SRI International

Technical Report 505 • April 1991

ABDUCTION vs. CLOSURE IN CAUSAL THEORIES

Kurt Konolige

Artificial Intelligence Center
Computer and Engineering Sciences Division

The research reported in this paper was supported partially by the NTT Corporation, and partially by the Office of Naval Research Contract N00014-89-C-0095.

333 Ravenswood Avenue • Menlo Park, CA 94025-3493 • (415) 326-6200 • FAX: (415) 326-5512 • Telex: 334486

Abstract

There are two distinct formalizations for reasoning from observations to explanations, as in diagnostic tasks. The consistency based approach treats the task as a deductive one, in which the explanation is deduced from a background theory and a minimal set of abnormalities. In the other treatment, based on abduction, the explanations are considered to be sentences that, when added to the background theory, account for the observations. We show that there is a close connection between these two formalizations. Starting with a causal theory, explanations can be generated either by abductive reasoning, or by adding closure axioms and minimizing causation within a deductive framework. The latter method is strictly stronger than the former, but requires full knowledge of causation in a domain.

Contents

1 Introduction

Reasoning to the best explanation is common task in many areas of Artificial Intelligence. One of the clearest examples is diagnosis, in which one reasons from observations such as patient symptoms to their underlying causes, a disease or physiological malfunction. In the literature, there are two fundamentally different formalizations of this task [11; 9]. In one, the process of finding a cause is treated as a straightforward abductive task. Representative of this approach is the set-covering model of diagnosis [12], which assumes two disjoint sets, d a set of disorders, and m a set of manifestations. Disorders are assumed to "cause" manifestations, represented by a relation $d \times m$. The problem of diagnosis is recast as the problem of finding a minimal cover of observed manifestations $m' \subset m$, that is, a minimal subset of d that causes m'.

The competing formalization, the consistency based approach, is best represented by Reiter's theory of diagnosis from first principles [14]. In this theory, the functionality of a system containing a finite number of components is characterized by a set of first-order sentences, the background theory. The special predicate $ab(c)$ is used to state that the component c is abnormal or not functioning correctly. The current behavior of the system is given by a set of observation sentences. A diagnosis of the behavior is a minimal set of abnormality assumptions that is consistent with the observations and the background theory.

These two formalizations seem fundamentally different. The abductive approach looks for a set of causes that will imply the observations; the consistency based approach looks for a set of abnormality assumptions that are consistent with the observations. Nevertheless there is a connection between the two: Reiter showed how to express the set-covering model within his framework. Recently, Console [2] and Poole [9] have shown that either formalization can be used in restricted settings to compute the same explanations for diagnosis. In the abductive framework, the domain theory has axioms that relate causes and their effects, e.g., $c_i \supset e$ would be used to say that the effect e is a result of cause c_i. A corresponding consistency based theory is created by adding closure axioms stating that the *only* way to achieve an effect is by the set of causes given $(e \supset c_1 \vee c_2 \vee \cdots)$. The closure axioms are local in that they are easily derived by looking at all the implications that have a common head atom. The explanations computed

by the two methods are the same, as long as the domain theory contains just horn-clause implications from causes to effects, and is acyclic.

This result applies to diagnostic tasks that require explanations, that is, the unexpected observations must be predicted or explained from the assumed malfunctions. In the literature, explanatory diagnosis is usually signalled by the presence of fault models [15; 4]. Reiter's framework may also be used for a weaker form of diagnosis, which could be called *excusing* diagnosis: identify components that, if malfunctioning, would cancel or excuse predicted normal behavior of the system that conflicts with the observations. Here we look only at the case of explanatory diagnosis (and causal explanation in general), since excusing diagnosis has no analog in the abductive framework.

The restrictions on the domain theory for the Console/Poole result are very tight; in particular, there can be no correlation information (e.g., that two causes are mutually exclusive, or that one effect is the negation of another) or uncertainty (e.g., a cause implying a disjunction of effects). In this paper we will examine the connection between abduction and closure in the setting of explanation in general causal models, allowing correlations, uncertainty, and acyclicity in the causal structure. We answer the following questions.

- Is there a notion of explanatory closure that is appropriate for the more general domain theory? Is there an equivalent local closure?
- Is consistent explanatory closure of a general domain theory possible?
- When consistent closure is possible, does minimization of causes in the closed theory compute the same explanations as does abduction in the original theory?

There are both positive and negative results. With an appropriate notion of explanatory closure, given certain technical conditions, the consistency based approach will compute the same explanations as the abductive approach. However, the utility of the former method is open to question, since local closure will no longer suffice for explanatory closure: there seems to be no way to close the domain theory other than by computing all explanations. Further, the consistency based method is strictly stronger than the abductive one in explanatory diagnosis tasks, and the answers it produces may have elements that are not relevant to a causal explanation.

