Artificial Intelligence 53 (1992) 255–272
Elsevier

Research Note

Abduction versus closure in causal theories

Kurt Konolige

Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA

Received September 1990
Revised April 1991

Abstract

Konolige, K., Abduction versus closure in causal theories (Research Note), Artificial Intelligence 53 (1992) 255–272.

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. The abductive method, on the other hand, treats explanations as sentences that, when added to the background theory, derive the observations. We show that there is a close connection between these two formalizations in the context of simple causal theories: domain theories in which a set of sentences are singled out as the explanatorily relevant causes of observations. There are two main results, which show that (with certain caveats) the consistency based approach can emulate abductive reasoning by adding closure axioms to a causal theory; and that abductive techniques can be used in place of the consistency based method in the domain of logic based diagnosis. It is especially interesting that in the latter case, the abductive techniques generate only relevant explanations, while diagnoses may have irrelevant elements.

1. Introduction

Reasoning to the best explanation is a 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 [9, 11]. 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 for 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 logic based theory of diagnosis [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 domain theory. The special predicate $ab(c)$ is used to state that the component c is abnormal or not functioning correctly. The observed behavior of the system is represented by a set of sentences. A diagnosis of the behavior is a minimal set of abnormality assumptions that is consistent with the observations and the domain 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 diagnostic tasks. 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 [4, 15]. 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 diagnostic tasks, and the answers it produces may have elements that are not relevant to a causal explanation.

A second area that we address is whether abductive methods may be used in the setting of logic based diagnosis with fault models. This area is closely related to the previous one, except that we assume that there is already a closed theory, and that the causes take on a specific form, namely normality and abnormality assumptions about components. Our main result here is that the abductive method produces kernel diagnoses,[1] but without any of the irrelevant causes that may be present in the latter.

The next three sections of this note describe simple causal theories, and define abductive and consistency based methods in this context. Section 5 develops the concept of explanatory closures, and Section 6 gives the main results on emulating abduction with the consistency based method. Section 7 describes how abductive methods can perform logic based diagnosis.

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.

[1] This term is from [3], and is defined in Section 7.

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 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 $\mathcal{L}$, are the *causes*;
- E, a set of sentences of $\mathcal{L}$, are 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 the conjunction of some subset of C.[2]

Effects E are those aspects of the domain that we might observe, and for 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.

Here is a simple causal theory that will be used as an example in the rest of the paper; a graphical presentation appears in Fig. 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:	$rain \supset wet\text{-}road$, $rain \supset wet\text{-}lawn$, $sun \equiv \neg rain$
	$sprinkler \supset wet\text{-}lawn$, $sun \wedge warm \supset sprinkler$.

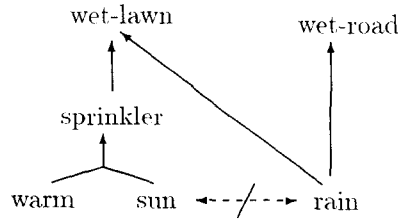

Fig. 1. A sample causal theory.

[2] 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.

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

$$A \vee B \quad \text{for} \quad (a_1 \wedge a_2 \wedge \cdots) \vee (b_1 \wedge b_2 \wedge \cdots),$$

$$\neg A \quad \text{for} \quad \neg(a_1 \wedge a_2 \wedge \cdots).$$

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 2. Let $\langle C, E, \Sigma \rangle$ be a simple causal theory. An *abductive explanation* (or ABE) 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 over sets satisfying the first two conditions.

If O has a nonzero finite number of ABE's 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 that there is no other set of causes for O consistent with the domain theory that is a proper subset. This is a relevancy condition, since it excludes from the explanation[3] elements of C that are not relevant to deriving O. 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 after a sequence 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.

[3] A note on terminology: we will use the simple term *explanation* to refer to abductive explanations when no ambiguity is possible, and *abductive explanation* or ABE when we want to distinguish them from explanations derived by the consistency based approach.

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 $\Sigma \cup A_1 \not\vdash A_2 \vee A_3 \vee \cdots \vee A_n$.

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 has been most clearly developed in the domain of diagnosis, especially in [3, 14]. In this section we will modify the terminology slightly to apply to the more general causal theories of Section 2, and to make comparison to the abductive approach easier. The particulars of the diagnostic task are discussed later, in Section 7.

