In *Challenges to Decision Support in a Changing World, Papers from the 2005 AAAI Spring Symposium*, Marek J. Druzdzel and Tze-Yun Leong (eds.), Technical Report SS-05-02, pages 73-79, Menlo Park, CA: AAAI Press, 2005

Mechanism-based Causal Models for Adaptive Decision Support

Tsai-Ching Lu
HRL Laboratories, LLC
Malibu, CA 90265
tlu@hrl.com

Marek J. Druzdzel
Decision Systems Laboratory
School of Information Sciences and
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
marek@sis.pitt.edu

Abstract

We propose a framework for decision support in a changing world that rests on mechanism-based causal models. A causal model is a self-contained set of simultaneous structural equations representing a decision problem. In our framework, a causal model is instantiated from a mechanism knowledge base, where domain knowledge is stored as mechanisms along with generic non-intervening and intervening actions. A change in the world is modeled as applying an action on the causal model. The framework supports reasoning about the effect of actions to deduce the causal model representing the world brought about by the actions.

Introduction

Probabilistic graphical models such as Bayesian networks and influence diagrams have been successfully applied in decision support systems where reasoning about uncertainty is essential. There also exist learning algorithms that enable such systems to update the model structure and its parameters to integrate the data observed in the field. However, to the best of our knowledge, there has been no work on how often a model should be updated and what events should trigger such updates. Furthermore, modifying a model to reflect changes resulting from intervening actions has not been studied either, as plain probabilistic graphical models do not encode information required for predicting the effects of intervening actions.

In this paper, we propose a framework based on causal mechanisms, for reasoning about changes resulting from intervening and non-intervening actions. A model based on causal mechanisms is a set of self-contained structural equations, each of which represents a causal mechanism active in a modeled system. A causal model is constructed by instantiating mechanisms from mechanism knowledge bases, where mechanisms are represented as structural equations and actions are represented as local modifications on causal models. We model changes to a system under study by actions applied to the corresponding causal model. This framework provides support for reasoning about what a system will be if an action is applied to it. Decision makers can, therefore, predict effects of actions executed by other agents or find the most effective sequence of actions to perform in order to reach a given objective.

Causal Models

A causal model is a set of self-contained structural equations, each of which represents a distinct mechanism active in a system.[1] More formally, we denote the set of variables appearing in an equation e as $\mathbf{Vars}(e)$, and in a set of equations $\mathbf{E}$ as $\mathbf{Vars}(\mathbf{E}) \equiv \cup_{e \in \mathbf{E}} \mathbf{Vars}(e)$. A causal model $M = \langle \mathbf{X}, \mathbf{E} \rangle$ consists of a set of self-contained structural equations $\mathbf{E}$ over a set of variables $\mathbf{X} \equiv \mathbf{Vars}(\mathbf{E})$. Each structural equation $e \in \mathbf{E}$ is generally written in its implicit form $e(X_1, X_2, \ldots, X_n) = 0$ where $X_i \in \mathbf{Vars}(e)$. We say that a variable $X_i \in \mathbf{X}$ is *exogenous* to M if its value is determined by factors outside the system, i.e., if there exists a structural equation $e \in \mathbf{E}$, $e(X_i) = 0$, and *endogenous* otherwise. In other words, the set of variables $\mathbf{X}$ consists of two disjoint sets: exogenous variables $\mathbf{U}$ and endogenous variables $\mathbf{V}$. Therefore, a causal model is sometimes denoted as $M = \langle \mathbf{U}, \mathbf{V}, \mathbf{E} \rangle$. Let $\mathsf{D}(X_i)$ be the domain of a variable X_i, and $\mathsf{D}(\mathbf{X}) = \mathsf{D}(X_1) \times \ldots \times \mathsf{D}(X_n)$ be the domain of the set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$. Given $\mathbf{u} \in \mathsf{D}(\mathbf{U})$, the solutions for endogenous variables $\mathbf{Y} \subseteq \mathbf{V}$, denoted as $\mathbf{Y}_M(\mathbf{u})$ or $\mathbf{Y}(\mathbf{u})$, in a causal model M can always be determined uniquely. The pair $\langle M, \mathbf{u} \rangle$ is called a *causal world,* or simply world. Given a probability distribution $\Pr(\mathbf{u})$ defined over $\mathsf{D}(\mathbf{U})$, the pair $\langle M, \Pr(\mathbf{u}) \rangle$ is called a *probabilistic* causal model where for each $Y \in \mathbf{V}$, $\Pr(Y = y) \triangleq \sum_{\{\mathbf{u} | Y(\mathbf{u}) = y\}} \Pr(\mathbf{u})$.