2 Simple Causal Theories

We are interested in domains in which there is a concept of cause and effect. Much of our commonsense view of the world can be cast into this form. Typical here is reasoning about actions or events and their results, usually formalized in the situation calculus or some variant [7]. Other domains include medical diagnosis with diseases as causes, symptoms as effects; mechanical or electrical systems with components and inputs as causes, outputs as effects; and planning domains with plans as causes, actions as effects.

While there is a great deal of complexity and controversy in defining causation, for this paper most of these problems can be bypassed because we are interested in a formal representation of the simplest aspects of causal consequence, given by the following definition.

DEFINITION 2.1 *Let $\mathcal{L}$ be a first-order language. A* simple causal theory *is a tuple $\langle C, E, \Sigma \rangle$ where*

- *C, a set of atomic sentences of $\mathcal{L}$, is the* causes.
- *E, a set of sentences of $\mathcal{L}$, is the* effects.
- *Σ, a set of sentences of $\mathcal{L}$, is the* domain theory.

The set C contains those atomic propositions which represent the possible causative agents of the domain. If we are looking for an answer to the question of "what caused e?", then an acceptable answer is some subset of C.[1]

Effects E are those aspects of the domain that we might observe and about which we want to know the cause. Note that E and C need not be disjoint; an observed cause may require no further explanation.

The domain theory Σ contains information about the relation between causes and effects. For example, in the situation calculus we might take C to be occurrences of actions, E to be properties of the final state, and Σ to hold information about the initial state and the way in which actions affect properties of situations.

[1] Allowing only atoms simplifies the analysis, but is not restrictive, since we can include equivalences such as $c \equiv \phi$, where ϕ is a complex sentence.

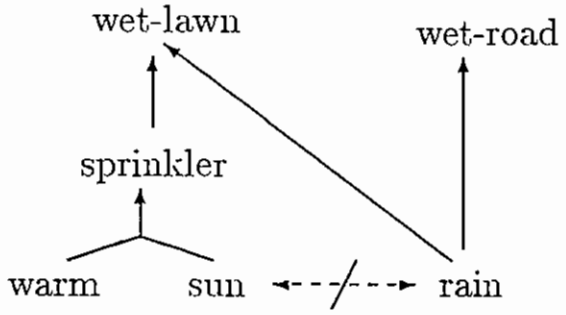

Figure 1: A sample causal theory

Here is a simple causal theory that will be used as an example in the rest of the paper; a graphical presentation appears in Figure 1. The intended meaning of the predicates should be obvious from their names.

Causes:	*rain*, *sun*, *warm*, *sprinkler*
Effects:	*wet-lawn*, *wet-road*
Domain theory:	*rain* $\supset$ *wet-road*, *rain* $\supset$ *wet-lawn*, *sun* $\equiv \neg$*rain* *sprinkler* $\supset$ *wet-lawn*, *sun* $\wedge$ *warm* $\supset$ *sprinkler*

A notational convention: a finite set of sentences will often be taken as a conjunction, e.g., if A and B are such sets, we write

$A \vee B$ for $(a_i \wedge a_2 \cdots) \vee (b_1 \wedge b_2 \cdots)$

$\neg A$ for $\neg(a_1 \wedge a_2 \cdots)$.

3 The Abductive Approach

Given a simple causal theory, the problem of reasoning from observations to causes can be expressed formally using abduction. The account of logical abduction we give here draws on ideas already present in the literature (e.g., [9]).

DEFINITION 3.1 *Let $\langle C, E, \Sigma \rangle$ be a simple causal theory. An* explanation *of a set of observations $O \subseteq E$ is a finite set $A \subseteq C$ such that*

- *A is consistent with Σ.*
- *$\Sigma \cup A \vdash O$.*
- *A is subset-minimal.*

If O has a nonzero finite number of explanations, then the cautious explanation *is their disjunction: $\bigvee_i A_i$.*

Remarks. A must be a minimal set of members of C; by minimal is meant there is no other explanation that is a proper subset. If these atoms constitute valid causes of the observed effects, and if there is an explanation that contains fewer causes, it should be preferred. Other than this we say nothing about preferences among multiple explanations. It is obvious that often such preferences will be required for reasoning, e.g., we may want the most specific explanation, or the most normal (where we partition causes into ones that normally occur and ones that do not), or the X-est, where X is some measure on explanations. The preference could be expressed mathematically by a partial order on the subsets of C. Since such an order will be closely related to the domain of application, and we have no way of making any general statements about the order, we omit it from further consideration here.

In a given problem domain, we may be interested in the best explanation, or the cautious explanation, or even any (satisficing) explanation. For example, if we want to predict the possible states of the world under a series of events, then the cautious explanation might be most appropriate, while tasks like plan recognition usually require the best explanation. And for some problems there is no ordering of solutions, and any one would be acceptable.

