Specifying causality in action theories: A default logic approach ⋆

Yan Zhang [1]

School of Computing and Information Technology
University of Western Sydney, Nepean
Kingswood, NSW 2747, Australia

Abstract

Recent research on reasoning about action has shown that the traditional logic form of domain constraints is problematic to represent ramifications of actions that are related to causality of domains. To handle this problem properly, as proposed by some researchers, it is necessary to describe causal relations of domains explicitly in action theories. In this paper, we address this problem from a new point of view. Specifically, unlike other researchers viewing causal relations as some kind of inference rules, we distinguish causal relations between defeasible and non-defeasible cases. It turns out that a causal theory in our formalism can be specified by using Reiter's default logic. Based on this idea, we propose a causality-based minimal change approach for representing effects of actions, and argue that our approach provides more plausible solutions for the ramification and qualification problems compared with other related work. We also describe a logic programming approximation to compute causal theories of actions which provides an implementational basis for our approach.

1 Introduction

A sentence like "heavy rainfall *normally causes* floods" represents some causal relation between "heavy rainfall" and "floods". Naturally, such causal relation can be used to represent some causal prediction. For example, if we know that there is a heavy rainfall, we would expect to have floods. From a semantical point of view, a causal relation is different from a logical implication. Suppose

⋆ This research is supported in part by a grant from University of Western Sydney, Nepean.

[1] Email address: yan@cit.nepean.uws.edu.au

we use logical formulas to represent causal relations, then the causal relation "heavy rainfall causes floods" would be likely represented by a logical formula *heavy-rainfall* $\supset floods$. Therefore, if *heavy-rainfall* is true, then from this formula, it follows that $floods$ is true. However, note that *heavy-rainfall* $\supset floods$ is also logically equivalent to the formula $\neg floods \supset \neg$*heavy-rainfall*. It is also the case that $\neg$*heavy-rainfall* follows from $\neg floods$. But it would not be normally regarded that no flooding is a *cause* of the absence of rain.

In many of current formulations of reasoning about action and change, properties about the domain are normally characterized by a set of logical formulas called *domain constraints*, and there is no difference between general logical constraints and causal relations. Any changes to the state of the world must be consistent with domain constraints. However, as we will see next, ignoring such difference may result in difficulties in describing effects of actions.

Consider the simplified stuffy room example of Ginsberg and Smith [6]. There is a constraint about the room which says that if both ducts in the room are blocked, then the room becomes stuffy. This constraint is expressed by the following formula:

$$blocked(duct_1) \wedge blocked(duct_2) \supset stuffy(room). \tag{1}$$

Suppose the initial state S_0 is given by the following:

$$S_0 = \{\neg blocked(duct_1), blocked(duct_2), \neg stuffy(room)\}.$$

We will define *state* formally in section 3. At the moment, it is sufficient to think that a state is simply an interpretation of a first order language.

Now consider an action of making $duct_1$ blocked (i.e. the effect of the action is $blocked(duct_1)$). To predict the state resulting from the performance of this action, we use Winslett's PMA [18] where a *minimal change principle* is employed for modeling state change. Informally, the minimal change principle in the PMA says that *under the condition of satisfying domain constraints and the direct effect of the action, the difference between the possible resulting state and the initial state should be minimal with respect to set inclusion.* Therefore, we get the following two possible resulting states:

$$S_1 = \{blocked(duct_1), blocked(duct_2), stuffy(room)\},$$
$$S_2 = \{blocked(duct_1), \neg blocked(duct_2), \neg stuffy(room)\}.$$

Note that both S_1 and S_2 satisfy the constraint (1) and the effect of the action $blocked(duct_1)$, and have minimal differences with the initial state S_0. But intuitively, only S_1 is our desired resulting state, while S_2 is quite implausible in some sense — why does the performance of action to make $duct_1$ blocked imply an indirect effect of resulting in $duct_2$ unblocked? This is the so-called *ramification problem* – the problem of how to derive the indirect effects of actions.

In fact, constraint (1) is also logically equivalent to the following formula

$$blocked(duct_1) \wedge \neg stuffy(room) \supset \neg blocked(duct_2). \quad (2)$$

It is clear now that both S_1 and S_2 are forced by (1) and (2) respectively to have minimal differences with S_0 with respect to the effect of action $blocked(duct_1)$. But semantically, (1) should not simply be a logical implication. It should represent a causal relation between $blocked(duct_1)$, $blocked(duct_2)$ and $stuffy(room)$ – the fact that $duct_1$ and $duct_2$ blocked *causes* the fact that the room becomes stuffy, which, for reasons similar to the sentence "heavy rainfall *causes* floods", is *not* logically equivalent to formula (2).

In fact, the problem of ignoring the difference between the causal relation and logical constraint in action theories has been realized by some researchers. It has been argued that to describe the indirect effect of actions properly, it is necessary to specify the causal relation of the domain explicitly. More recently, Brewka and Hertzberg, McCain and Turner, and Lin have independently developed alternative action theories to solve this problem [2,13,11]. In their action theories, causality is explicitly taken into account. In particular, both Brewka and Hertzberg's and McCain and Turner's approaches view causal relations as kinds of inference rules, while Lin's method employs a specific predicate $Caused$ to represent causality in the circumscription-based action theory.

However, as we will argue later in this paper, all of their methods, more or less, suffer from some restrictions and hence are not entirely satisfactory. The goal of this paper is to propose a new formal approach to specify causality in action theories, and show how our approach provides some new insights for solving the ramification and qualification problems by combining causal relations into action theories.

The paper is organized as follows. Section 2 defines causal relations and causal theories formally, and uses Łukaszewicz's default model theory to provide formal semantics for causal theories. Section 3 proposes a new principle for reasoning about action called the *causality-based* minimal change principle. A formulation of actions based on this principle is presented. Section 4 illustrates further examples to show the application of our method in reasoning about action. Section 5 extends our approach to deal with sequence of actions. Section 6 compares our method with related work in detail. Section 7 describes a logic programming approximation to compute causal theories of actions that provides an implementational basis for our method. Finally, section 8 concludes the paper with some remarks.

2 Semantics for causal theories

Let $\mathcal{L}$ be a finite propositional language with two additional symbols $\leadsto$ and $\Rightarrow$ which we will use to express causal relations. A *causal relation* is any expression of the form $A \leadsto B$ or $A \Rightarrow B$, where A and B are propositional formulas of $\mathcal{L}$. Specifically, an expression $A \leadsto B$ is read "the fact that A *normally causes* the fact that B", while an expression $A \Rightarrow B$ is read "the fact that A *causes* the fact that B". For both cases, we will call A is a *cause* of $A \leadsto B$ (or $A \Rightarrow B$) or B is an *effect* of $A \leadsto B$ (or $A \Rightarrow B$).

The above two forms of expressions $A \leadsto B$ and $A \Rightarrow B$ represent *defeasible* and *non-defeasible* causal relations respectively in our problem domain. For example, consider a causal relation which says that "the fact that a match has been struck normally causes the fact that it has lit". However, if the match is wet and we know that striking a wet match normally does not light, then this causal relation should not be triggered during our reasoning. So this causal relation is defeasible. On the other hand, a statement that "the fact that some one is dead causes the fact that this person is not walking" should not be defeasible from our commonsense.

A *causal theory* CT is defined as a pair $(\mathcal{A}, \mathcal{C})$, where $\mathcal{A}$ is a set of propositional formulas of $\mathcal{L}$ and $\mathcal{C}$ a set of causal relations. Note that $\mathcal{C}$ includes both forms of causal relations as we described earlier. To simplify our formulation, we restrict both $\mathcal{A}$ and $\mathcal{C}$ to be finite. Intuitively, a causal theory represents knowledge about a domain while $\mathcal{A}$ is viewed as basic facts or/and constraints of the domain, and $\mathcal{C}$ represents the causality about this domain.

Now we try to provide a model theoretic semantics for the causal theory. In fact, we can treat our causal relations as defaults in Reiter's default logic [17] under the propositional case. In particular, a defeasible causal relation $A \leadsto B$ can be viewed as a normal default $A : B/B$, while a non-defeasible causal relation $A \Rightarrow B$ is viewed as a default $A : /B$ (i.e. a default without justification). Therefore, a causal relation like *heavy-rainfall* $\leadsto$ *floods* can be also interpreted as "if there is a heavy rainfall and it is consistent to believe that there are floods, then one may believe that there are floods caused by a heavy rainfall". Hence, in our formalism, causal reasoning turns out to be a kind of default reasoning, and a causal theory corresponds to a propositional default theory [17].

Treating causal theories as propositional default theories, we then can use the model default theory originally proposed by Etherington and extended by Lukaszewicz [12] to provide formal semantics for our causal theories.

Let Φ be a class of interpretations of language $\mathcal{L}$ and A a propositional formula. $\Phi(A) = \{\varphi \mid \varphi \in \Phi \text{ and } \varphi \models A\}$. Φ is called *elementary* iff Φ is the class of all models of some set of propositional formulas $\mathcal{F}$.

Definition 1 *Let Φ be a class of interpretations of language $\mathcal{L}$ and Π a set of propositional formulas. The pair $<\Phi, \Pi>$ is called a* structure *iff*

(i) Φ is elementary;
(ii) For every $A \in \Pi$, $\Phi(A) \neq \emptyset$.