Simon (1953) developed an algorithm that explicates the asymmetries among variables in a causal model M and represents them as a *causal graph* $G(M)$. A causal model M is *recursive* if the associated $G(M)$ is a directed acyclic graph, where each node corresponds to a variable, and each family (a node with its parents in $G(M)$) a structural equation (Druzdzel & Simon 1993). In other words, each structural equation $e(X_1, \ldots, X_n) = 0$ is expressed in its explicit functional form $X_i = f_{X_i}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and is depicted graphically as a family with arcs from nodes representing argu-

[1] Please see Pearl (2000) or Spirtes et al. (2000) for general introductions to causal models.

ments of f_{X_i} (i.e., $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$) to X_i. A causal model M is *non-recursive* if there exists strongly-coupled components in a causal graph $G(M)$ generated by the causal ordering algorithm. This indicates that variables in a strongly-coupled components need to be solved simultaneously. Simon's causal ordering algorithm is in the worst case exponential time algorithm. Lu (2003) developed the worst case polynomial time algorithm COA_{BGM}, based on bipartite graph matching, to generate causal graphs for self-contained and under-constrained sets of structural equations.

Reversibility

A *Reversible mechanism* is a mechanism in which causal relations among variables may be reversed when it is embedded in different systems. Traditionally, reversible mechanisms are discussed mainly in deterministic mechanical and physical relations (Wold & Jureen 1953, pp. 325), since the invertibility of a function is a necessary condition for the reversibility. A functional relation may be reversible in *functional* sense, but may not be reversible in *causal* sense. For example, ideal gas law and Ohm's law are given in (Wold & Jureen 1953, pp. 40) and (Nayak 1994, pp. 10) respectively as examples of only partially reversible mechanisms, even though their functional relations are invertible. Druzdzel and van Leijen (2001) demonstrated that under some special conditions probability distribution tables in a causal Bayesian network can be reversed in both functional and causal senses.

We explicitly represent the reversibility of a mechanism to assist predictions of the effect of actions. We define the reversibility of a mechanism semantically on the set of possible effect variables of a mechanism. Assuming that the number of variables in a mechanism is finite, the number of possible effect variables for a mechanism is also finite. We can classify mechanisms into four categories according to their *reversibility*: (1) *completely reversible* (CR): every variable in the mechanism can be an effect variable, (2) *partially reversible* (PR): some of the variables in the mechanism can be effect variables, (3) *irreversible* (IR): exactly one of the variables in the mechanism can be an effect variable, and (4) *unknown* (UN): the reversibility of the mechanism is unspecified, i.e., the modeler only asserts that variables in a mechanism are relevant, but has not yet resolved how they relate to each other causally.

Definition 1 (reversibility)
Let $\mathbf{Vars}(e)$ *be the set of variables in a structural equation* e*. Let* $\mathbf{EfVars}(e) \subseteq \mathbf{Vars}(e)$ *be the set of all possible effect variables in a structural equation* e*. The reversibility of a mechanism represented by* e *is*

1. completely reversible *if* $\mathbf{EfVars}(e) = \mathbf{Vars}(e)$ *and* $|\mathbf{EfVars}(e)| > 1$,
2. partially reversible *if* $1 < |\mathbf{EfVars}(e)| < |\mathbf{Vars}(e)|$,
3. irreversible *if* $|\mathbf{EfVars}(e)| = 1$, *and*
4. unknown *if* $|\mathbf{EfVars}(e)| = \emptyset$.

We emphasize that the notion of reversibility of a mechanism is a semantic one since it is defined with respect to the set of effect variables of a mechanism with a-priori assumptions. In addition, reversibility is defined as the property of a mechanism, but not as a derived property of a mechanism when it is embedded in a system. In other words, the set of effect variables of a mechanism is assumed a-priori before we decide which system the mechanism will be embedded. Which effect variable is *active* will be determined as soon as we know which system the mechanism has been embedded.

The Framework

Given a causal model which represents the system of a decision problem, we aim to develop a framework that can perform the following reasoning: (1) given future events on the modeled system, the framework can automatically revise the causal model to reflect the changes such that queries on the projected system can be answered, (2) given decision objectives for the modeled system, the framework can automatically recommend a sequence of actions such that the system is likely to yield desired outcomes, and (3) given a history of events happened to the modeled system, the framework can automatically recover likely causal models such that the origin of the system can be examined.

We conceive changes to a system as applying intervening or non-intervening actions on the corresponding causal model. We first define an action operator that can reason about the effect of actions on a causal model. Second, we discern the difference among local effects, persistence and response brought about by an action. Third, we introduce the generic action, by which users can specify local effects of an action at domain level, and an algorithm that can infer persistence and response to instantiate a generic action for a particular causal model.