Finally, it is possible that one explanation A will imply another A' in a simple causal theory. For example, *sun* and *warm* implies *sprinkler* in the sample theory. More generally, let $A_1 \vee A_2 \vee \cdots \vee A_n$ be any disjunction of explanations for O; A_1 is *independent for O in the theory Σ* if $\Sigma \cup A_1 \not\models A_2 \vee A_3 \vee \cdots \vee A_n$.

Using the example causal theory of the previous section, there are three explanations of $O = \{wet\text{-}lawn\}$, namely $\{rain\}$, $\{sprinkler\}$, and $\{sun, warm\}$. The cautious explanation is $rain \vee sprinkler \vee (sun \wedge warm)$, which simplifies in the domain theory to $rain \vee sprinkler$. The explanation $sun \wedge warm$ is not independent, since it implies *sprinkler*. The observation set $O = \{wet\text{-}road, wet\text{-}lawn\}$ has the single explanation $\{rain\}$, which is also its cautious explanation.

4 The Consistency Based Approach

The consistency based approach to explanation is fundamentally different from abduction in its formal specification. We start with a structure $\langle C, E, \Pi \rangle$ similar to a simple causal theory, except that the domain theory Π now gives necessary causal conditions for the occurrence of an effect; we comment on this more below. We call this theory an inverse causal theory.

DEFINITION 4.1 *Let $\langle C, E, \Pi \rangle$ be an inverse causal theory, and O a subset of E (the observation set). A* denial set for O *is a maximal subset $D \subseteq C$ such that*

$$\Pi \cup O \cup \{\neg d \mid d \in D\} \text{ is consistent.}$$

The complement of the denial set with respect to C $(C - D)$ is called a diagnosis for O.[2]

Because a denial set is subset-maximal, the diagnosis is a deductive consequence of the domain theory, the observations, and the denial set, as proven in [14]:

$$\Pi \cup O \cup \{\neg d \mid d \in D\} \vdash C - D\,. \tag{1}$$

The difference between the consistency based approach and abduction is twofold. First, the form of inference is distinct: rather than abducing causes that imply the observations O given the domain theory Σ, the consistency approach tries to minimize the extent of the causation set C by denying as many of its elements as possible. Second, these methods encode knowledge of the domain differently: in the abductive framework, there are implications from the causes to the effects, while in the consistency based systems, the most important information seems to be the implication from observations to possible causes. For example, in reconstructing the set-covering model of diagnosis, Reiter [14] uses axioms of the form:

$$\text{OBSERVED}(m) \supset \text{PRESENT}(d_1) \vee \cdots \vee \text{PRESENT}(d_n),$$

[2] In the original paper ([14]), a diagnosis was defined differently and then proven to be the complement of the denial set.

where m is the observed symptom and d_i are diseases that cause the symptom. These axioms give necessary conditions for the observation, namely, that one of a set of diseases be present. An inverse causal theory contains statements of this form. In fact, under certain conditions on the domain theory, diagnoses and explanations coincide.

THEOREM 4.1 (CONSOLE, POOLE)[3] *Let $\langle C, E, \Sigma \rangle$ be a simple causal theory over a propositional language, with Σ a set of nonatomic definite clauses whose directed graph is acyclic. Let C be a set of atoms that do not appear in the head of any clause of Σ, and E any set of atoms. Let Π be the Clark completion [1] of Σ. Then the diagnoses of $\langle C, E, \Pi \rangle$ are exactly the explanations of $\langle C, E, \Sigma \rangle$.*

The simple causal theory of Figure 1 does not satisfy the conditions, because it contains the equivalence $sun \equiv \neg rain$, and *sprinkler* is a cause that appears as the head of a clause. If we eliminate these anomalies, then the Clark completion of the domain theory is:

$$\begin{aligned} &wet\text{-}road \equiv rain \\ &wet\text{-}lawn \equiv rain \vee sprinkler \\ &sprinkler \equiv warm \wedge sun \end{aligned} \tag{2}$$

The diagnoses of *wet-lawn* are $\{rain\}$ and $\{sun, warm\}$; the explanation $\{sprinkler\}$ is missing.

For more complicated domain theories, Clark completion does not give the required closure over explanations. If the theory has cycles, for example $\{a \supset b, a' \supset b, b \supset a\}$, then the completion will only pick out a subset of the explanations (in this case, $b \equiv a$). If there is disjunction in the head of a clause, the completion is undefined.

In the next few sections we will extend the scope of Theorem 4.1 by considering a more general notion of completion for a simple causal theory, that of explanatory closures.

[3]Neither of these authors states the theorem in this form, although Poole [10] is close. It is clear that the theorem follows from their results.