Intuitively, a structure $<\Phi, \Pi>$ is interpreted as a class of all models of some belief set (i.e. Φ) and a set of justifications (i.e. Π) for these beliefs that are consistent with these beliefs.

Definition 2 *Let Φ be an elementary class of interpretations of language $\mathcal{L}$ and $c = A \leadsto B$ or $c = A \Rightarrow B$ be a causal relation. We say that c is* applicable *with respect to Φ iff*

(i) $\varphi \models A$ for every $\varphi \in \Phi$;
(ii) If $c = A \leadsto B$, there is some $\varphi \in \Phi$ such that $\varphi \models B$.

Definition 3 *Let $<\Phi, \Pi>$ be a structure. The* characteristic function *f_c of a causal relation $c = A \leadsto B$ or $c = A \Rightarrow B$ with respect to $<\Phi, \Pi>$ is defined as follows:*

$$f_c(<\Phi, \Pi>) = \begin{cases} <\Phi(B), \Pi \cup \{B\}> & \textit{if } c = A \leadsto B \textit{ and is applicable} \\ & \textit{with respect to } \Phi \\ <\Phi(B), \Pi> & \textit{if } c = A \Rightarrow B \textit{ and is applicable} \\ & \textit{with respect to } \Phi \\ <\Phi, \Pi> & \textit{if } c \textit{ is not applicable} \\ & \textit{with respect to } \Phi \end{cases}$$

Note that a characteristic function f_c changes a structure $<\Phi, \Pi>$ to another pair $<\Phi', \Pi'>$, which may or may not be a structure, by applying causal relation c as shown in the definition. The following definition defines the stability of a structure in an obvious way.

Definition 4 *Let $<\Phi, \Pi>$ be a structure and let $\mathcal{C}$ be a set of causal relations. We say that $<\Phi, \Pi>$ is* stable *with respect to $\mathcal{C}$ iff*

$f_c(<\Phi, \Pi>) = <\Phi, \Pi>$ *for all $c \in \mathcal{C}$.*

Definition 5 *Let $<\Phi, \Pi>$ be a structure and $<c_i>$ a sequence of causal* *relations. By* $f_{<c_i>}(<\Phi,\Pi>)$ *we denote:*

$$f_{<c_i>}(<\Phi,\Pi>) = \begin{cases} <\Phi,\Pi> & \textit{if } <c_i>=<> \textit{ (the empty sequence)} \\ <\bigcap \Phi_i, \bigcup \Pi_i> & \textit{otherwise} \end{cases}$$

where $<\Phi_0,\Pi_0>=<\Phi,\Pi>$*, and for* $i = 0, 1, \cdots$

$<\Phi_{i+1},\Pi_{i+1}>= f_{c_i}(<\Phi_i,\Pi_i>)$.

Definition 6 *Let* $<\Phi,\Pi>$ *be a structure,* Ψ *be an elementary class of interpretations of language* $\mathcal{L}$*. Suppose* $<c_i>$ *is a sequence of causal relations. We say that* $<\Phi,\Pi>$ *is* $f_{<c_i>}$*-accessible from* Ψ *iff* $<\Phi,\Pi>= f_{<c_i>}(<\Psi,\emptyset>)$*. We say that* $<\Phi,\Pi>$ *is* accessible *from* Ψ *with respect to* $\mathcal{C}$ *iff there is a sequence (possible empty)* $<c_i>$ *of causal relations from* $\mathcal{C}$ *such that* $<\Phi,\Pi>$ *is* $f_{<c_i>}$*-accessible from* Ψ*.*

Now we are ready to give the definition of the class of models of a causal theory. A detailed explanation of the following definition is referred to Łukaszewicz's work [12].

Definition 7 *Let* $CT = (\mathcal{A},\mathcal{C})$ *be a causal theory,* $Mod(\mathcal{A})$ *the class of all models of* $\mathcal{A}$*, and* Θ *the set of all subclasses of* $Mod(\mathcal{A})$ *such that for each* $\Phi \in \Theta$*, there exists a set of propositional formulas* Π *satisfying*

(i) $<\Phi,\Pi>$ *is a structure;*
(ii) $<\Phi,\Pi>$ *is stable with respect to* $\mathcal{C}$*;*
(iii) $<\Phi,\Pi>$ *is accessible from* $Mod(\mathcal{A})$ *with respect to* $\mathcal{C}$*.*

Then, the class of all models of CT*, denoted as* $Mod(CT)$*, is defined as*

$$Mod(CT) = \bigcup_{\Phi\in\Theta} \Phi$$

Here we present some intuitions behind the above definition. Recall that our causal theory $(\mathcal{A},\mathcal{C})$ is a special form of Reiter's default theory. As Łukaszewicz showed in [12], each Φ in Θ above actually is a class of all models of some extension of the corresponding default theory. Hence, $\bigcup_{\Phi\in\Theta} \Phi$ is the class of all models of all extensions of the corresponding default theory. We then specify this class to be the class of all models of our causal theory.

Example 1 Consider a causal theory

$$CT = (\{\neg floods\}, \{\textit{heavy-rainfall} \leadsto floods\}).$$

Obviously, $<Mod(\{\neg floods\}),\emptyset>$ is a structure. Further, it is easy to see that
$<Mod(\{\neg floods\}),\emptyset>$ is stable with respect to causal relation $heavy\text{-}rainfall$

$\rightsquigarrow floods$. Therefore, from Definition, 7, it follows that

$$Mod(CT) = Mod(\{\neg floods\}).$$

■

Example 2 Consider a causal theory

$$CT = (\{blocked(duct_1), blocked(duct_2)\},\\ \{blocked(duct_1) \wedge blocked(duct_2) \Rightarrow stuffy(room)\}).$$

Since $blocked(duct_1) \wedge blocked(duct_2) \Rightarrow stuffy(room)$ is applicable with respect to $Mod(\{blocked(duct_1), blocked(duct_2)\})$, it is easy to see that $< \{blocked(duct_1), blocked(duct_2), stuffy(room)\}, \emptyset >$ is the only stable and accessible structure from $Mod(\{blocked(duct_1), blocked(duct_2)\})$ with respect to the causal relation. So we have the following solution:

$$Mod(CT) = \{\{blocked(duct_1), blocked(duct_2), stuffy(room)\}\}.$$

■

Example 3 Consider a causal theory

$$CT = (\{bird(tweety), brokenwing(tweety)\},\\ \{bird(tweety) \rightsquigarrow fly(tweety),\\ brokenwing(tweety) \Rightarrow \neg fly(tweety)\}).$$

Ignoring the detail, this causal theory has a unique model:

$$\{bird(tweety), brokenwing(tweety), \neg fly(tweety)\}.$$

■

Example 4 Consider a causal theory

$$CT = (\{smoke, sport\}, \{smoke \rightsquigarrow \neg health, sport \rightsquigarrow health\}).$$

It is not difficult to see that there are two structures,

$< \{smoke, sport, \neg health\}, \{\neg health\} >$, and
$< \{smoke, sport, health\}, \{health\} >$,

which are stable and accessible from $< Mod(\{smoke, sport\}), \emptyset >$ with respect to causal relations $smoke \rightsquigarrow \neg health$ and $sport \rightsquigarrow health$. Hence models of this causal theory are

$\{smoke, sport, \neg health\}$, and
$\{smoke, sport, health\}$.

If we replace causal relations $smoke \leadsto \neg health$ and $sport \leadsto health$ by $smoke \Rightarrow \neg health$ and $sport \Rightarrow health$ respectively, it is clear that the new causal theory does not have a model according to Definition 7. ■

3 A causality-based minimal change approach

3.1 Preliminaries

In this section we propose a formulation of actions in which a causality-based minimal change principle is employed. To begin with, we first introduce some useful notations and concepts. We use an interpretation S of the language $\mathcal{L}$ to represent the *state of the world*. We assume that there is a set of propositional formulas $\mathcal{D}$ to represent *domain constraints* about the world, which should be satisfied by each state of the world, i.e. $S \models \mathcal{D}$.

We also consider a set of causal relations $\mathcal{C}$ to represent the causality of the domain while a causal relation in $\mathcal{C}$ is expressed as $A \leadsto B$ or $A \Rightarrow B$ as described in last section. Let $c : A \leadsto B$ or $A \Rightarrow B$ be a causal relation. A state S is called *c-consistent* if and only if $S \models A$ implies $S \models B$. S is called *$\mathcal{C}$-consistent* if and only if for each $c \in \mathcal{C}$, S is c-consistent. We assume that every initial state is $\mathcal{C}$-consistent before any actions are performed on it. However, as it will be shown later, the resulting state may not be $\mathcal{C}$-consistent after some action is performed on the initial state.

An *action act* is specified by a post-condition $Post(act)$, where $Post(act)$ is a propositional formula and represents the direct effect of act. Informally, action act changes the state of the world from one to another, and in the resulting state $Post(act)$ is satisfied. Note that here we do not specify the pre-condition of action act in the sense that the executability of an action on a state can be derived from our approach. This is the so-called *derived qualification* [13] that will be addressed in sections 4 and 6.