Action Operator

In the recent literature, Pearl (2000, pp. 225) suggested to use the notation $do(q)$, where q is a proposition, to denote an action, since people use phrases such as "reduce tax" in daily language to express actions. More precisely, an *atomic action*, denoted as $do(V = v)$ in Pearl (2000) and $manipulate(v)$ in Spirtes et al. (2000), is invoked by an external force that manipulates on a variable V by imposing a probability distribution or holding it at a constant value, v, and replacing the causal mechanism, $V = f(PA(V))$, which directly governs V in a causal model. The corresponding change in the causal graph is depicted as the arc-cutting operation by which all incoming arcs to the manipulated variable V are removed (Spirtes, Glymour, & Scheines 2000; Pearl 2000). Notice that the implicit assumption behind the arc-cutting operation is that the manipulated variable is governed by an *irreversible* mechanism, i.e., only V can be an effect variable in mechanism $e(V, PA(V)) = 0$. In order to ensure that the manipulated causal model is self-contained, the irreversible mechanism, which governs the manipulated variable in the model before manipulation, has to be replaced. However, when the manipulated variable is governed by a *reversible* mechanism, the manipulated model derived from the arc-cutting operation may not be consistent with our conception of the manipulated system. We argue

that an action in causal modeling should be defined at the level of mechanisms, not at the level of propositions.

In econometric literature (Simon 1953; Strotz & Wold 1960), a system is represented as a SEM, a set of structural equations, and actions are modeled as "scraping invalid equations" and "replacing them by new ones." In STRIPS language (Fikes & Nilsson 1971), a situation is represented by a state, conjunctions of function-free ground literals (propositions), and actions are represented as PRE, ADD, and DEL lists which are conjunctions of literals. There is a clear analogy between these two modeling formalisms, where the effects of actions are modeled explicitly as adding or deleting fundamental building blocks, which are structural equations in SEM and propositions in STRIPS. We propose to explicitly translate the operations of "scraping invalid equations" as specifying equations in $\mathbf{E}_{\mathbf{del}}$ list and "replacing them by new ones" as specifying equations in $\mathbf{E}_{\mathbf{add}}$ and define an action operator as follows.

Definition 2 (action operator)
An action operator $\mathbf{Act}(\mathbf{E}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$ *represents an action on a system represented by a SEM* $\mathbf{E}$*, where* $\mathbf{E}_{\mathbf{pre}}$ *is the precondition of the action, and* $\mathbf{E}_{\mathbf{add}}$ *and* $\mathbf{E}_{\mathbf{del}}$ *are the changes brought about by the action on* $\mathbf{E}$ *:*

1. $\mathbf{E}_{\mathbf{pre}}$*: a set of conditions that must be true before the action can be applied to* $\mathbf{E}$*.*
2. $\mathbf{E}_{\mathbf{add}}$*: a set of structural equations added to* $\mathbf{E}$*.*
3. $\mathbf{E}_{\mathbf{del}}$*: a set of structural equations removed from* $\mathbf{E}$*.*

Given an action $A = \mathbf{Act}(\mathbf{E}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$ on a SEM $\mathbf{E}$, the modified model $\mathbf{E}_A$ is a set of structural equations $\mathbf{E}_A = \mathbf{E} \cup \mathbf{E}_{\mathbf{add}} \setminus \mathbf{E}_{\mathbf{del}}$. We will see in the later sections how this definition of action operator allows us to reason about the effect of actions with reversible mechanisms.

Actions in general will bring new mechanisms into a system; however, non-intervening actions will not remove any mechanisms from a system. In other words, a non-intervening action $\mathbf{Act}(\mathbf{E}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$ by definition will have $\mathbf{E}_{\mathbf{del}} = \emptyset$.

Since an action $\mathbf{Act}(\mathbf{E}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$ can be applied to a SEM $\mathbf{E}$ as long as the precondition $\mathbf{E}_{\mathbf{pre}}$ is satisfied, the modified model $\mathbf{E}_A$ is not necessary self-contained after the manipulation even though $\mathbf{E}$ is self-contained. $\mathbf{E}_A$ could be under-constrained or over-constrained, depending on the structural equations specified in $\mathbf{E}_{\mathbf{add}}$ and $\mathbf{E}_{\mathbf{del}}$. To support the action deliberation in the following sections, we say an action $A = \mathbf{Act}(\mathbf{E}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$ on a system represented by a SEM $\mathbf{E}$ is *self-contained* if the manipulated model $\mathbf{E}_A$ is self-contained and indeed represents the manipulated system. In other words, applying a self-contained action on a self-contained model will result in a self-contained manipulated model. For example, an atomic action defined in (Spirtes, Glymour, & Scheines 2000; Pearl 2000) is one of self-contained actions in our definition.

Persistence and Response

Given the action operator defined in Definition 2, one may immediately raise the question about how to specify an action that is not subject to a specific model. We will derive such specification by the following example.