5 Explanatory closures

Let $\langle C, E, \Sigma \rangle$ be a simple causal theory, and suppose $g \in E$ has a cautious explanation $\bigvee_i A_i$. Now consider the statement

$$g \supset A_1 \vee A_2 \vee \cdots \vee A_n \,, \tag{3}$$

where we understand each A_i to be the conjunction of its elements. This expression says that whenever g is present, it must have been caused by one of the A_i; we call this expression the *explanatory closure* of g with respect to the simple causal theory $\langle C, E, \Sigma \rangle$; it is abbreviated $\gamma(g)$. If the explanatory closures of all effects E exist, then the theory $\langle C, E, \Pi \rangle$ formed by adding the closures to Σ is called the *closure* of $\langle C, E, \Sigma \rangle$.

By forming the closure of a causal theory we can deduce the cautious explanation from any given effect. One immediate question is whether we should add something stronger or weaker to close the theory. If we add a stronger closure, then we have excluded some original explanation from consideration; e.g., if the explanatory closure is $g \supset a_1 \vee a_2$, and we use $g \supset a_1$ instead, then a_1 is the only explanation for g. On the other hand, suppose instead we add $g \supset (a_1 \vee a_2 \vee \delta)$ for some arbitrary sentence δ. If we try to derive explanations by minimizing causes, then since $\neg(a_1 \vee a_2)$ is consistent with the closure, we could assume it, and derive δ as the "explanation" for g, which is certainly not intended.

Another question is whether explanatory closures are always consistent with the original causal theory, and if so, whether the original explanations remain unchanged. Unfortunately, the answer to both parts of this question is "no."

EXAMPLE 5.1 Let $\langle \{a_1, a_2, a_3\}, \{g_1, g_2, g_3\}, \Sigma \rangle$ be a simple causal theory, with Σ equal to the conjunction of

$$\begin{array}{lll} a_1 \wedge a_2 \supset g_1 & a_1 \wedge a_2 \supset g_2 \vee g_3 & \neg a_1 \vee \neg a_2 \vee \neg a_3 \\ a_2 \wedge a_3 \supset g_2 & a_2 \wedge a_3 \supset g_1 \vee g_3 & g_1 \vee g_2 \vee g_3 \\ a_3 \wedge a_1 \supset g_3 & a_3 \wedge a_1 \supset g_1 \vee g_2 & \end{array} .$$

The closures of this theory are

$$\begin{array}{l} g_1 \supset a_1 \wedge a_2 \\ g_2 \supset a_2 \wedge a_3 \\ g_3 \supset a_3 \wedge a_1 \,. \end{array}$$

It is easy to show that the conjunction of these closures is inconsistent with Σ.[4]

The technical conditions for inconsistency are somewhat complicated, and it takes some work to create a causal theory that will have inconsistent closures; e.g., the example of Figure 1 can be consistently closed, but it was not originally designed with this property in mind. The necessary conditions involve interacting effects and causes such that in the causal theory at least one of the effects is true, and one of the causes false. The following proposition states this more precisely.

PROPOSITION 5.1 *Let $\{\gamma(g_i) \mid 0 < i \leq n\}$ be a set of closures for $\langle C, E, \Sigma\rangle$. For each $i \leq n$, let p_i be either g_i or $\neg A_i$, where A_i is any explanation for g_i. Each sentence*

$$p_1 \vee p_2 \vee \cdots \vee p_n$$

must be a theorem of Σ for these closures to be inconsistent with Σ.

Proof. For the closures to be inconsistent, $\bigvee_i \neg(\gamma(g_i))$ must be a theorem of Σ. We have:

$$\begin{aligned} \bigvee_i \neg(\gamma(g_i)) &\equiv \bigvee_i (g_i \wedge \neg E_{g_i}) \\ &\equiv \bigvee_i (g_i \wedge \neg A_i^1 \wedge \neg A_i^2 \cdots) \end{aligned}$$

where E_{g_i} is the cautious explanation for g_i, and the A_i's are all explanations for it. The proposition follows by tautological consequence.

We now turn to the question of how adding closures can modify explanations.

EXAMPLE 5.2 Let $\langle\{a_1, a_3, a_4\}, \{g_1, g_2, g_3\}, \Sigma\rangle$ be a simple causal theory, with Σ equal to the conjunction of

[4] It was suggested by a reviewer that the sentence $g_1 \supset (a_1 \wedge a_2) \vee (a_2 \wedge a_3 \wedge \neg g_3) \vee (a_1 \wedge a_3 \wedge \neg g_2)$ and similar ones for the other effects be used; these closures are consistent with the domain theory. However, as noted above, this would generate an anomalous explanation: by asserting $\neg a_1$, we derive $a_2 \wedge a_3$, which is not an explanation for g_1.

$$a_1 \supset g_1 \qquad \neg a_1 \vee \neg a_3$$
$$a_1 \supset g_2 \qquad a_3 \supset g_1 \vee g_2$$
$$a_3 \supset g_3$$
$$a_4 \supset g_3$$

The closures of this theory are

$$g_1 \supset a_1$$
$$g_2 \supset a_1$$
$$g_3 \supset a_3 \vee a_4 .$$

If the first two closures are added to Σ, a_1 becomes true, a_3 becomes false, and the only explanation for g_3 is a_4.