Let S_1 and S_2 be two interpretations. We say that S_1 and S_2 *differ* on a propositional letter p if p appears in exactly one of S_1 and S_2. $Diff(S_1, S_2)$ denotes the set of all different propositional letters between S_1 and S_2. Let S be an interpretation, Ψ be a set of interpretations. We define the set of all *minimal different interpretations* of Ψ with respect to S as follows:

$$Min(S, \Psi) = \{S' \mid S' \in \Psi, \text{ and there does not exist other } S'' \in \Psi \text{ such that } Diff(S, S'') \subset Diff(S, S')\}.$$

Let S be an initial state, act an action and $\mathcal{D}$ a set of domain constraints.

Then the PMA [18] defines the set of possible resulting states as follows:

$$Res^{\mathcal{D}}(S, act) = Min(S, Mod(\mathcal{D} \cup \{Post(act)\})).$$

3.2 The approach

Taking causality into account, we need to consider indirect effects of actions which are related to causal relations of the domain. As was showed in section 1, the traditional minimal change approach, in which only logical constraints are considered, cannot represent these action effects properly. Here we propose a *causality-based minimal change principle*. The basic idea is: during a state transition, the minimal change principle is applied generally but with exception on the *causally indirect effects* of actions. For instance, in the simplified stuffy room example, suppose that we replace constraint (1) with a causal relation

$$blocked(duct_1) \wedge blocked(duct_2) \Rightarrow stuffy(room), \tag{3}$$

and the current state is:

$$S = \{\neg blocked(duct_1), blocked(duct_2), \neg stuffy(room)\}. \tag{4}$$

If an action of making $duct_1$ blocked, i.e. the post-condition is $blocked(duct_1)$, is performed on S, and there is no other domain constraint, intuitively, we would expect the resulting state to be

$$S' = \{blocked(duct_1), blocked(duct_2), stuffy(room)\}.$$

To derive this intuitive resulting state from S, we may use the minimal change principle on literal $blocked(duct_2)$ but with exception on literal $\neg stuffy(room)$.

The reason is that since $blocked(duct_2)$ is consistent with the direct effect $blocked(duct_1)$ of the action, according to the minimal change principle, $blocked(duct_2)$ should persist during the state change. On the other hand, because of causal relation (3), we know that from $blocked(duct_1)$ and $blocked(duct_2)$, we should *causally* derive $stuffy(room)$. That is, $stuffy(room)$ is an causally indirect effect of this action. So, the minimal change principle should not apply to literal $\neg stuffy(room)$. Now, we are ready to formalize the idea described above. The following two definitions give the formal description of the causality-based minimal change approach.

Definition 8 *Let $\mathcal{C}$ and $\mathcal{D}$ be the sets of causal relations and constraints of the domain respectively, S a state of the world, and act an action. Let δ be a maximal subset of S such that*

(i) $\mathcal{D} \cup \{Post(act)\} \cup \delta$ *is consistent,*
(ii) for any $A \rightsquigarrow B$ *or* $A \Rightarrow B$ *in* $\mathcal{C}$, $\delta \not\models \neg B$,
(iii) let $\delta = \emptyset$ *if there does not exist a* δ *satisfying conditions 1 and 2 above.*

The set of all such maximal subsets of S *is denoted as* $Max(S, act)$.

Definition 9 *Let* $\mathcal{C}$, $\mathcal{D}$, S, act, δ *and* $Max(S, act)$ *be defined as above. A causal theory* $CT^{act}_{S,\delta}$ *is specified as follows:*

$$CT^{act}_{S,\delta} = (\mathcal{A}, \mathcal{C}), \text{ where } \mathcal{A} = \mathcal{D} \cup \{Post(act)\} \cup \delta, \\ \text{and } \delta \in Max(S, act).$$

Then the set of all possible resulting states from performing act on S, $Res^{\mathcal{C},\mathcal{D}}(S, act)$, *is defined as*

$$Res^{\mathcal{C},\mathcal{D}}(S, act) = \bigcup_{\delta \in Max(S,act)} Min(S, Mod(CT^{act}_{S,\delta})). \tag{5}$$

Let us examine the above two definitions more closely. In Definition 8, δ is the maximal subset of S with respect to action *act* such that (i) δ is consistent with domain constraints $\mathcal{D}$ and the direct effect of action *act* $\{Post(act)\}$, and (ii) δ is also consistent with the effect of each causal relation in $\mathcal{C}$, or (iii) $\delta = \emptyset$ if there is no such δ satisfying (i) and (ii). Obviously, condition (i) is the basic requirement for a state change. Condition (ii) means that those facts in δ together with $\mathcal{D}$ and $\{Post(act)\}$ *may* causally derive some causal effects according to causal relations in $\mathcal{C}$. In this case, if negations of these causal effects are entailed by S, they should be changed after this action. To make such causal reasoning successfully, it is necessary to require that δ be consistent with the effect B of any causal relation $A \rightsquigarrow B$ or $A \Rightarrow B$ in $\mathcal{C}$.

It may be argued that Condition 2 in Definition 8 is too conservative since it might be the case that for some $A \rightsquigarrow B$ or $A \Rightarrow B$ in $\mathcal{C}$, $S \not\models A$ and B may not be causally derived. For instance, suppose $S = \{\neg A, \neg B, C\}$, $\mathcal{C} = \{A \rightsquigarrow B\}$ (or $\mathcal{C} = \{A \Rightarrow B\}$) and $Post(act) \equiv \neg C$. In this case, $\delta = \{\neg A\}$. However, note that this will not affect the persistence of $\neg B$ as the minimal change will guarantee this as Definition 9 states.

It should be also noted that Condition 3 in Definition 8 is necessary because in some situations, there is no δ satisfying conditions 1 and 2. For example, consider a situation that we have a causal relation $A \Rightarrow False$, domain constraint $\mathcal{D} = \emptyset$, an initial state $S = \{\neg A\}$ and an action's post-condition $\neg A$. Intuitively, this action does not change S so the resulting state should be still S. However, as the effect of the causal relation is $False$, there is no δ satisfying conditions 1 and 2. In this case, Condition 3 ensures that a background knowledge $\mathcal{A}$ about the domain with respect to the action, which is needed in the causal theory $CT^{act}_{S,\delta}$ defined in Definition 9, can be still formed by assigning

$\delta = \emptyset$.

Now consider the stuffy room example again where $\mathcal{D} = \emptyset$, $\mathcal{C}$ consists of a single causal relation (3), and the initial state S is (4). The direct effect of action of making $duct_1$ blocked is $blocked(duct_1)$. According to the above definition, we have $\delta = \{blocked(duct_2)\}$. Note that $\neg stuffy(room)$ is not included in δ since δ needs to be consistent with $stuffy(room)$. Therefore, from $\{blocked(duct_1)\} \cup \delta$ together with causal relation (3), we may derive a causal effect $stuffy(room)$.

The derivation of causally indirect effects of act with respect to $\mathcal{C}$ is achieved by specifying a causal theory $CT^{act}_{S,\delta}$ in Definition 9 and obtaining all models $Mod(CT^{act}_{S,\delta})$ of $CT^{act}_{S,\delta}$. Then we specify the set of final possible resulting states to be the union of those models in $Mod(CT^{act}_{S,\delta})$ for any $\delta \in Max(S, act)$, such that models in $Mod(CT^{act}_{S,\delta})$ have minimal differences with the the initial state S (i.e. (5)).

Continue considering the stuffy room example presented above. From Definition 8, $Max(S, act) = \{\delta\} = \{\{blocked(duct_2)\}\}$, and from Definition 9, we have

$$CT^{act}_{S,\delta} = (\{blocked(duct_1),\ blocked(duct_2)\},\ \{(3)\}), \text{ and}$$

$$\begin{aligned} Res^{\mathcal{C},\mathcal{D}}(S, act) &= Min(S, Mod(CT^{act}_{S,\delta})) \\ &= \{\{blocked(duct_1), blocked(duct_2), stuffy(room)\}\}, \end{aligned}$$

which gives the desired solution as we expect.

It is easy to see that if the set of causal relations is empty, our approach described above reduces to the PMA, as the following theorem states.

Theorem 1 *$Res^{\mathcal{D}}(S, act) = Res^{\emptyset,\mathcal{D}}(S, act)$.*

Proof We first prove $Res^{\mathcal{D}}(S, act) \subseteq Res^{\emptyset,\mathcal{D}}(S, act)$. Let $S' \in Res^{\mathcal{D}}(S, act)$. From the definition of $Res^{\mathcal{D}}(S, act)$, we have $S' \in Mod(\mathcal{D} \cup \{Post(act)\})$ and for any $S'' \in Mod(\mathcal{D} \cup \{Post(act)\})$, $Diff(S, S'') \not\subset Diff(S, S')$. This follows that $S \cap S'$ is a maximal subset of S such that $\mathcal{D} \cup \{Post(act)\} \cup S \cap S'$ is consistent. So $S \cap S' \in Max(S, act)$. Let $\delta = S \cap S'$. That is, $S' \in Mod(\mathcal{D} \cup \{Post(act)\} \cup \delta)$. Since $Mod(\mathcal{D} \cup \{Post(act)\} \cup \delta) \subseteq Mod(\mathcal{D} \cup \{Post(act)\})$ and for any $S'' \in Mod(\mathcal{D} \cup \{Post(act)\})$, $Diff(S, S'') \not\subset Diff(S, S')$, it follows that for any $S'' \in Mod(\mathcal{D} \cup \{Post(act)\} \cup \delta)$, $Diff(S, S'') \not\subset Diff(S, S')$. So $S' \in Min(S, Mod(\mathcal{D} \cup \{Post(act)\} \cup \delta))$. That is, $S' \in Res^{\emptyset,\mathcal{D}}(S, act)$ from Definition 9 (note that since $\mathcal{C} = \emptyset$, $Mod(CT^{act}_{S,\delta}) = Mod(\mathcal{D} \cup \{Post(act)\} \cup \delta)$).