Definition 3. Let $\langle C, E, \Sigma \rangle$ be a simple causal theory, and O (the observations) a subset of E. A *denial set for O* is a subset $D \subseteq C$ such that

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

When a denial set is maximal (that is, there is no other denial set that contains it), no more negative causes can be consistently added to it, and thus it is possible to deduce a set of (positive) causes from the maximal denial set, the domain theory, and the observations [14]:

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

In this case, we call $C - D$ a *consistency based explanation for O*, or CBE for short. CBE's are called *diagnoses* in Reiter's original paper, and they were used mainly for producing excusing diagnoses in the domain of electronic circuits. However, Reiter recognized that the consistency based approach was also capable of producing explanatory diagnoses, as long as the domain theory contained implications from effects to causes. For example, in reconstructing the set-covering model of (explanatory) diagnosis, Reiter used axioms of the form:

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

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. Later papers refer to implications of this sort as fault models [4, 15].

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, if we want to derive explanatory rather than excusing diagnoses, the most important information seems to be the implication from observations to possible causes.

Despite these differences, it is known that, under certain conditions on the domain theory, abductive and consistency based explanations coincide.

Theorem 4 (Console, Poole).[4] *Let* (C, E, Σ) *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 CBEs of* $\langle C, E, \Pi \rangle$ *are exactly the ABEs of* $\langle C, E, \Sigma \rangle$.

The simple causal theory of Fig. 1 does not satisfy the conditions, because it contains the equivalence *sun* $\equiv$ *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:

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

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

For more complicated domain theories, Clark completion does not give the required closure over abductive 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 abductive explanations (in this case, $b \equiv a$). If there is disjunction in the head of a clause, the completion is undefined.

In the next sections we will extend the scope of Theorem 4 by considering a more general notion of completion for a simple causal theory, that of explanatory closures.

[4] 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. Poole's theorem as stated seems to have a broader application, but personal correspondence with him disclosed that the conditions of application are as given here.

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 abductive 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 will be the only CBE for g. On the other hand, suppose instead we add something weaker, e.g., $g \supset (a_1 \vee a_2 \vee \delta)$ for some arbitrary sentence δ. If we try to derive CBEs 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 abductive explanations remain unchanged. Unfortunately, the answer to both parts of this question is "no".

Example 5. 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

$$a_1 \wedge a_2 \supset g_1 , \quad a_1 \wedge a_2 \supset g_2 \vee g_3 , \quad \neg a_1 \vee \neg a_2 \vee \neg a_3 ,$$

$$a_2 \wedge a_3 \supset g_2 , \quad a_2 \wedge a_3 \supset g_1 \vee g_3 , \quad g_1 \vee g_2 \vee g_3 ,$$

$$a_3 \wedge a_1 \supset g_3 , \quad a_3 \wedge a_1 \supset g_1 \vee g_2 .$$

The closures of this theory are

$$g_1 \supset a_1 \wedge a_2 ,$$

$$g_2 \supset a_2 \wedge a_3 ,$$

$$g_3 \supset a_3 \wedge a_1 .$$

It is easy to show that the conjunction of these closures is inconsistent with Σ.[5] 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 Fig. 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 6. *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 abductive 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_is are all abductive explanations for it. The proposition follows by tautological consequence. □

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

Example 7. 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

$$\begin{aligned} &a_1 \supset g_1, \qquad \neg a_1 \vee \neg a_3, \\ &a_1 \supset g_2, \qquad a_3 \supset g_1 \vee g_2, \\ &a_3 \supset g_3, \\ &a_4 \supset g_3. \end{aligned}$$

The closures of this theory are

[5] 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 $a_2 \wedge a_3 \wedge \neg g_3$, which, although it derives g_1 in the domain theory, is not an abductive explanation for g_1.

$$g_1 \supset a_1 ,$$

$$g_2 \supset a_1 ,$$

$$g_3 \supset a_3 \vee a_4 .$$

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 abductive explanations. However, no truly "new" explanations are created by the addition of closures, since every explanation must be the subset of one of the original ones.

Proposition 8. *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 abductive explanation of g w.r.t. Π is a subset of some abductive 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 6, 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.[6] 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 abductive explanations for g_i to be true, since r_i is an alternative. Further, augmented theories do not change their original abductive explanations at all when closures are added.

Proposition 9. *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*

[6] Escape causes are the same idea as the unknown faults of [4, 15].

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* $A \subseteq D$ *is an abductive explanation of* g *w.r.t.* Π *if and only if it is an abductive explanation of* g *w.r.t.* Σ.