This example shows that some causes may become true or false, thus modifying the available explanations. However, new explanations, which are not subsets of old ones, do not arise from the addition of closures.

PROPOSITION 5.2 *Let $\langle C, E, \Sigma \rangle$ be a simple causal theory, and $\{\gamma(g_i)\}$ a set of explanatory closures with respect to it. Suppose $\Pi = \Sigma \cup \{\gamma(g_i)\}$ is a consistent set. For an arbitrary effect g, every explanation of g w.r.t. Π is a subset of some explanation of g w.r.t. Σ.*

Proof. Assume A is an explanation for g w.r.t. Π, but there is no $A' \supseteq A$ such that A' is an explanation for g w.r.t. Σ. Using a technique similar to that of Proposition 5.1, the following must be theorems of Σ, where each p_i is either g_i or $\neg A_i$ for any explanation A_i of g_i:

$$p_1 \vee p_2 \vee \cdots \vee p_n \vee \neg A \vee g .$$

Choosing each p_i to be $\neg A_i$, this is a sentence which contradicts the original assumption.

5.1 Augmented Domain Theories

Rather than trying to determine if a causal theory has a consistent closure, we might find it useful to modify the theory so that it does. The simplest way to do this is to add an escape cause for each effect: a new cause r_i is

included in C for each g_i, and the sentence $r_i \supset g_i$ is added to Σ.[5] The new causes are sufficiently isolated from the original domain theory so that inconsistency cannot result. In effect, the closure conditions no longer force one of the original explanations for g_i to be true, since r_i is an alternative. Further, augmented theories do not change their original explanations at all when closures are added.

PROPOSITION 5.3 *Let $\langle C', E, \Sigma' \rangle$ be a simple causal theory formed from $\langle C, E, \Sigma \rangle$ by adding r_i to C and $r_i \supset g_i$ to Σ for each $g_i \in E$; call this an* augmented causal theory. *Suppose that $\{\gamma(g_i)\}$ is a set of explanatory closures with respect to the augmented theory, and let $\Pi = \Sigma' \cup \{\gamma(g_i)\}$. Then Π is consistent, and for an arbitrary effect g, a subset $A \subseteq C$ is an explanation of g w.r.t. Π if and only if it is an explanation of g w.r.t. Σ.*

Proof. By Proposition 5.1, if the closure of Σ' is to be inconsistent, the following must be theorems of Σ', where each p_i is either g_i or $\neg r_i$:

$$p_1 \vee p_2 \vee \cdots \vee p_n \ .$$

Because the only expressions containing r_i are of the form $r_i \supset g_i$, the above sentences are theorems of Σ' only if there are corresponding theorems of Σ with each $\neg r_i$ replaced by $\neg g_i$. This is impossible, since such a set is unsatisfiable.

Assume $A \subseteq C$ is an explanation for g w.r.t. Π, but not w.r.t. Σ. By reasoning similar to that in the proof of Proposition 5.3, the following must be theorems of Σ', where each p_i is either g_i or $\neg g_i$:

$$p_1 \vee p_2 \vee \cdots \vee p_n \vee \neg A \vee g \ .$$

By tautological consequence, these sentences imply $A \supset g$, contradicting the initial assumption.

[5] Escape causes are the same idea as the unknown faults of [4: 15].

5.2 Local Closure

Cautious explanations for a proposition g are defined by reference to the entire contents of the causal theory Σ. Is there a way of deriving these explanations in a local manner, that is, by looking only at the sentences of Σ in which g occurs? From Theorem 4.1, Clark completion works for a restricted language. But if arbitrary correlations are allowed in Σ, then adding cautious explanations by a local closure operation is not possible. The simplest example showing this contains loops in the implication structure; e.g., let Σ be

$$\begin{array}{lll} a \supset g & a' \supset b & b \supset g \\ g \supset c & c \supset b & \end{array} \quad . \tag{4}$$

Let a and a' be the causes. Adding the local closure $g \supset a \vee b$ is insufficient, because it is subsumed by $g \supset c \supset b$, so that a as a cause of g will never be inferred. Any local closure for g cannot find the connection between c and b, and thus has the chance of being incorrect.

Loops in the implication structure also cause problems for other global closure methods such as circumscription, which is equivalent to Clark completion for the restricted language [13]. In the case of the above example, minimizing g while holding the causes fixed yields $g \supset b$, which is again stronger than the explanatory closure.