Now we prove $Res^{\emptyset,\mathcal{D}}(S, act) \subseteq Res^{\mathcal{D}}(S, act)$. Let $S' \in Res^{\emptyset,\mathcal{D}}(S, act)$. Therefore, there exists some $\delta \in Max(S, act)$ and $S' \in Min(S, Mod(\mathcal{D} \cup \{Post(act)\} \cup \delta))$. Hence, for any $S'' \in Mod(\mathcal{D} \cup \{Post(act)\} \cup \delta)$, $Diff(S, S'') \not\subset Diff(S, S')$. Clearly, S' is also in $Mod(\mathcal{D} \cup \{Post(act)\})$. We need to show that for any $S'' \in Mod(\mathcal{D} \cup \{Post(act)\})$, $Diff(S, S'') \not\subset Diff(S, S')$.

Firstly, it is obvious that $\delta = S \cap S'$ since $\delta \subseteq S \cap S'$ and from the maximum of δ we have $S \cap S' \subseteq \delta$.

Suppose that there is some $S'' \in Mod(\mathcal{D} \cup \{Post(act)\})$ such that $Diff(S, S'') \subset Diff(S, S')$. That is $S \cap S' \subset S \cap S''$. But as $\delta = S \cap S' \in Max(S, act)$, this is contradictory. So for any $S'' \in Mod(\mathcal{D} \cup \{Post(act)\})$, $Diff(S, S'') \not\subset Diff(S, S')$, which means $S' \in Res^{\mathcal{D}}(S, act)$. $\square$

4 Further examples

In this section, we discuss more examples to show how the approach described previously represents causal effects of actions properly.

Example 5 Consider a baby room domain. In this domain, $\mathcal{D}$ consists of the following constraints:

$$ontable(cup) \supset safe(cup), \tag{6}$$

$$ontable(doll) \supset safe(doll), \tag{7}$$

Intuitively, (6) and (7) say that both the cup and doll on the table are safe to the baby.

This domain reveals a problem so-called the *persistence of derived information* first examined by Myers and Smith [15]. Specifically, $safe(cup)$ (and $safe(doll)$) is derived information from $ontable(cup)$ (and $ontable(doll)$ respectively). But there are different requirements for the persistence of $safe(cup)$ and $safe(doll)$ with respect to the change of $ontable(cup)$ and $ontable(doll)$ respectively. For example, consider an initial state while both the cup and doll are on the table (i.e. $ontable(cup)$ and $ontable(doll)$), and from (6) and (7) this follows that both the cup and doll are safe to the baby (i.e. $safe(cup)$ and $safe(doll)$). According to the minimal change principle, removing the cup or doll from the table will not affect the persistence of $safe(cup)$ and $safe(doll)$ respectively. However, from our commonsense on the other hand, removing the cup from the table (eg. put the cup on the floor) may cause it to be no longer safe to the baby, while removing the doll from the table should not affect the doll's safety to the baby.

This problem can be avoided with our approach if we specify a causal relation (i.e. $\mathcal{C}$ consists of just one causal relation) in the domain:

$$\neg ontable(cup) \rightsquigarrow \neg safe(cup). \tag{8}$$

Formally, let the initial state be:

$$S = \{ontable(cup), ontable(doll), safe(cup), safe(doll)\}.$$

Suppose action act_1 of removing the cup from the table (i.e. $Post(act_1) \equiv \neg ontable(cup)$) is performed on S. To predict the possible resulting state, we use the approach presented in section 3. According to Definition 8, it is not difficult to see that

$$Max(S, act_1) = \{\delta_1\} = \{\{ontable(doll), safe(doll)\}\}.$$

Then we have a causal theory as follows:

$$CT^{act_1}_{S,\delta_1} = (\{(6), (7)\} \cup \{\neg ontable(cup)\} \cup \delta_1, \{(8)\}).$$

The set of models of $CT^{act_1}_{S,\delta_1}$ is

$$Mod(CT^{act_1}_{S,\delta_1}) = \{\varphi_1\}, \text{ where}$$
$$\varphi_1 = \{\neg ontable(cup), ontable(doll), \neg safe(cup), safe(doll)\}.$$

Finally, from (5) in Definition 9, we have the possible resulting state as follows:

$$Res^{\mathcal{C},\mathcal{D}}(S, act_1) = Min(S, Mod(CT^{act_1}_{S,\delta_1})) = \{\varphi_1\},$$

which represents our intuitive solution – after removing the cup from the table, it is no longer safe to the baby.

Now consider that still at the same initial state S, another action act_2 of removing the doll from the table (i.e. $Post(act_2) \equiv \neg ontable(doll)$) is performed on S. Similarly, we have

$$Max(S, act_2) = \{\delta_2\} = \{\{ontable(cup), safe(cup), safe(doll)\}.$$

Ignoring the detail, we finally get the possible resulting state as follows:

$$Res^{\mathcal{C},\mathcal{D}}(S, act_2) = \{\varphi_2\}, \text{ where}$$
$$\varphi_2 = \{ontable(cup), \neg ontable(doll), safe(cup), safe(doll)\},$$

which also gives the intuitive solution – removing the doll from the table does not affect its safety to the baby. ■

Example 6 From our commonsense, we know that if a baby is awake, he/she normally is not quiet. But if the baby is cuddled by his/her mother, he/she normally is quiet. These can be described by the following causal relations respectively:

$$babyawake \rightsquigarrow \neg quiet, \tag{9}$$

$$babycuddled \rightsquigarrow quiet. \tag{10}$$

Suppose currently the baby is awake and is not cuddled by mother and then he/she is crying. Hence the initial state is

$S = \{babyawake, \neg quiet, \neg babycuddled\}$.

Now assume that the mother cuddles the baby. Our question is: Is the baby quiet after cuddled by his/her mother? Obviously, using our method presented in section 3, we have two possible resulting states

$S_1 = \{babyawake, \neg quiet, babycuddled\}$, and
$S_2 = \{babyawake, quiet, babycuddled\}$,

which seem reasonable as both *babyawake* and *babycuddled* being true leads to two possible statuses of the baby's quietness according to causal relations (9) and (10) respectively. ∎

Example 7 The light-power domain. Consider a domain with two causal relations as following respectively:

$$on \rightsquigarrow light, \tag{11}$$

$$\neg on \vee \neg power \Rightarrow \neg light. \tag{12}$$

Intuitively, (11) says that the fact that switch is on normally causes the fact that the lamp lights, while (12) says that the fact that the switch is off or there is no power will cause the fact that the lamp can not light. For causal relation (11), people may think that a precise causal relation should be $on \wedge power \rightsquigarrow light$. But we argue that a more precise one can be $on \wedge power \wedge \neg broken\text{-}lamp \wedge circuit\text{-}ok \wedge \cdots \rightsquigarrow light$. Since (11) is a defeasible causal relation, it is sufficient to represent (11) in our domain. It is also important to note that causal relation (12) is not defeasible in the sense that the fact that the switch is off or there is no power will definitely cause the fact that the lamp can not light.

Now suppose the initial state $S_1 = \{light, on, power\}$. Consider an action of cutting off the power (i.e. $Post(act_1) \equiv \neg power$) is performed on S_1. Obviously using our method presented in section 3, we have a unique resulting state $S' = \{\neg light, on, \neg power\}$. On the other hand, suppose the initial state $S_2 = \{\neg light, \neg on, \neg power\}$ and an action of turning on the switch (i.e.

$Post(act_2) \equiv on$) is performed on S_2. Clearly, there is also a unique resulting state $S'' = \{\neg light,\ on,\ \neg power\}$. Note that the resulting state is not consistent with respect to causal relation $on \rightsquigarrow light$. ■

Example 8 Consider a domain with a causal relation:

$$\neg alive \Rightarrow \neg walking, \tag{13}$$

which says that the fact that some one is not alive causes the fact that he/she is not walking. Suppose the initial state $S = \{\neg alive, \neg walking\}$ to represent that Fred is not alive and he is not walking. Assume that an action act with the postcondition $walking$ is intended to perform on S. Using our method, we will get $Res^{\{(13)\},\emptyset}(S, act) = \emptyset$, that means that the action act can not be performed on S. Intuitively, this solution is reasonable from our commonsense that if Fred is not alive then he can not walk any more. Technically, this result is achieved by specifying causal relation (13) as a non-defeasible rule. This is also an example of a derived qualification. ■