Consider a causal model that describes the relations among heart disease (HD), blood pressure (BP), and headache (HA) as mechanisms $f_{HD}(HD) = 0$, $f_{BP}(HD, BP) = 0$, and $f_{HA}(BP, HA) = 0$. Assume that a patient's utility ($Utility$) directly depends on headache (HA). The causal graph for this example is depicted in Figure 1(a) and the corresponding set of structural equations is listed as follows:

$$\begin{cases} f_{HD}(HD) = 0 \\ f_{BP}(BP, HD) = 0 \\ f_{HA}(HA, BP) = 0 \\ f_{Utility}(Utility, HA) = 0 \end{cases}$$

An example of a non-intervening action would be measuring blood pressure (MBP), which brings the variable blood pressure reading (BPR) and the mechanism describing how the blood pressure is measured, $f_{BPR}(BP, BPR, MBP) = 0$, into the model. We represent the cost of measuring blood pressure ($\mathsf{CO}(MBP)$) as a value function of MBP, $\mathsf{U}(MBP)$. The non-intervening action is represented as $A_{MBP} \triangleq \mathbf{Act}(\mathbf{E}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$ where $\mathbf{E}_{\mathbf{pre}} = \{BP \in \mathbf{Vars}(\mathbf{E})\}$, $\mathbf{E}_{\mathbf{add}} = \{f_{BPR}, f_{\mathsf{CO}(MBP)}\}$, and $\mathbf{E}_{\mathbf{del}} = \emptyset$. The modified model $\mathbf{E}_{\mathbf{A}_{MBP}}$ is shown as follows.

$$\begin{cases} f_{HD}(HD) = 0 \\ f_{BP}(BP, HD) = 0 \\ f_{HA}(HA, BP) = 0 \\ f_{Utility}(Utility, HA) = 0 \\ f_{BPR}(BPR, MBP, BP) = 0 \\ f_{MBP}(MBP) = 0 \\ f_{\mathsf{CO}(MBP)}(\mathsf{CO}(MBP), MBP) = 0 \end{cases}$$

We see that A_{MBP} brought f_{BPR} and $f_{\mathsf{CO}(MBP)}$ into the model but did not intervene into any of the existing mechanisms: f_{HD}, f_{BP}, f_{HA}, and $f_{Utility}$. Furthermore, A_{MBP} is applicable only when the variable is about to be observed is in the model, i.e., $\mathbf{E}_{\mathbf{pre}} = \{BP \in \mathbf{Vars}(\mathbf{E})\}$.

After examining the reading of a patient's blood pressure, a doctor may prescribe a medicine to control the blood pressure such that the symptom of headache can be eased. Consequently, we need to augment the model to represent the *persistency* of heart disease, which has not been treated, and the *response* of blood pressure and headache relative to the prescribed blood pressure control medicine. It is common sense that the previous reading of blood pressure becomes invalid after taking the blood pressure control medicine. However, the reading of blood pressure before taking the medicine should affect our belief of the severity of the heart disease and its persistence.

To model the intervening action of taking a blood pressure control medicine (D_{bp}) conditioned on the non-intervening action (MBP) and its reading (BPR), we apply the intervening action $A_{D_{bp}} \triangleq \mathbf{Act}(\mathbf{E}_{\mathbf{A}_{MBP}}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$ where $\mathbf{E}_{\mathbf{pre}} = \{MBP \in \mathbf{Vars}(\mathbf{E}) \wedge MBP = true, BPR \in \mathbf{Vars}(\mathbf{E}) \wedge BPR = high, BP \in \mathbf{Vars}(\mathbf{E})\}$, $\mathbf{E}_{\mathbf{add}} = \{f_{D_{bp}}, f_{HD'}, f_{BP'}, f_{HA'}, f_{Utility'}, f_{\mathsf{CI}(D_{bp})}\}$, and $\mathbf{E}_{\mathbf{del}} = \{f_{HA}, f_{Utility}\}$. The modified model is shown as follows.