6 Closure + Minimization Implies Abduction

The closure of a causal theory contains the explanatory closure

$$g \supset A_1 \vee A_2 \vee \cdots \vee A_n$$

of each effect g. Suppose the closed theory is consistent, and we observe g. Then $A_1 \vee A_2 \vee \cdots \vee A_n$ is true in all models of g and the closed theory. If we now try to minimize causes, that is, to assert $\neg A_i$ for as many explanations as possible, we will eliminate possible explanations from the disjunction, until we are left with a single one. Thus we can perform abductive reasoning in the consistency based approach.

There is one caveat to this reasoning:[6] if an explanation A_1 is not independent, then it will not be found by closure and minimization. Suppose there is another A_2 that is implied by A_1 and the domain theory; then A_1 will be shadowed from the minimization by A_2: we cannot assert $\neg A_2$ without concluding $\neg A_1$. Thus using closure and minimization will only produce the independent explanations.

This discussion is made more precise with the following proposition.

THEOREM 6.1 *Let $\langle C, E, \Sigma \rangle$ be a simple causal theory, and suppose that $\langle C, E, \Pi \rangle$, its closure, is consistent and does not entail an effect g. Let A be an explanation for g in Σ, and suppose that A is consistent with Π and independent in Π. Then A is a subset (not neccssarily proper) of some diagnosis for g in Π.*

Conversely, every diagnosis for g in Π is a superset (not necessarily proper) of some explanation for g in Σ.

Proof. Suppose A is an explanation of g (in Σ), and let X be the disjunction of the rest of g's explanations: $A_2 \vee A_3 \vee \cdots \vee A_n$. $\neg X$ is consistent with Π, or else $\Pi \models X$ and so $\Pi \models g$, contradicting the assumptions. Also, by assumption, $\Pi \cup A$ is consistent, and since A is independent in Π, $\Pi \cup A \cup \neg X$ is consistent, and hence so is $\Pi \cup \{g\} \cup \neg X$. Let m be a model of $\Pi \cup \{g\} + \neg X$, and let $D =$

[6] I am indebted to Eunok Paek for pointing out this problem.

$\{a_2, a_3, \ldots a_n\}$ be a set of elements, one from each explanation of X, that are false in m. Let $\sim D = (\neg a_2 \wedge \neg a_3 \wedge \ldots \neg a_n)$. Now $\Pi \cup \{g\} \cup \sim D$ is consistent, and because of the presence of the closure of g, A is a consequence of it. D can be extended to some maximal set D' that is a denial set for g, and its complement w.r.t. C contains A.

For the converse part, let $\Pi \cup \{g\} \cup \sim D$ be consistent for some denial (maximal) set $D \subseteq C$. Suppose the associated diagnosis H is not a superset of any explanation of g in Σ. Then for any explanation A_i of g, $\neg A_i$ is consistent with $\Pi \cup \{g\} \cup \sim D$, and so D by maximality must contain some element of each of A_i. Thus $\sim D \supset \neg E_g$, where E_g is the cautious explanation for g. This is a contradiction, since $g \supset E_g$ is a sentence of Π.

Remarks. This theorem shows the general correspondence between abductive and consistency methods. If an inverse causal theory is formed by closing a causal theory, then, with several restrictions, consistency based diagnoses and abductive based explanations are isomorphic to one another. The restrictions have to do with the problems encountered in adding closures to a causal theory; given the results of the last section, it may not be possible to do so consistently, as explanations may change, and so forth.

This is not to say that the two approaches are equivalent, however. The consistency based method in general entails more than the abductive one, as a consequence of adding the closures.

COROLLARY 6.2 *Same conditions as Theorem 6.1 above. For some denial set D, every consequence of Σ and A is a consequence of Π and D. On the other hand, some consequence of D and Π may not be consequences of Σ and A.*

In Example 5.2, $\neg a_3$ is a consequence of Π, but not of any abductive explanation for g_3.

6.1 Representational Issues

The results of this paper must be interpreted in light of both the domain information available and the task at hand. We will briefly examine three areas: temporal projection, plan recognition, and diagnosis.

Closure conditions may be given directly as part of the axiomatization of a domain. This is the case with many diagnostic tasks in which complete fault models [15] are given, characterizing the exact relation between the input/output behavior of a system and its internal states. A similar analysis can be given for the system of [8] on temporal projection, where the "internal state" is the set of event occurrences and the axiomatization specifies exactly what events must occur given a sequence of states, and vice versa.

On the other hand, often one has information about causal effects, together with some noncausal correlations (e.g., forbidden states) and would like to generate explanatory diagnoses. In order to employ the consistency based approach, the explanatory closures must be generated and added. Here the form of the causation axioms can be exploited. If they are horn, definite and acyclic, then local closure (Clark completion) can be used. For more complicated theories, a technique such as circumscription may be appropriate. An example here is the theory of plan recognition in [5]. The domain theory is a hierarchical set of actions; the causes are the goals at the highest level of the hierarchy, the so-called *END* events. Relations between actions at different levels in the hierarchy are given by a first-order domain theory. Circumscription is used to close off the axioms, producing the explanatory closure axioms. Given a set of observed actions, minimizing over the *END* events produces an explanation of the observations.[7]

Another good example of the derivation of closure axioms is from the theory of temporal projection in [6]; in effect, this theory is similar to that of [8] above, with the following differences. First, the sequence of actions is fully specified by the *result* function, but exceptions to the actions are allowed, in the form of miracles: these are the assumable atoms. Second, there is a theory of causation for action types, which is used to generate the closure conditions by circumscription. That is, the causation axioms state what must follow if the preconditions of an action hold and the action takes place; circumscription then generates the closure axioms. Minimization over miracles gives the desired explanations.