Proof. By Proposition 6, 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 abductive explanation for g w.r.t. Π, but not w.r.t. Σ. By reasoning similar to that in the proof of Proposition 8, 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. □

5.2. *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, 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

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

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 abductive 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:[7] if an abductive 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 abductive explanations.

This discussion is made more precise with the following theorem.

Theorem 10. *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 ABE 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 CBE for g in Π.*

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

Proof. Suppose A is an ABE of g (in Σ), and let X be the disjunction of the rest of the ABEs of g: $A_2 \vee A_3 \vee \cdots \vee A_n$. $\neg X$ 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 $\Pi \cup \{g\} \cup \neg X$, and let $D = \{a_2, a_3, \ldots, a_n\}$ be a set of elements, one from each ABE of X, that are false in m. Let $\sim D = (\neg a_2, \neg a_3, \ldots, \neg a_n)$. 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 maximal denial set $D \subseteq C$. Suppose the associated CBE H is not a superset of any ABE of g in Σ. Then for any ABE 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

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

$\sim D \supset \neg E_g$, 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 and abductive 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 abductive explanations may change, and so forth.

This 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 11. *Assume the same conditions as in Theorem* 10 *above. For some maximal 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 a consequence of* Σ *and A.*

In Example 7, $\neg a_3$ is a consequence of Π, but not of any abductive explanation for g_3.

6.1. Representational issues

The designer of a domain theory who wishes to employ the consistency based approach can use the results of the previous sections to help determine how to formalize the domain. If the closure conditions are given directly as part of the available knowledge of the domain, then the consistency based approach may be used to generate explanations. For example, in the system of Morgenstern on temporal projection [8], the axiomatization gives a closed causal theory, since it specifies exactly what events must occur given a sequence of states, and vice versa.

On the other hand, often the designer only has information about causal effects, together with some noncausal correlations (e.g., forbidden states). 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. A good example from the domain of temporal projection is the work of Lifschitz [6]; in effect, this theory is similar to that of Morgenstern 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.

Another good example of the derivation of closure axioms by circumscriptive techniques is Kautz's theory of plan recognition [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* event produces an explanation of the observations.[8]

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 appropriate closure can be tested by checking whether it produces the explanatory closure axioms, and whether adding these axioms changes the set of abductive 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.

7. Logic based diagnosis

In the previous sections we considered how it was possible to derive explanations in a causal theory using consistency based methods. Here we consider the converse question in a more particular setting: can abductive methods be used to perform logic based diagnosis? In logic based diagnosis, 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 example, consider the double inverter of Fig. 2. The proposition in_i means

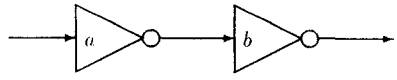

Fig. 2. A double inverter.

[8] The accounts of these systems are of necessity somewhat simplified, but the basic structure is correct.

that the input of device i is a logical 1, and $\neg in_i$ that it is a 0 (similarly for out_i and output). The domain axioms are

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

for $i = a, b$. Each inverter can either have normal behavior, or have a short circuit so that input and output are the same. In this example, both normal and abnormal behaviors of the system are fully specified, that is, there is an exhaustive fault model. It has been recognized that exhaustive fault models are required for explanatory diagnosis; in our terms, the domain theory contains the explanatory closures for all possible input/output behaviors.

In the most recent formulation of logic based diagnosis [3], the presence of exhaustive fault models has made it necessary to modify Reiter's original definition of diagnoses. Let $\langle C, E, \Sigma \rangle$ be a simple causal theory, with C containing a set of abnormality predicates and their negations.[9] In the example above, $C = \{ab_a, ab_b, \neg ab_a, \neg ab_b\}$.

A *partial diagnosis* G for an observation set $O \subseteq E$ is a subset of C that is consistent with Σ and O, such that every noncontradictory way of extending G with elements of C is also consistent with Σ and O. For example, for $O = \{in_a, out_b\}$, there are two partial diagnoses, $\{\neg ab_a, \neg ab_b\}$ and $\{ab_a, ab_b\}$. Only the first of these would be considered as a diagnosis in the original theory.

A *kernel diagnosis* is a subset-minimal partial diagnosis. For $O = \{in_b, out_b\}$, there are three partial diagnoses, $\{ab_b\}$, $\{ab_b, ab_a\}$, and $\{ab_b, \neg ab_a\}$; the first of these is subset-minimal, and so is a kernel diagnosis. Note that kernel diagnoses eliminate some of the irrelevancy present in partial diagnosis, since the state of inverter a is not relevant to the observed behavior.