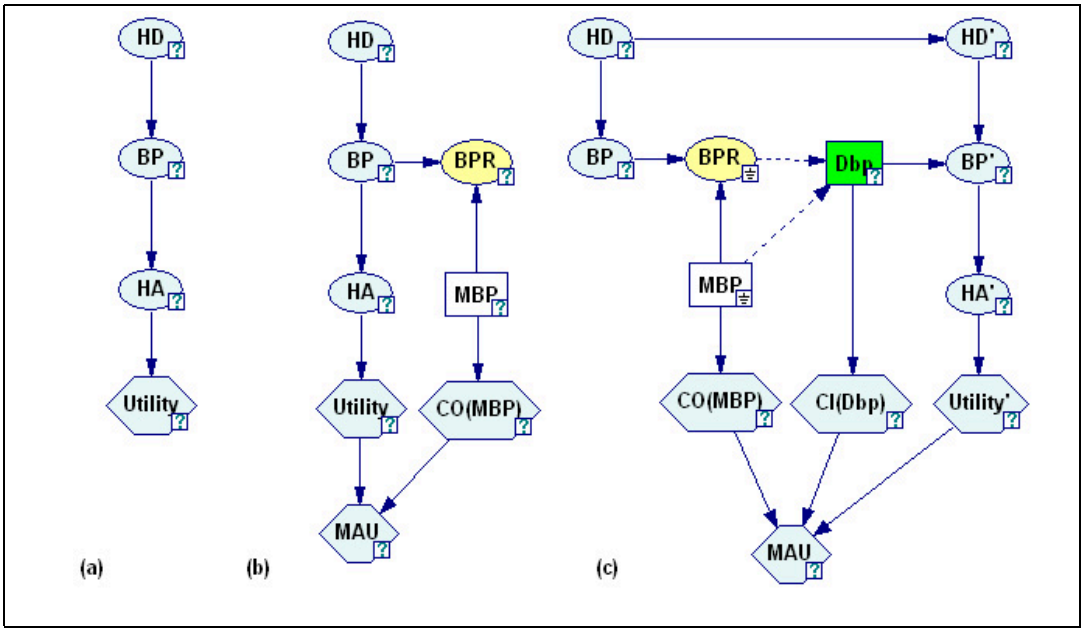

Figure 1: (a) depicts relations among heart disease (HD), blood pressure (BP), and headache (HA); (b) depicts the augmented model for the non-intervening action – measuring blood pressure (MBR), its reading (BPR) and cost ($\mathsf{CO}(MBR)$); (c) depicts the augmented model for considering taking the blood pressure control medicine (D_{bp}) and its cost ($\mathsf{CI}(D_{bp})$), after measuring the blood pressure

$$\begin{cases} f_{HD}(HD) = 0 \\ f_{BP}(BP, HD) = 0 \\ f_{BPR}(BPR = high, MBP = true, BP) = 0 \\ f_{MBP}(MBP) = 0 \\ f_{\mathsf{CO}(MBP)}(\mathsf{CO}(MBP), MBP) = 0 \\ f_{D_{bp}}(D_{bp}, BPR = high, MBP = true) = 0 \\ f_{HD'}(HD', HD) = 0 \\ f_{BP'}(BP', HD', D_{bp}) = 0 \\ f_{HA'}(HA', BP') = 0 \\ f_{Utility'}(Utility', HA') = 0 \\ f_{\mathsf{CI}(D_{bp})}(\mathsf{CI}(D_{bp}), D_{bp}) = 0 \end{cases}$$

As we can see, the persistency of heart disease between time slice is modeled by the structural equation $f_{HD'}$. The previous reading of blood pressure ($BPR = high$) updates our belief of the severity of heart disease in the next time slice (HD') through the path of persistence ($HD \rightarrow HD'$). The decision of taking blood pressure control medicine (D_{bp}) depends on the decision of measuring the blood pressure ($MBP = true$) and the reading of blood pressure ($BPR = high$). The blood pressure after the intervention is governed by the structural equation $f_{BP'}$. The headache at the next time slice (HA') is governed by the structural equation $f_{HA'}$. The causal graph for the manipulate system is shown in Figure 1(c).

If we examine the actions A_{MBP} and $A_{D_{bp}}$, we shall find the difference in their applicabilities. The non-intervening action A_{MBP} is applicable to all models in the same domain as long as the precondition $\mathbf{E_{pre}} = \{BP \in \mathbf{Vars}(\mathbf{E})\}$ is held, since it simply brings in new mechanisms into a model without intervening on the rest. On the other hand, we can see that $A_{D_{bp}}$ is not directly applicable to all models in the domain, since we have represented the *persistence* and *response* into $\mathbf{E_{add}}$ and $\mathbf{E_{del}}$ of $A_{D_{bp}}$. In order to make the intervening action $A_{D_{bp}}$ model independent, we shall specify the relations that are locally relevant to $A_{D_{bp}}$, namely $f_{D_{bp}}$, $f_{BP'}$, and $f_{\mathsf{CI}(D_{bp})}$ in $\mathbf{E_{add}}$. The relation representing persistence, $f_{HD'}$ in $\mathbf{E_{add}}$, should be inferred from $\mathbf{E}_{A_{MBP}}$. Similarly, the relations representing response, $f_{HA'}$ and $f_{Utility'}$ in $\mathbf{E_{add}}$ and the $\mathbf{E_{del}}$ of $A_{D_{bp}}$, should be inferred from $\mathbf{E}_{A_{MBP}}$.

Generic Action

To represent actions that are applicable in the same domain for different models, we introduce the representations of persistence relations and generic actions into our mechanism knowledge bases. A *persistence relation* is represented by a structural equation in the form of $e_{X_i'}(X_i', \mathbf{X_{pre}}) = 0$ where $\mathbf{X_{pre}}$ is a set of exogenous variables and $X_i \in \mathbf{X_{pre}}$. The persistence relation describes how a subset of variables in the current time slice, $\mathbf{X_{pre}}$, affects the variable X_i' at the next time slice. We represent *generic actions* in a knowledge base as $\mathbb{A}_{X_i} \triangleq \mathbf{Act}(\mathbb{E}, \mathbb{E}_{\mathbf{pre}}, \mathbb{E}_{\mathbf{add}}, \mathbb{E}_{\mathbf{del}})$ where $\mathbb{A}_{X_i}$ will be instantiated into $A_{X_i} \triangleq \mathbf{Act}(\mathbf{E}, \mathbf{E_{pre}}, \mathbf{E_{add}}, \mathbf{E_{del}})$ when $\mathbb{A}_{X_i}$ is about to be applied on a model $\mathbf{E}$; $\mathbb{E}_{pre}$ is the precondition that can invoke $\mathbb{A}_{X_i}$; $\mathbb{E}_{\mathbf{add}}$ consists of local mechanisms that will be brought about by the action $\mathbb{A}_{X_i}$ and will be augmented into $\mathbf{E_{add}}$ when the action $\mathbb{A}_{X_i}$ is instantiated on a specific model $\mathbf{E}$; similarly $\mathbb{E}_{\mathbf{del}}$ is initially an empty set and will be instantiated into $\mathbf{E_{del}}$ when the ac-