5 Representing action sequences

It is straightforward to generalize our approach presented in section 3 to the case of a sequence of actions. Formally, let $\bar{A}$ be an action sequence $act_1 \cdot act_2 \cdots act_n$ and act_{n+1} an action, we denote $\bar{A} \cdot act_{n+1} = act_1 \cdot act_2 \cdots act_n \cdot act_{n+1}$. An empty action sequence is denoted as ϵ. Suppose S is the initial state in which $\bar{A}$ starts to be executed sequentially. Then the set of possible states resulting from performing action sequence $\bar{A}$ on S sequentially, $Res^{\mathcal{C},\mathcal{D}}(S, \bar{A})$, is defined as follows:

$$Res^{\mathcal{C},\mathcal{D}}(S, \bar{A}) = \begin{cases} S & \text{if } \bar{A} = \epsilon \\ \bigcup_{S' \in Res^{\mathcal{C},\mathcal{D}}(S, \bar{A}_1)} Res^{\mathcal{C},\mathcal{D}}(S', act) & \bar{A} = \bar{A}_1 \cdot act \end{cases}$$

Let us consider the light-power domain described in Example 7 again, where $\mathcal{C} = \{(11), (12)\}$ and $\mathcal{D} = \emptyset$. Suppose the initial state is

$$S = \{\neg light, \neg on, power\},$$

and three actions of turning on the switch ($Post(act_1) \equiv on$), cutting off the power ($Post(act_2) \equiv \neg power$), and providing power ($Post(act_3) \equiv power$) are sequentially performed on S. According to the definition of $Res^{\mathcal{C},\mathcal{D}}(S, \bar{A})$, we have the following results:

$$Res^{\mathcal{C},\mathcal{D}}(S, act_1) = \{\{light, on, power\}\},$$

$Res^{\mathcal{C},\mathcal{D}}(S, act_1 \cdot act_2) = \{\{\neg light, on, \neg power\}\}$, and
$Res^{\mathcal{C},\mathcal{D}}(S, act_1 \cdot act_2 \cdot act_3) = \{\{light, on, power\}\}$.

6 Comparisons with related work

The idea of introducing causal connections into action theories is not new. Hayes' work on the logic of actions [7] probably was the first formalization of considering domain causality in an action theory. Some other researchers have also used the notion of causality to formalize the effects of actions (eg. [9]). However, as pointed by Lin [11], most of them only considered the direct effects *caused* by actions. So their approaches did not represent the indirect effects of actions properly. Recently, Brewka and Hertzberg [2], McCain and Turner [13] and Lin [11] have independently proposed new theories of actions in which the domain causality was explicitly taken into account. In this section, we will specifically compare our approach with McCain and Turner's and Brewka and Hertzberg's approaches. We will also briefly discuss Lin's method.

6.1 McCain and Turner's approach

In McCain and Turner's method, a causal relation is presented as an (non-defeasible) inference rule. For instance, expression $A \Rightarrow B$ indicates a causal relation between propositional formulas A and B, eg. the statement that Fred not being alive *causes* Fred not being walking can be expressed as $\neg alive \Rightarrow \neg walking$, which has the same form of the non-defeasible causal relation in our approach. Let $\mathcal{C}$ be a set of such causal laws, Γ a set of propositional formulas. Γ is *closed under* $\mathcal{C}$ if for every $A \Rightarrow B \in \mathcal{C}$, if $A \in \Gamma$ then $B \in \Gamma$. For any formula A, $\Gamma \vdash_{\mathcal{C}} A$ denotes that A belongs to the *smallest* set of formulas containing Γ that is closed with respect to propositional logic and closed under $\mathcal{C}$. Then the set of possible states resulting from the performance of action act on the initial state S, $Res^{\mathcal{C}}(S, act)$, is specified to be the set of interpretations such that

$$Res^{\mathcal{C}}(S, act) = \{S' \mid S' = \{l \mid (S \cap S') \cup \{Post(act)\} \vdash_{\mathcal{C}} l\}\}, \tag{14}$$

where l is a propositional literal.

Note that in McCain and Turner's method, the possible resulting states are defined by satisfying a fixpoint condition with respect to the inference relation $\vdash_{\mathcal{C}}$. To illustrate the difference between McCain and Turner's method and ours, let us first consider the following example.

Example 9 The light-power domain revisited. Let us reconsider Example 7 within McCain and Turner's formalism. As all causal relations in their formalism are represented as (non-defeasible) inference rules, the causal relation (11) needs to be represented as

$$on \Rightarrow light. \tag{15}$$

Therefore, the set of causal relations of the light-power domain here is $\mathcal{C} = \{(12), (15)\}$. Suppose that the initial state $S_1 = \{light, on, power\}$, and the action of cutting off the power (i.e. $Post(act_1) \equiv \neg power$) is performed on S_1. From (14), we have the result $Res^{\mathcal{C}}(S_1, act_1) = \emptyset$, i.e. the action of cutting off the power can not occur. On the other hand, if the initial state $S_2 = \{\neg light, \neg on, \neg power\}$ and the action of turning on the switch (i.e. $Post(act_2) \equiv on$) is performed on S_2. Similarly, we still have $Res^{\mathcal{C}}(S_2, act_2) = \emptyset$.

Now we argue that both solutions above seem not quite reasonable from our intuition. For instance, it is necessary for a security robot to be able to cut off the power for the safety if there is a fire alarm. Similarly, no matter if there is power, the action of turning on the switch should be executable usually. From this sense, McCain and Turner's method is quite restricted. In order to obtain the desired solutions as described in Example 7, within McCain and Turner's formalism, we have to rewrite causal relation (15) to the following form

$$on \wedge power \Rightarrow light. \tag{16}$$

However, as we mentioned in Example 7, it is always arguable that a more precise causal relation can be obtained by adding more causes in (16), like $on \wedge power \wedge \neg broken\text{-}lamp \Rightarrow light$ and so on. Obviously this is not always practicable to represent causal relations. ■

From the above example, we can see that distinguishing causal relations as defeasible and non-defeasible seems more natural and flexible than just representing them as unified non-defeasible inference rules. In fact, the following result shows that McCain and Turner's method can be treated as a special case in our system.

Theorem 2 *Let $\mathcal{C}$ be a set of causal relations in which every causal relation has the form $A \Rightarrow B$, S a state, and act an action. Then $Res^{\mathcal{C}}(S, act) = Res^{\mathcal{C},\emptyset}(S, act)$.*

Proof We first prove $Res^{\mathcal{C}}(S, act) \subseteq Res^{\mathcal{C},\emptyset}(S, act)$. If $Res^{\mathcal{C}}(S, act) = \emptyset$, then the result holds. Now suppose $Res^{\mathcal{C}}(S, act) \neq \emptyset$. For any $S' \in Res^{\mathcal{C}}(S, act)$, $S' = \{l \mid (S \cap S') \cup \{Post(act)\} \vdash_{\mathcal{C}} l\}$, this follows that S' is the interpretation such that
(i) $S' \models Post(act)$,

(ii) S' is $\mathcal{C}$-consistent, and
(iii) for any other interpretation S'' satisfying (i) and (ii), $S \cap S' \not\subset S \cap S''$.
This is because that if there exists some S'' satisfying (i), (ii) and $S \cap S' \subset S \cap S''$, we will have $\{l \mid (S \cap S') \cup \{Post(act)\} \vdash_{\mathcal{C}} l\} \subset \{l \mid (S \cap S'') \cup \{Post(act)\} \vdash_{\mathcal{C}} l\}$. But S' is an interpretation and $S' = \{l \mid (S \cap S') \cup \{Post(act)\} \vdash_{\mathcal{C}} l\}$. So it is contradictory.

Since S' satisfies conditions (i), (ii) and (iii), it follows that $S \cap S'$ is the maximal subset of S such that $\{Post(act)\} \cup S \cap S'$ is consistent and $S \cap S'$ is $\mathcal{C}$-consistent. Therefore, there exists some $\delta \in Max(S, act)$ such that $\delta \subseteq S \cap S'$. Let $\mathcal{A} = \{Post(act)\} \cup \delta$. Then $S' \in Mod(\mathcal{A})$. We specify a causal theory $CT^{act}_{S,\delta} = (\mathcal{A}, \mathcal{C})$. From Definition 3-7, it is easy to see that $S' \in Mod(CT^{act}_{S,\delta})$. Since from Definition 3 and Definition 7, if $S' \in Mod(\mathcal{A})$ but $S' \notin Mod(CT^{act}_{S,\delta})$, then there must be some $A \Rightarrow B$ in $\mathcal{C}$ such that $S' \models A$ and $S' \not\models B$. This is contradictory to S''s $\mathcal{C}$-consistency.

We then need to show $S' \in Min(S, Mod(CT^{act}_{S,\delta}))$ according to Definition 9. Assume that $S' \notin Min(S, Mod(CT^{act}_{S,\delta}))$. It means that there exists some $S'' \in Mod(CT^{act}_{S,\delta})$ such that $Diff(S, S'') \subset Diff(S, S')$, which leads to $S \cap S' \subset S \cap S''$. Since $S'' \models Post(act)$ and S'' is $\mathcal{C}$-consistent, it turns out that S' does not satisfy conditions (i), (ii) and (iii) presented above, and obviously this is contradictory.