The motivation behind the multiple circumscription in systems such as [5; 6] has often been obscure. Given the results of this paper, it should be clear that the circumscriptions are performing abduction by using closure and minimization. Whether the circumscription corresponds to an appropri-

[7]This account is of necessity somewhat simplified.

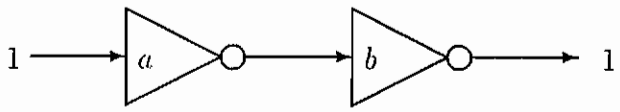

Figure 2: A double inverter

ate closure can be tested by checking whether it produces the explanatory closure axioms, and whether adding these axioms changes the set of explanations. The examples and propositions of Section 5 should be helpful in this regard; for example, by adding escape causes it is always possible to retain the original causal structure. In general, if there are cycles in the implication structure of the causal domain theory, then neither circumscription nor local closure will work correctly in generating explanatory closures.

6.2 Logic Based Diagnosis

In logic based diagnosis methods derivative of [14], the domain theory takes on a restricted form, with a distinguished set of abnormality predicates ab_i used to describe the expected behavior of a system. For explanatory diagnosis, a complete fault model is required, so that both normal and abnormal behaviors are fully specified. For example, consider the double inverter of Figure 2. The domain axioms are

$$\begin{aligned} &\neg ab_i \supset (in_i \equiv \neg out_i) \\ &ab_i \supset (in_i \equiv out_i) \\ &out_a \equiv in_b \ . \end{aligned} \tag{5}$$

Each inverter can either have normal behavior, or have a short circuit so that input and output are the same. Let $C = \{ab_i\}$. If we observe $\{in_a, out_b\}$, then there is one diagnosis, the empty set. To compare this to the abductive approach, we must modify the definition of explanation slightly, allowing $\neg ab_i$ as nonatomic causes.[8] Let $C' = C \cup \{\neg c \mid c \in C\}$, and define a *default explanation* of O in a theory $\{C', E, \Sigma\}$ to be an explanation of O minimal in C-atoms, restricted to just these atoms. In this example, the explanations are $\{\neg ab_a, \neg ab_b\}$ and $\{ab_a, ab_b\}$; the first of these is minimal in ab-atoms,

[8] We could also change the vocabulary and add $ok_i \equiv \neg ab_i$ as a new set of causes, and then use Theorem 6.1. But this would change the diagnoses in logic based diagnosis, by eliminating the distinction between normal and abnormal behavior.

and the default explanation is the empty set, i.e., the diagnosis. In fact, we can show in general that the default explanations correspond to diagnoses.

THEOREM 6.3 *Let $C' = C \cup \{\neg c \mid c \in C\}$, and let $\langle C', E, \Sigma\rangle$ be a simple causal theory. Suppose Π, the explanatory closure of Σ, is consistent and does not entail an effect g. Let A be a default explanation for g in Σ, and suppose that A is consistent with Π and independent in Π. Then A is a subset (not necessarily proper) of some diagnosis for g in $\langle C, E, \Pi\rangle$.*

Conversely, every diagnosis for g in $\langle C, E, \Pi\rangle$ is a superset (not necessarily proper) of some default explanation for g in $\langle C', E, \Sigma\rangle$.[9]

Proof. Consider the theory Π' formed from Π by adding $c' \equiv \neg c$ for each $c \in C$. The diagnoses of $\langle C, E, \Pi\rangle$ are just the diagnoses of $\langle C \cup \{c'\}, E, \Pi'\rangle$ that are minimal in C, keeping only atoms in C. Similarly, default explanations of $\langle C', E, \Sigma\rangle$ are just explanations of $\langle C \cup \{c'\}, E, \Sigma'\rangle$ that are minimal in C, keeping only atoms in C. The result follows from the connection between diagnoses of $\langle C \cup \{c'\}, E, \Pi'\rangle$ and explanations of $\langle C \cup \{c'\}, E, \Sigma'\rangle$ given by Theorem 6.1.