tion $\mathbb{A}_{X_i}$ is instantiated. Please note that $X_i \in \mathbf{Vars}(\mathbb{E})$ should be in $\mathbb{E}_{\mathbf{pre}}$ as default, since the action will be applied to the variable X_i.

Given the definitions of persistence relations and generic actions, we introduce the procedure for instantiating a generic action. We first explain the use of time slice in our modeling. The concept of time slice in our models is derived from the necessity of reasoning about the effects of intervening actions. Consider the blood pressure variables BP and BP' in Figure 1. Both variables refer to the blood pressures of a patient; however, they refer to the blood pressures of a patient before and after taking the blood pressure control medicine (D_{bp}). Such distinguishment allows us to correctly specify the intervening action D_{bp} which is indirectly conditioned by the BP, and directly influences the "same" variable BP'. In other words, when we apply an intervening action on a direct or an indirect cause of an observed variable, by which the intervening action is conditioned, we trigger the procedure to model the system into two consequent times slices. How to model the system into two consequent time slices is accomplished by the following inferred modeling of persistence and response.

After deciding on modeling a system into two consequent time slices, the procedure needs to infer on which variables the modeling of persistence relations should be applied. Since all endogenous variables are determined within the model, the procedure only applies persistence relations for exogenous variables at the next time slice, except the case that an exogenous variable is the manipulated variable. We emphasize that persistence relations should be specified purely by their *evolutional influences* with respect to their domain. Please also note that the persistence relations serve as the path way for carrying the information observed from the current time slice to the next time slice.

The procedure also needs to infer the responses brought about by a generic action. Modeling a system into two consequent time slices is one of the responses, for which the procedure will first copy the existing mechanisms in the current time slice into the next time slice and link two time slices by persistence relations, i.e., adding all these newly created mechanisms into $\mathbf{E}_{\mathbf{add}}$. When copying existing mechanisms into the next time slice, the procedure will leave out (1) mechanisms of invalid observations at the current time slice, which are mechanisms associated with descendants of the manipulated variable (e.g., f_{BPR}, f_{MBP}, and $f_{\mathsf{CO}(MBP)}$ in Figure 1(b)), and (2) mechanisms for exogenous variables which have newly created persistence relations (e.g., f_{HD} in Figure 1(b)). Finally,the procedure will also remove those mechanisms that are not needed for decisions at the next time slice from the model in the current time slice. Such mechanisms are those govern the nodes which are not ancestors of the evidences and are d-separated from the evidences given the manipulated variable (e.g., f_{HA} and $f_{Utility}$ in Figure 1(b)).

We outline the procedure for instantiating generic intervening actions on systems containing reversible mechanisms in Figure 2. The procedure $InstantiateActionRev$ takes a reversible system $\mathbf{E}$, a mechanism knowledge base $\mathcal{K}$, and a generic intervening action $\mathbb{A}_{X_i} \triangleq \mathbf{Act}(\mathbb{E}, \mathbb{E}_{\mathbf{pre}}, \mathbb{E}_{\mathbf{add}}, \mathbb{E}_{\mathbf{del}})$ in $\mathcal{K}$ as inputs and outputs an instantiated action $A_{X_i} \triangleq \mathbf{Act}(\mathbf{E}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$ when the precondition of $\mathbb{A}_{X_i}$ is satisfied. In Lines 6-16, mechanisms for the next time slice are added into $\mathbf{E}_{\mathbf{add}}$. In Line 8, mechanisms for persistence relations are added into $\mathbf{E}_{\mathbf{add}}$. In Lines 17-22, valid observations are created into the sys-

Procedure $InstantiateActionRev(\mathbf{E}, \mathcal{K}, \mathbb{A}_{X_i})$

Input: A reversible system $\mathbf{E}$, a mechanism knowledge base $\mathcal{K}$, and a generic action $\mathbb{A}_{X_i} \triangleq \mathbf{Act}(\mathbb{E}, \mathbb{E}_{\mathbf{pre}}, \mathbb{E}_{\mathbf{add}}, \mathbb{E}_{\mathbf{del}})$ in $\mathcal{K}$.

Output: `true`: $\mathbb{A}_{X_i}$ is instantiated action into $A_{X_i} \triangleq \mathbf{Act}(\mathbf{E}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$; or `false`.