We would like to know if we can produce the same kernel diagnoses with the abductive approach. The following theorem shows that this is possible, except (as in the case of Theorem 10) that the ABEs are, in general, more compact than kernel diagnoses.

Theorem 12. *Let $\langle C, E, \Sigma \rangle$ be a simple causal theory, with C a set of abnormality predicates and their negations, and Σ closed and consistent. Let A be an independent ABE for g in Σ. Then A is a subset (not necessarily proper) of some kernel diagnosis for g.*

Conversely, every kernel diagnosis for g is a superset (not necessarily proper) of some ABE for g.

[9] We could also change the vocabulary and add $ok_i \equiv \neg ab_i$ as a new set of causes. But we are trying to stay as close as possible to the logic based diagnosis terminology.

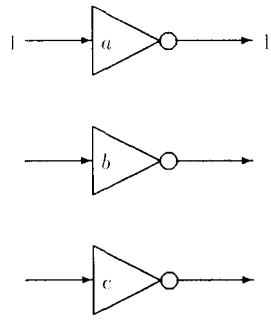

Fig. 3. An unconnected circuit.

Proof. Let A be an independent ABE for g. Using an argument similar to the first part of the proof of Theorem 10, construct the set D, such that $\Sigma \cup \{g\} \cup \sim D \vdash A$. Now $\sim D \cup A$ is a subset of some partial diagnosis, and therefore A must be a subset of some kernel diagnosis, since $\Sigma \cup \sim D \cup \neg A$ is inconsistent.

For the converse part, let K be a kernel diagnosis for g. Assume no ABE A_i is contained in K. Then K can be consistently extended by adding one negated element from each A_i, thus contradicting $g \supset E_g$, which must be in Σ. □

In general, although kernel diagnoses get rid of some irrelevancies with respect to their corresponding ABEs, they still assume more than is really needed for an explanation: hence the necessity of subset/superset relations in the correspondence theorem. As an example, consider the three unconnected inverters of Fig. 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 $\{in_a, out_a\}$, there are two kernel diagnoses, $\{ab_a, ab_b\}$ and $\{ab_a, ab_c\}$. There is only one ABE, namely $\{ab_a\}$. The kernel diagnoses must reflect the constraint that one of b or c is abnormal, which is not a relevant factor in explaining the observations. The abductive explanation is just the set of 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. On the other hand, the information that either b or c is abnormal is still derivable in the abductive system, as a consequence of the domain theory and the ABE; but it is not a part of the explanation itself.

8. 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 abductive 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 abductive 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.

In logic based diagnosis, using an abductive method may be appropriate if it is important to distinguish causes relevant to producing the observations from those that are mere side-effects. But one must be careful here in giving too much weight to the term "causally relevant", since the simple causal theories we have introduced give only a very narrow interpretation of the complex concept of causation.

Acknowledgement

I would like to thank David Poole, Eunok Paek, Oskar Dressler, and Nicolas Helft for many helpful discussions on the evolving draft. The two anonymous referees contributed many useful comments that I have used in revising the paper. The research reported here was supported partially by the NTT Corporation, and partially by the Office of Naval Research under Contract No. N00014-89-C-0095.

References

[1] K. Clark, Negation as failure, in: *Logic and Data Bases* (Plenum, 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 AAAI-90*, Boston, MA (1990).

[4] J. de Kleer and B.C. Williams, Diagnosis with behavioral modes, in: *Proceedings IJCAI-89*, Detroit, MI (1989).

[5] H. Kautz, A formal theory for plan recognition, Tech. Report TR-215, University of Rochester (1987).
[6] V. Lifschitz and A. Rabinov, Miracles in formal theories of action, *Artif. Intell.* **38** (2) (1989) 225–237.
[7] J. McCarty, 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 AAAI-88*, St. Paul, MN (1988) 518–523.
[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, Tech. Report, Department of Computer Science, University of Waterloo, Waterloo, Ontario (1988).
[11] D. Poole, Normality and faults in logic-based diagnosis, in: *Proceedings IJCAI-89*, 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 AAAI-82*, Pittsburg, PA (1982) 418–420.
[14] R. Reiter, A theory of diagnosis from first principles, *Artif. Intell.* **32** (1987) 57–95.
[15] P. Struss and O. Dressler, Physical negation – integrating fault models into the general diagnostic engine, in: *Proceedings IJCAI-89*, Detroit, MI (1989) 1318–1323.