating the result of EXPLAIN as a set of sentences.

In practice, one would not want to iterate an abductive procedure that takes the trouble of putting its answer into DNF, since the next step of the iteration requires an argument that is in CNF.

[18] Similar iterative techniques, we suspect, will lead to a procedure for full (implicit) abductive reasoning, as a controlled alternative to the ATMS itself.

in state e_k.[19] With a deductive architecture of this form, we would have that a sentence was implicitly believed iff at some point in the future it would be explicitly believed. Interestingly enough, to represent at the symbol level the epistemic state e_k, it is not necessary (though certainly sufficient and perhaps desirable) to use a set of sentences KB^k (calculated from some initial representation KB_0). We might just as well represent the state e_k by a pair of symbols $<KB_0, k>$, since this pair also determines whether or not $e_k \models \mathbf{B}_\mathbf{E}\alpha$, for any *a.* One nice property of our knowledge-level characterization is that it does not commit us to representing what is known using a set of sentences.

7 Conclusion

There are a number of questions left open by this research: how should nested beliefs be handled? What about quantified beliefs? Is there indeed (as hinted above) an account of limited belief that leads to tractable abduction for any sentence and yet can be iterated to produce all implicit beliefs? Can explicit abduction be used to produce implicit abduction?

On another front, one possibility suggested by this work is a new deductive operation. If abductive reasoning asks "what would tell you that β is true," we might also consider asking "what would *a* tell you/' and expect to find a simplest sentence (or a set of them) that captures what or adds to what is believed. This would be very useful information when doing "what-if" experiments, for example. In the medical domain, we could ask what we should expect to see if *hepatitis* were present and get *jaundice* as the only (non-trivial) answer. This is just the dual of abduction, and the two operations should be interdefmable. Indeed, the appropriate definition appears to be something like $-\bigvee \text{EXPLAIN}_\lambda[\![c, \neg\alpha]\!]$.

8 Acknowledgements

1 have been mulling over the ideas presented here for some time. I wish to thank Jim des Rivieres, Greg McArthur, and Calvin Ostrum for general discussion on this topic, Bart Selman for exploring the complexity issues, Syd Hurtubise for the specious reasoning, and Ron Brachman, Russ Greiner, Gerhard Lakemeyer, Joe Nunes, Ray Reiter and Rick Sobiesiak for commenting on an earlier draft. But I especially want to thank Ray Reiter and Johan de Kleer for writing the first ATMS paper I could get a real handle on.

References

[Allemand *et* a/., 1987] D. Allemand, M. Tanner, T. Bylander, and J. Josephson. On the computational complexity of hypothesis assembly. In *IJCAI-87,* Milan, Italy, August 1987.

[19]This is not unlike the so-called step-logics of [Drapkin and Pedis, 1986], but without the requirement that the beliefs available at any time be a finite set. Here, the requirement is only that the beliefs at any step be readily computable.

[Anderson and Belnap, 1975] A. Anderson and N. Belnap. *Entailment: The Logic of Relevance and Necessity.* Princeton University Press, Princeton, NJ, 1975.

[de Kleer, 1986] J. de Kleer. An assumption-based TMS. *Artificial Intelligence,* 28(2):127-162, 1986.

[Dixon and de Kleer, 1988] M. Dixon and J. de Kleer. Massively parallel assumption-based truth maintenance. In *AAAI-88,* pages 199-204, Saint Paul, MN, August 1988.

[Drapkin and Perlis, 1986] J. Drapkin and D. Perlis. Step-logics: an alternative approach to limited reasoning. In *Proc. 7th ECAI,* pages 160-163, Brighton, England, July 1986.

[Dunn, 1976] M. Dunn. Intuitive semantics for first-degree entailments and 'coupled trees'. *Philosophical Studies,* 29:149-168, 1976.

[Eshghi and Kowalski, 1988] K. Eshghi and R. A. Kowalski. Abduction as deduction. Technical report, Dept. of Computing, Imperial College of Science and Technology, London, England., 1988.

[Frisch, 1988] A. Frisch. Knowledge retrieval as specialized inference. Technical Report UIUCDCS-R-88-1404, Dept. of Computer Science, University of Illinois, Urbana, Illinois, 1988.

[Gardenfors, 1988] P. Gardenfors. *Knowledge in flux: modeling the dynamics of epistemic states.* MIT Press, Cambridge, MA, 1988.

[Levesque and Brachman, 1986]
11. Levesque and R. Brachman. Knowledge level interfaces to information systems. In M. Brodie and J. Mylopoulos, editors, *On Knowledge Base Management Systems: Integrating Artificial Intelligence and Database Technologies,* pages 13 34. Springer-Verlag, New York, 1986.

[Levesque, 1984] H. Levesque. A logic of implicit and explicit belief. In *AAAI-84,* pages 198-202, Austin, TX, 1984.

[Levesque, 1988] H. Levesque. Logic and the complexity of reasoning. *The Journal of Philosophical Logic,* 17:355-389, 1988.

[Newell, 1982] A. Newell. The knowledge level. *Artificial Intelligence,* 18(1):87-127, 1982.

[Poole, 1988] D. Poole. A methodology for using a default and abductive reasoning system. Technical report, Dept. of Computer Science, University of Waterloo, Waterloo, Ont., 1988.

[Reggia, 1983] J. Reggia. Diagnostic expert systems based on a set-covering model. *International Journal of Man Machine Studies,* 19(5):437-460, 1983.

[Reiter and de Kleer, 1987] R. Reiter and J. de Kleer. Foundations of assumption-based truth maintenance systems: preliminary report. In *AAAI-87,* pages 183-188, Seattle, WA, August 1987.

[Reiter, 1987] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence,* 32(I):57-96, 1987.