1. $\mathbb{E} := \mathbf{E}$; $\mathbf{E}_{\mathbf{pre}} := \mathbb{E}_{\mathbf{pre}}$; $\mathbf{E}_{\mathbf{add}} := \emptyset$; $\mathbf{E}_{\mathbf{del}} := \emptyset$;
2. **if** ($\mathbf{E}_{\mathbf{pre}}$ is not `true` in $\mathbf{E}$) **return** `false`; **end if**
3. Apply COA_{BGM} on $\mathbf{E}$ and generate the corresponding graph $G(\mathbf{E}) = \langle \mathbf{N}, \mathbf{A} \rangle$;
4. Let $\mathbf{O} \subset \mathbf{N}$ be the set of all observed variables in $\mathbf{Vars}(\mathbf{E})$;
5. $\mathbf{P} := \emptyset$;
6. **for each** $N_j \in \mathbf{N}$ where $N_j \neq X_i$ and $N_j \notin \mathbf{O}$
7. **if** ($N_j \in \mathbf{ExVars}(\mathbf{E})$ and $\exists e_{N'_j}$ in $\mathcal{K}$)
8. $\mathbf{E}_{add} := \mathbf{E}_{add} \cup e_{N'_j}$; $\mathbf{P} := \mathbf{P} \cup N_j$;
9. **else if** ($N_j = \mathbf{Vars}(\mathbb{E}_{\mathbf{add}})$)
10. Copy e_{N_j} into $e_{N'_j}$ where $e_{N_j} \in \mathbb{E}_{\mathbf{add}}$;
11. $\mathbf{E}_{add} := \mathbf{E}_{add} \cup e_{N'_j}$;
12. **else if** ($N_j \neq N_i$ where $e_{N_i} \in \mathbb{E}_{\mathbf{del}}$)
13. Copy e_{N_j} into $e_{N'_j}$ where e_{N_j} in $\langle e_{N_j,N_j} \rangle$;
14. $\mathbf{E}_{add} := \mathbf{E}_{add} \cup e_{N'_j}$;
15. **end if**
16. **end for each**
17. **for each** ($N_j \in \mathbf{O}$)
18. **if** ($N_j \notin DES(X_i)$) and
19. no ($\exists P_i \in ANS(N_j)$ and $P_i \in \mathbf{P}$)
20. Copy e_{N_j} into $e_{N'_j}$ where e_{N_j} in $\langle e_{N_j,N_j} \rangle$;
21. $\mathbf{E}_{add} := \mathbf{E}_{add} \cup e_{N'_j}$; **end if**
22. **end for each**
23. Find $\mathbf{D} \subset \mathbf{N}$ where $\mathbf{D} \nsubseteq ANS(X_i)$ and $Independent(\mathbf{D}, \mathbf{O}|X_i)$;
24. **for each** $D_j \in \mathbf{D}$
25. $\mathbf{E}_{del} := \mathbf{E}_{del} \cup e_{D_j}$ where $e_{D_j} \in \mathbf{E}$ and
26. e_{D_j} is the mapping of D_j in $\langle e_{D_j}, D_j \rangle$.
27. **end for each**
28. **return** `true` with $A_{X_i} \triangleq \mathbf{Act}(\mathbf{E}, \mathbf{E}_{\mathbf{pre}}, \mathbf{E}_{\mathbf{add}}, \mathbf{E}_{\mathbf{del}})$

Figure 2: Procedure for instantiating a generic intervening action $\mathbb{A}_{X_i}$ on a system containing reversible mechanisms $\mathbf{E}$ using knowledge in $\mathcal{K}$