Now we prove $Res^{\mathcal{C},\emptyset}(S, act) \subseteq Res^{\mathcal{C}}(S, act)$. Obviously if $Res^{\mathcal{C},\emptyset}(S, act) = \emptyset$, then the result holds. Suppose $Res^{\mathcal{C},\emptyset}(S, act) \neq \emptyset$. For any $S' \in Res^{\mathcal{C},\emptyset}(S, act)$, there exists some $\delta \in Max(S, act)$ such that $\delta \subseteq S \cap S'$ and $S' \in Mod(CT^{act}_{S,\delta})$, where $CT^{act}_{S,\delta} = (\mathcal{A}, \mathcal{C})$ and $\mathcal{A} = \{Post(act)\} \cup \delta$. Therefore, we have $\{l \mid \delta \cup \{Post(act)\} \vdash_{\mathcal{C}} l\} \subseteq S'$. Furthermore, since $S' \in Min(S, Mod(CT^{act}_{S,\delta}))$ and then $Diff(S, S'') \not\subset Diff(S, S')$ for any S'' in $Mod(CT^{act}_{S,\delta})$, $S' - \{l \mid \delta \cup \{Post(act)\} \vdash_{\mathcal{C}} l\} \subseteq S \cap S'$. So we have $S' = \{l \mid \delta \cup \{Post(act)\} \vdash_{\mathcal{C}} l\} \cup S \cap S'$. On the other hand, as $\delta \subseteq S \cap S'$, we have $\{l \mid \delta \cup \{Post(act)\} \vdash_{\mathcal{C}} l\} \subseteq \{l \mid S \cap S' \cup \{Post(act)\} \vdash_{\mathcal{C}} l\}$. So $S' = \{l \mid \delta \cup \{Post(act)\} \vdash_{\mathcal{C}} l\} \cup S \cap S'$ $\subseteq \{l \mid S \cap S' \cup \{Post(act)\} \vdash_{\mathcal{C}} l\} \cup S \cap S' = \{l \mid S \cap S' \cup \{Post(act)\} \vdash_{\mathcal{C}} l\}$. That is: $S' \subseteq \{l \mid S \cap S' \cup \{Post(act)\} \vdash_{\mathcal{C}} l\}$. As S' is an interpretation, we have $S' = \{l \mid S \cap S' \cup \{Post(act)\} \vdash_{\mathcal{C}} l\}$, and hence $S' \in Res^{\mathcal{C}}(S, act)$. $\Box$

6.2 Brewka and Hertzberg's approach

Brewka and Hertzberg proposed an approach to deal with the problem of reasoning about actions with causal and disjunctive effects [2]. In their approach, like McCain and Turner's method, causal relations are explicitly represented as some inference rules. The difference between Brewka and Hertzberg's approach and McCain and Turner's is that Brewka and Hertzberg also considered general logical constraints as parts of background knowledge of the domain. In the rest of this section, we will first review Brewka and Hertzberg's method and then discuss the problem of their approach by illustrating a simple example. To keep our discussion to be consistent with our previous descriptions, we will slightly modify Brewka and Hertzberg's formulation.

Let $\mathcal{C}$ be a set of causal relations in which each causal relation has the form $A \Rightarrow B$ (A and B are propositional formulas), and $\mathcal{D}$ a set of propositional formulas to represent the set of domain constraints. A *background knowledge* K of the domain is defined as $K = Theory(\mathcal{C}) \cup \mathcal{D}$, where $Theory(\mathcal{C}) = \{A \supset B \mid A \Rightarrow B \in \mathcal{C}\}$. Every model in $Mod(K)$ is called a *possible world* of K. The $\mathcal{C}$-closure of some set of formulas H is the smallest deductively closed set that is also closed under the rules in $\mathcal{C}$. We denote $H \vdash_{\mathcal{C}} f$ if f is in the $\mathcal{C}$-closure of H. If F is a set of formulas in which each element of F is in the $\mathcal{C}$-closure of H, we denote $H \vdash_{\mathcal{C}} F$.

Given two possible worlds S_1 and S_2. $Diff_{BH}(S_1, S_2)$ denotes the set of literals true in S_2 but not in S_1. We should mention that $Diff_{BH}(S_1, S_2)$ is different from $Diff(S_1, S_2)$ defined in section 3. A set of literals P is called a *causal change set* of (S_1, S_2) if and only if P is a minimal subset of $Diff_{BH}(S_1, S_2)$ such that $P \cup (S_1 \cap S_2) \vdash_{\mathcal{C}} S_2'$ and $S_2' \cup \mathcal{D} \vdash S_2$ for some $S_2' \subseteq S_2$. Intuitively, a causal change set is a minimal set of changes that transforms S_1 to S_2, given the background knowledge K. The remaining changes between S_1 and S_2 are caused by the causal change set with respect to $\mathcal{C}$ and also forced by the constraint $\mathcal{D}$.

Suppose S, S_1 and S_2 are possible worlds. S_1 is called *closer* to S than S_2, denoted as $S_1 \prec_S S_2$, if and only if (i) $Diff_{BH}(S, S_1) \subset Diff_{BH}(S, S_2)$, or (ii) every causal change set of (S, S_1) is a subset of a causal change set of (S, S_2), and not vice versa.

Based on the above definitions, the possible world S' resulting from S by performing action act on S is defined as follows: S' is $\prec_S$-minimal among the worlds in $Mod(K)$ satisfying $Post(act)$. The set of all possible resulting worlds is denoted as $Res^{\mathcal{C},\mathcal{D}}_{BH}(S, act)$.

The intuitive idea of Brewka and Hertzberg's formalization is quite clear: a combination of the traditional minimal change principle on models (i.e. condition (i) described above) and a *minimal causal change principle* (i.e. condition (ii) above) is used in reasoning about action. Unfortunately, as it will be shown next, Brewka and Hertzberg's minimal causal change principle is actually not appropriate to capture causal effects of actions if *mutual effects* between causal relations and domain constraints are involved in the reasoning.

Example 10 Consider a domain consisting of two causal relations

$$on_1 \wedge on_2 \Rightarrow light, \tag{17}$$

$$bright(room) \Rightarrow visible(painting), \tag{18}$$

and a single logical constraint

$$light \supset bright(room). \tag{19}$$

Intuitively, (17) says that if two switches are on, then the lamp will light, and (18) says that if the room is bright, then the painting hung on the wall is visible, while constraint (19) simply states that the fact that lamp is lighting implies the fact that the room is bright. Note that we do not specify (19) as a causal relation since formula $\neg bright(room) \supset \neg light$ is also true from our commonsense. Now suppose the initial state is

$$S = \{\neg bright(room), \neg light, \neg on_1, on_2, \neg visible(painting)\}$$

and an action act of turning on the first switch is performed on S (i.e. $Post(act) \equiv on_1$). Since $K = Theory(\{(17), (18)\} \cup \{(19)\}$, there are six possible worlds of K satisfying $Post(act) \equiv on_1$ as follows:

$$\begin{aligned}
S_1 &= \{bright(room), light, on_1, on_2, visible(painting)\},\\
S_2 &= \{\neg bright(room), \neg light, on_1, \neg on_2, \neg visible(painting)\},\\
S_3 &= \{\neg bright(room), \neg light, on_1, \neg on_2, visible(painting)\},\\
S_4 &= \{\neg bright(room), \neg light, on_1, \neg on_2, \neg visible(painting)\},\\
S_5 &= \{bright(room), \neg light, on_1, \neg on_2, visible(painting)\},\\
S_6 &= \{bright(room), light, on_1, \neg on_2, visible(painting)\}.
\end{aligned}$$

From the definition of causal change set presented above, it is not difficult to see that $Diff_{BH}(S, S_1) \not\subset Diff_{BH}(S, S_2)$, and both $P_1 = \{on_1, visible(room)\} \subseteq Diff_{BH}(S, S_1)$ and $P_2 = \{on_1, \neg on_2\} \subseteq Diff_{BH}(S, S_2)$ are two minimal causal change sets, and therefore both S_1 and S_2 are $\prec_S$-minimal. So we have

$$Res_{BH}^{\{(17),(18)\},\{(19)\}}(S, act) = \{S_1, S_2\}.$$

Clearly S_1 is our desired result. But it is surprising that the unintuitive world S_2 is also a possible solution. ■

From the above example, we can see that Brewka and Hertzberg's approach still suffers from the difficulty of deriving indirect effects of actions originally pointed by Lifschitz [10], even if the causality was introduced in their formulation. In fact, the major problem of their formulation is that the definition of causal change set does not capture their intuition of minimal causal change principle properly. For instance, they did not consider the possibility of mutual effects occurring between causal relations and logical constraints. In Example 10, when literal $\neg on_1$ is changed to on_1, causal relation (17) is triggered and this causes literal $light$ to be true. Then constraint (19) should be applied to derive $bright(room)$. After that, the second causal relation (18) is triggered to derive literal $visible(painting)$. Clearly, the definition of causal change set only covers the first two steps and misses the last step.

On the other hand, if we use our approach described in section 3 to deal with Example 10, it is not difficult to see that the unique desired possible solution S_1 can be derived, where both causal relations (17) and (18) are still non-defeasible in our formalism.

6.3 Lin's approach

Differently from McCain and Turner's, Brewka and Hertzberg's and our approaches, Lin considered the causality problem within the situation calculus framework [14], where causality was treated as a predicate called *Caused* [11]. For instance, the formula

$$blocked(duct_1, s) \wedge blocked(duct_2, s) \supset Caused(stuffy, true, s)$$

represents a causal rule saying that if $duct_1$ is blocked and $duct_2$ is blocked, then it will cause the fluent $stuffy$ to be true. The essential principle employed in Lin's formalization for specifying the effects of actions is that "anything changed during a state transition must be *caused*". It means that any change must be related to some causal rules in the causal theory. To derive the indirect effects of actions, a circumscription mechanism is used to minimize predicate *Caused* in the causal theory.