In general, diagnoses assume more than is really needed for an explanation: hence the necessity of subset/superset relations in the correspondence theorems. As an example, consider the three unconnected inverters of Figure 3. Suppose the basic axioms governing the inverters are the same as before; there are no connections between the inverters, and the faults are coupled by the axiom $ab_a \supset (ab_b \vee ab_c)$. If we observe $\{in_a, out_b\}$, there are two diagnoses, $\{ab_a, ab_b,\}$ and $\{ab_a, ab_c\}$. There is only one default explanation, namely $\{ab_a\}$. The diagnoses all give a complete state of the system, whereas the default explanation is only the set of abnormal causes that account for the observed behavior. The abductive approach distinguishes between direct causes of the observations and irrelevant causes, while the consistency based approach does not.

[9] While default explanations correspond to diagnoses, explanations themselves correspond to the kernel diagnoses of [3], which cover the set of possible diagnoses. We do not prove this result here because of space limitations.

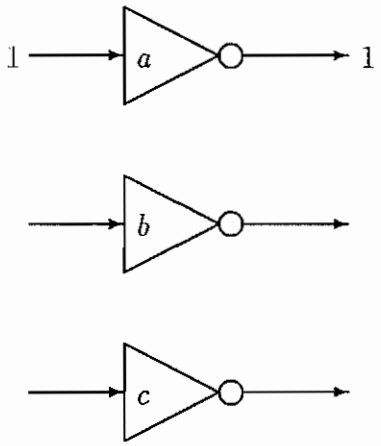

Figure 3: An unconnected circuit

7 Conclusion

We have shown how to extend the correspondence between abductive and consistency based methods to the case of causal theories that have arbitrary first-order relations between causes and effects. The correspondence requires that a domain theory expressing how causes produce effects be closed, that is, contain statements that the only causes are the known ones. The appropriate closure axioms are identified in this paper as explanatory closures. The main result of the paper is that minimization of causes in the closed theory produces almost the same explanations as abduction in the original causal theory. The caveat is that the abductive explanations are generally weaker than their consistency based counterparts. There are two reasons for this: adding closures may change the available explanations; and the consistency based method can conclude causes that are intuitively irrelevant to the observed behavior.

If one is interested in the representation of domain knowledge, then the abductive approach offers several advantages. It does not require the assumption of complete knowledge of causation, and it is not necessary to assert the explanatory closures. Adding the closures can lead to inconsistency and change the available explanations (although it will not add new ones). The computational aspect of adding closures is also discouraging, since there is no general local method that accomplishes the addition. Stronger global

methods such as circumscription will work only in special circumstances.

8 Acknowledgments

I would like to thank David Poole, Eunok Paek, Oskar Dressler, and Nicolas Helft for many helpful discussions on the evolving paper.

References

[1] K. Clark, Negation as failure, in: *Logic and Data Bases* (Plenum Press, New York, 1978) 293–322.

[2] L. Console, D. T. Dupre, and P. Torasso, Abductive reasoning through direct deduction from completed domain models, in: Z. W. Ras, ed., *Methodologies for Intelligent Systems 4* (North-Holland, New York, 1988) 175–182.

[3] J. de Kleer, A. Mackworth, and R. Reiter, Characterizing diagnoses, in: *Proceedings of the American Association of Artificial Intelligence*, Boston, MA (1990).

[4] J. de Kleer and B. C. Williams, Diagnosis with behavioral modes, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI (1989).

[5] H. Kautz, A formal theory for plan recognition, Technical Report TR–215, University of Rochester (1987).

[6] V. Lifschitz and A. Rabinov, Miracles in formal theories of action, *Artificial Intelligence* **38** (2) (1989).

[7] J. McCarthy, First order theories of individual concepts and propositions, in: B. Meltzer and D. Michie, eds., *Machine Intelligence 9* (Edinburgh University Press, Edinburgh, 1979) 120–147.

[8] L. Morgenstern and L. A. Stein, Why things go wrong: A formal theory of causal reasoning, in: *Proceedings of the American Association of Artificial Intelligence*, Minneapolis, MN (1988).

[9] D. Poole, Representing knowledge for logic-based diagnosis, in: *Proceedings of the International Conference on Fifth Generation Computing Systems*, Tokyo (1988) 1282–1290.

[10] D. Poole, A methodology for using a default and abductive reasoning system, Technical report, Department of Computer Science, University of Waterloo, Waterloo, Ontario (1988).

[11] D. Poole, Normality and faults in logic-based diagnosis, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI (1989).

[12] J. A. Reggia, D. S. Nau, and Y. Wang, A formal model of diagnostic inference I. Problem formulation and decomposition, *Inf. Sci.* **37** (1985).

[13] R. Reiter, Circumscription implies predicate completion (sometimes), in: *Proceedings of the American Association of Artificial Intelligence*, Pittsburgh, PA (1982) 418–420.

[14] R. Reiter, A theory of diagnosis from first principles, *Artificial Intelligence* **32** (1987).

[15] P. Struss and O. Dressler, Physical negation – integrating fault models into the general diagnostic engine, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI (1989) 1318–1323.