Although Lin's approach is quite different superficially from ours, the basic motivations are similar. Both methods use the minimization idea to specify the state change, and in order to derive the causal effects of actions, both methods employed some nonmonotonic mechanisms (e.g., circumscription and default logic respectively). However, it is not difficult to see that a similar problem with McCain and Turner's method illustrated by Example 9 also exists in Lin's approach. At this moment, it is not clear yet if there exists some formal relationship between Lin's approach and ours.

7 A logic programming approximation

So far, we have not addressed the implementational issue of our approach yet. It is obvious that to put a theory to be applicable, reasonable computational costs should be provided. From the construction of our approach described in section 3, we know that causal theories of actions specified in Definition 9 play a central role in generating resulting states. Furthermore, we have also indicated that our causal theories of actions actually correspond to special forms of default theories. Unfortunately, it is well known that the computation of default theory, even for the propositional case, is intractable generally.

To overcome this difficulty, in this section, we will describe a logic programming approximation for our approach. Specifically, we will translate a causal theory of action defined in Definition 9 into a corresponding extended logic program [5] so that the computation of the causal theory reduces to evaluate a relevant extended logic program. We will also show the the soundness of such translation.

7.1 Extended logic programs: A review

Similarly to the previous case, we restrict our discussion to the propositional case. A *rule* is a statement of the form

$$L_0 \leftarrow L_1, \cdots, L_m, not L_{m+1}, \cdots, not L_n, \tag{20}$$

where L_i $(0 \leq i \leq n)$ is a propositional literal. L_0 is the *head* of the rule, while $L_1, \cdots, L_m, not L_{m+1}, \cdots, not L_n$ is the *body* of the rule. Obviously, the body of a rule could be empty. The connective *not* represents a *negation as failure*. Therefore, (20) may contain two types of negations: classic negation $\neg$ and negation as failure *not*. Informally, rule (20) can be interpreted as follows: if facts $L_1, \cdots, L_m$ are true in our current knowledge base/belief set, and facts $L_{m+1}, \cdots, L_n$ are *not* presented in our current knowledge base/belief set, then fact L_0 is true in our current knowledge base/belief set.

An *extended logic program* Π is a collection of rules [5]. To evaluate extended logic programs, Gelfond and Lifschitz defined an answer set semantics, as presented below. Let Π be an extended logic program not containing *not* and Lit the set of all propositional literals in the language of Π. The *answer set* of Π, denoted as $Ans(\Pi)$, is the smallest subset S of Lit such that

(i) for any rule $L_0 \leftarrow L_1, \cdots, L_m$ from Π, if $L_1, \cdots, L_m \in S$, then $L_0 \in S$;
(ii) if S contains a pair of complementary literals, then $S = Lit$.

Now consider Π be an extended logic program. For any subset S of Lit, let Π^S be the logic program obtained from Π by deleting

(i) each rule that has a formula *not* L in its body with $L \in S$, and
(ii) all formulas of the form *not* L in the bodies of the remaining rules.

We define that S is an *answer set* of Π, denoted $Ans(\Pi)$, iff S is an answer set of Π^S, i.e. $S = Ans(\Pi^S)$. Clearly, an extended logic program can have one, more than one or no answer sets at all. The following example shows how extended logic programs can be used for knowledge representation.

Example 11 Consider an extended logic program Π_1 consisting of the following rules:

$\neg quiet \leftarrow babyawake,\ not\ quiet,$
$quiet \leftarrow babycuddled,\ not\ \neg quiet,$
$babyawake \leftarrow,$
$babycuddled \leftarrow.$

According to the definition above, it is not difficult to conclude that Π_1 has two answer sets namely

$\{babyawake, babycuddled,\ \neg quiet\}$ and
$\{babyawake, babycuddled,\ quiet\}$.

■

7.2 Translating causal theory into extended logic program

Given sets of domain constraints and causal relations $\mathcal{D}$ and $\mathcal{C}$, an initial state S, an action act with post-condition $Post(act)$, and $Max(S, act)$ as defined in Definition 8, a causal theory of action act with respect to S and δ where $\delta \in Max(S, act)$, is specified as $CT^{act}_{S,\delta} = (\mathcal{D} \cup \{Post(act)\} \cup \delta, \mathcal{C})$. To derive the resulting states, we need to compute models of $CT^{act}_{S,\delta}$, i.e. $Mod(CT^{act}_{S,\delta})$. In the following, we will use extended logic programs to present a logic programming approximation for this computation.

Before we present the translation procedure, we need to make some assumptions. In causal theory $CT^{act}_{S,\delta}$, domain constraints $\mathcal{D}$ is explicitly represented. However, as we can see from section 7.1, arbitrary logic forms of constraints are clearly not expressible in extended logic programs. Therefore, we need to assume $\mathcal{D} = \emptyset$ in our translation. Moreover, in order to use extended logic programs to represent causal relations $\mathcal{C}$, we need to restrict each causal relation in $\mathcal{C}$ $A \Rightarrow B$ and $A \rightsquigarrow B$ to the form

$$L_1 \wedge \cdots \wedge L_m \Rightarrow L_0, \tag{21}$$

or

$$L_1 \wedge \cdots \wedge L_m \rightsquigarrow L_0, \tag{22}$$

where $L_0, L_1, \cdots, L_m$ are propositional literals. Finally, we restrict all actions' post-conditions to have the form

$$P_1 \wedge \cdots \wedge P_k, \tag{23}$$

where $P_1, \cdots, P_k$ are propositional literals. In other words, we only consider deterministic actions in our translation. But we will remark in section 8 that this restriction is easily released if we enhance extended logic programs by allowing disjunction occurrence in the head of a rule in logic programs.

With the above assumptions, we then can describe our translation as follows. Let S be a state, *act* an action with post-condition of the form (23), $\mathcal{C}$ a set of causal relations with forms (21) and (22), and $Max(S, act)$ as defined in Definition 8. For each $\delta \in Max(S, act)$, we specify an extended logic program $\Pi^{act}_{S,\delta}$ consisting of the following rules:
Action effect rules: for each P_i occurs in $P_1 \wedge \cdots \wedge P_k$, specify

$P_i \leftarrow$,

Persistent fact rules: for each $L \in \delta$, specify

$L \leftarrow$,

Causal rules: for each causal relation with the form (21) and each causal relation with the form (22) in $\mathcal{C}$, specify

$L_0 \leftarrow L_1, \cdots, L_m$,
$L_0 \leftarrow L_1, \cdots, L_m, not\ \neg L_0$,

respectively.

In the program above, action effect rules represent the post-condition of the action, persistent fact rules, on the other hand, represent facts of S that must be persistent with repsect to this action. From the intuitive semantics of causal relations described in section 2, it is quite obvious to translate non-defeasible causal relation (21) and defeasible causal relation (22) into logic rules $L_0 \leftarrow L_1, \cdots, L_m$ and $L_0 \leftarrow L_1, \cdots, L_m$, $not\ \neg L_0$, respectively. We call program $\Pi^{act}_{S,\delta}$ a *logic program approximation* of causal theory $CT^{act}_{S,\delta} = (\{P_1 \wedge \cdots \wedge P_k\} \cup \delta, \mathcal{C})$. The following result ensures that this approximation preserves the semantics of the original causal theory.

Theorem 3 *Let $\Pi^{act}_{S,\delta}$ be a logic program approximation of causal theory $CT^{act}_{S,\delta}$ as defined above. Then (i) $Mod(CT^{act}_{S,\delta}) = \emptyset$ iff $\Pi^{act}_{S,\delta}$ has a unique answer set Lit or $\Pi^{act}_{S,\delta}$ has no answer set; (ii) $S' \in Mod(CT^{act}_{S,\delta})$ iff there exists some answer set Ans of $\Pi^{act}_{S,\delta}$ such that $Ans \subseteq S'$.*

Proof To prove this theorem, we need to use the following two lemmas proved by Gelfond and Lifschitz, and Łukaszewicz respectively:

Lemma 1 [5] *For any extended logic program* Π,

(a) if Ans is an answer set of Π*, then the deductive closure of Ans is an extension of* Π*;*

(b) every extension of Π *is the deductive closure of exactly one answer set of* Π.

As Gelfond and Lifschitz showed in [5], each rule (20) in Π can be viewed as a default $L_1 \wedge \cdots \wedge L_m : \neg L_{m+1}, \cdots, \neg L_n / L_0$. Therefore, Π can be also viewed as a default theory.

Lemma 2 [12] *Let* $(\mathcal{A}, \mathcal{C})$ *be a default theory,* $Mod(\mathcal{A})$ *the class of all models of* $\mathcal{L}$ *and* Θ *the set of all subclasses of* $Mod(\mathcal{A})$ *such that for each* $\Phi \in \Theta$*, there exists a set of propositional formulas* Π *satisfying the conditions (i), (ii) and (iii) stated in Definition 7. Then*

(a) Φ *is the set of all models of some extension of* $(\mathcal{A}, \mathcal{C})$*;*

(b) for each extension E *of* $(\mathcal{A}, \mathcal{C})$*, there exists some* Φ *in* Θ *such that* Φ *is the set of all models of* E.

Suppose $CT^{act}_{S,\delta} = (\mathcal{A}, \mathcal{C})$. As we mentioned earlier, the causal theory $CT^{act}_{S,\delta}$ can be viewed as a default theory. Note that if we view each fact L in $\mathcal{A}$ as a default $T : /L$, and view $\Pi^{act}_{S,\delta}$ as a default theory by using Gelfond and Lifschitz's notation, then $\Pi^{act}_{S,\delta}$ is eventually identical to $CT^{act}_{S,\delta}$.

We first prove result (i). If $Mod(CT^{act}_{S,\delta}) = \emptyset$, from Lemma 2, it follows that $CT^{act}_{S,\delta}$ does not have an extension or only has an inconsistent extension. Then from Lemma 1, it is easy to see that $\Pi^{act}_{S,\delta}$ does not have an answer set or only has an answer set Lit.

Similarly, we can prove the other direction of result (i).

Now we prove result (ii). If $S' \in Mod(CT^{act}_{S,\delta})$, then from Lemma 2, it follows that S' is a model of an extension E of $CT^{act}_{S,\delta}$. Since $CT^{act}_{S,\delta}$ is identical to default theory $\Pi^{act}_{S,\delta}$, from Lemma 1, it is easy to see that E is the deductive closure of an answer set Ans of $\Pi^{act}_{S,\delta}$. It then follows $Ans \subseteq S'$.

Similarly, we can prove the other direction of result (ii). $\Box$

Having Theorem 3, the logic program approximation $\Pi^{act}_{S,\delta}$ provides a method to compute models of causal theory $CT^{act}_{S,\delta}$. It should be noted that the answer set semantics can be reduced to the *stable model semantics* of logic programs [4] without increasing the computational cost [5]. Although the computation for stable model semantics is still intractable generally, Ben-Eliyahu has recently identified the tractable subclass for computing stable models [1] and Niemela and Simons further implemented some efficient algorithms to compute stable model semantics [16]. Hence, these results are obviously applicable in the implementation of our approach.

Example 12 The light-power domain revisited. We slightly modify the causal relations in Example 7 by specifying $\mathcal{C} = \{on \leadsto light,\ \neg on \Rightarrow \neg light,\ \neg power \Rightarrow \neg light\}$. Let $S = \{light, on, power\}$ be the initial state and suppose an action of cutting off the power (i.e. $Post(act) \equiv \neg power$) is performed on S. Obviously, there exists unique $\delta = \{on\}$ which is a maximal subset of S satisfying the conditions that $\delta \cup \{Post(act)\}$ is consistent and $\delta \not\models \neg light$ and $\delta \not\models light$.

According to the above description, the logic program approximation $\Pi^{act}_{S,\delta}$ of causal theory $CT^{act}_{S,\delta}$ consists of the following rules:

$$
\begin{aligned}
&\neg power \leftarrow, \\
&on \leftarrow, \\
&light \leftarrow on,\ not\ \neg light, \\
&\neg light \leftarrow \neg on, \\
&\neg light \leftarrow \neg power.
\end{aligned}
$$

It is easy to see that $\Pi^{act}_{S,\delta}$ has a unique answer set $\{\neg light, on, \neg power\}$, which is also the unique resulting state. Clearly, this solution coincides with the result obtained in Example 7. ■

8 Conclusions

In this paper, we proposed a new approach for reasoning about action. We have seen that to specify effects of actions properly, we need to represent the domain causality explicitly in action theories, and distinguishing causal relations as defeasible and non-defeasible cases does provide us more flexible capabilities to handle problems of reasoning about action effects.

It should be noticed that specifying a causal relation to be defeasible or non-defeasible is domain-specific. It seems that there is no a general principle to define which causal relation should be defeasible and which should not be. Given the problem domain but with different causal relations, different solutions may be derived even for the same action. For instance, in Example 8 presented in section 4, if a causal relation that the fact that some one is not alive causes the fact that he or she is not walking is specified to be defeasible in the sense that even if a person is not alive, he or she can be moved by someone else, then we will not be able to derive the qualification of an action with effect *walking*. It is interesting to note that if only defeasible causal relations are specified in a domain, all actions then become executable to any state and no action qualifications may be derived (we assume that every action's post-condition is consistent with domain constraints).

The other feature of our approach is that domain constraints still have their conventional functions and logical forms in our formalism. We believe that having explicit representation for domain constraints is important in action theories, as in many situations, domain constraints describe some physical properties about the world, and these properties may not reflect a natural causality of the domain. For instance, in the blocks world, we usually have a domain constraint $\neg on(box, box)$ saying that any block can not be located on itself. The other function of the domain constraint is to force some ramifications during the change that may not be captured by causal relations. For example, if we have a constraint stating that any object can only be located on the table or on the floor, then an action of removing an object from the table will lead to the effect that the object will be on the floor. In such a domain, we do not need to specify an extra causal relation like that "the fact that an object is not on the table causes the fact that the object must be on the floor".

Finally, the logic programming approximation to compute causal theories of actions certainly provides an effective way to implement our approach. Here we should mention that, however, this approximation is not complete with the following restrictions: (a) as arbitrary forms of logic constraints are not expressible in extend logic programs, we therefore require $\mathcal{D} = \emptyset$ for this translation; (b) specific forms (21) and (22) of causal relations are also required; (c) actions should be deterministic. But we would argue that restrictions (a) and (b) are not real limitations from a practical point of view, because inference rules in logic programs (i.e. rules with the form $L_0 \leftarrow L_1, \cdots, L_m$) are usually strong enough to capture the intuitive meanings of the constraints that are used to characterize our domains, while many causal relations can be rewritten to forms (21) and (22). For instance, the causal relation $\neg on \vee \neg power \Rightarrow \neg light$ in Example 7 can be rewritten as two separate causal relations $\neg on \Rightarrow \neg light$ and $\neg power \Rightarrow \neg light$. Restriction (c), on the other hand, can be released if we enhance the extended logic programs by allowing disjunction occurrence in the head of a rule in logic programs [5]. With this augmentation, actions with disjunctive post-conditions will become expressible in extended logic programs.

References

[1] R. Ben-Eliyahu, A hierarchy of tractable subsets for computing stable models. *Journal of Artificial Intelligence Research*, **6** 1996 27–52.

[2] G. Brewka and J. Hertzberg, How do things with worlds: On formalizing actions and plans. *Journal of Logic and Computation*, **3(5)** (1993) 517–532.

[3] J.P. Delgrande, T. Schaub and W.K. Jackson, Alternative approaches to default logic. *Artificial Intelligence*, **70** (1994) 167–237.

[4] M. Gelfond and V. Lifschitz, The stable model semantics for logic programming. In *Proceedings of the 5 the International Conference on Logic Programming.* MIT Press, (1988) 1070–1080.

[5] M. Gelfond and V. Lifschitz, Classical negation in logic programs and disjunctive databases. *New Generation Computing*, **9** (1991) 365–386.

[6] M.L. Ginsberg and D.E. Smith, Reasoning about action I: A possible worlds approach. In *Proceedings of the 1987 Workshop on the Frame Problem in Artificial Intelligence.* F.M. Brown ed. Morgan Kaufmann Publishers, Inc. (1987) 233–258.

[7] P. Hayes, A logic of actions. *Machine Intelligence*, **13** (1971) 459–520.

[8] G.N. Kartha and V. Lifschitz, Actions with indirect effects (preliminary report). In *Proceeding of the Fourth International Conference on Principles of Knowledge Representation and Reasoning.* Morgan Kaufmann Publishers, Inc. (1994) 341–350.

[9] V. Lifschitz, Formal theories of action. In *Proceeding of the Tenth International Joint Conference on Artificial Intelligence (IJCAI'87).* Morgan Kaufmann Publishers, Inc. (1987) 966–972.

[10] V. Lifschitz, Frames in the space of situations. *Artificial Intelligence*, **46** (1990) 365–376.

[11] F. Lin, Embracing causality in specifying the indirect effects of actions. In *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI'95).* Morgan Kaufmann Publishers, Inc. (1995) 1985–1993.

[12] W. Lukaszewicz, *Non-Monotonic Reasoning.* Ellis Horwood, 1990.

[13] N. McCain and H. Turner, A causal theory of ramifications and qualifications. In *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI'95).* Morgan Kaufmann Publishers, Inc. (1995) 1978–1984.

[14] J. McCarthy and P. Hayes, Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, **4** (1969) 463–502.

[15] K. Myers and D. Smith, The persistence of derived information. In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI'88).* Morgan Kaufmann Publishers, Inc. (1988) 496–500.

[16] I. Niemela and P. Simons, Efficient implementation of the well-founded and stable model semantics. In *Proceedings of the 1996 Joint International Conference and Symposium on Logic Programming.* MIT Press (1996) 289–303.

[17] R. Reiter, A logic for default reasoning. *Artificial Intelligence*, **13** (1980) 81–132.

[18] M. Winslett, Reasoning about action using a possible models approach. In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI'88)*. Morgan Kaufmann Publishers, Inc. (1988) 89–93.

[19] Y. Zhang, Compiling causality into action theories. In *Proceedings of the Third Symposium on Logical Formalizations of Commonsense Reasoning (COMMONSENSE'96)*. (1996) 261–270. Stanford.