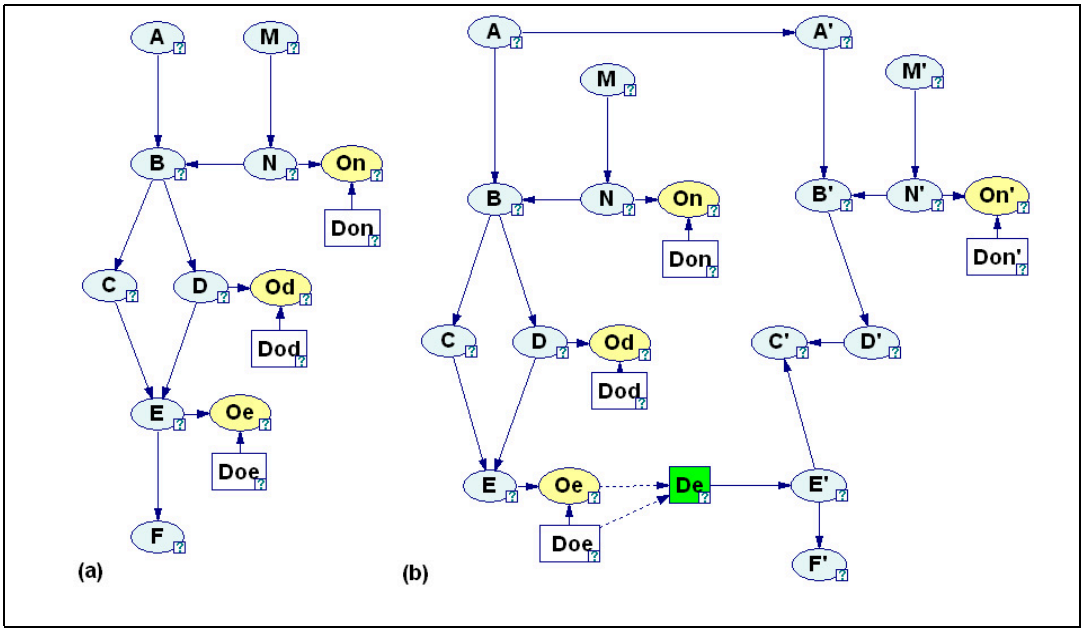

Figure 3: (a) depicts relations among variables A, B, C, D, E, F, M, N and non-intervening actions $D_{O_n}, D_{O_d}, D_{O_e}$ and their corresponding observations O_n, O_d, O_e. (b) depicts the augmented model for considering the intervening action which manipulates on E and releases the mechanism governing C. The system in (a) is augmented into the next time slice in (b). There is a persistence relation for A defined in knowledge base, but no persistence relation for M. The valid observation O_n is therefore kept in the next time slice

tem in the next time slice. In Lines 24-27, mechanisms that are independent to the decision at the next time slice are added into $\mathbf{E}_{\mathbf{del}}$. The procedure is worst-case polynomial time due to Line 3.

Figure 3(a) shows an example of instantiating a generic action on a reversible system. The model describes relations among variables A, B, C, D, E, F, M, N with non-intervening actions $D_{O_n}, D_{O_d}, D_{O_e}$ and their corresponding observations O_n, O_d, O_e. For the sake of presentation, we do not include mechanisms describing utility functions in the system. Consider the generic intervening action $\mathbb{A}_E \triangleq \mathbf{Act}(\mathbb{E}, \mathbb{E}_{\mathbf{pre}}, \mathbb{E}_{\mathbf{add}}, \mathbb{E}_{\mathbf{del}})$ where $\mathbb{E}_{\mathbf{pre}} = \{O_e \in \mathbf{Vars}(\mathbf{E}) \wedge D_{O_e} = \texttt{true}\}$, $\mathbb{E}_{\mathbf{add}} = \{f_E\}$, $\mathbb{E}_{\mathbf{del}} = \{f_C\}$. When we instantiate this generic action on the model $\mathbf{E}$ depicted in Figure 3(a), we add $f_{A'}$ into $\mathbf{E}_{\mathbf{add}}$ by persistence relations in $\mathcal{K}$. With respect to the response, we add $f_{M'}, f_{N'}, f_{B'}, f_{D'}, f_{F'}, f_{O'_n}, f_{D'_{O_n}}$ into $\mathbf{E}_{\mathbf{add}}$ by copying them directly from their corresponding mechanisms in previous time slice. We copy f_E into $\mathbf{E}_{\mathbf{add}}$ as $f_{C'}$ and do not copy f_C into $\mathbf{E}_{\mathbf{add}}$. And f_E specified in $\mathbb{E}_{\mathbf{add}}$ is instantiated into $f_{E'}$ which is added into $\mathbf{E}_{\mathbf{add}}$ together with f_{D_e} .

Discussion

In this paper, we viewed changes as actions on a causal model and defined generic actions as structural modifications on local mechanisms. We provided a procedure for instantiating a generic action into an actual action for a given causal model and mechanism knowledge bases. By applying such an instantiated action on a causal model, one can reason about the effect of actions. In other words, if one can encodes foreseeable changes into generic actions, one has a framework that can revise the model to reflect the changes and to answer queries based on the revised model.

However, as in any model-based reasoning, elicitation and encoding of domain knowledge such as changes into their proper representations remain the most difficult tasks. In the future, we plan to extend our framework to perform mechanism-based learning so that changes can be automatically detected and extracted from data.

In this paper, we have not demonstrated how the framework can support planning and explanation. Given decision objectives for a modeled system, the framework is capable of automatically planning a sequence of actions that is most likely to yield desired outcomes. In (Lu & Druzdzel 2002), we framed the problem as search for opportunities and showed a myopic search algorithm for finding the best next action to perform based on the value of intervention. We plan to extend this work to some practical domains. With regards to supporting explanation, we believe that the framework can be further extended to automatically recover likely causal models for a given history of events.

Acknowledgments

Our research was supported by the Air Force Office of Scientific Research under grant F49620–03–1–0187. The methods described in this paper are implemented in GeNIe and SMILE, available at `http://www2.sis.pitt.edu/~genie`

References

Druzdzel, M. J., and Simon, H. A. 1993. Causality in Bayesian belief networks. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI–93)*, 3–11. San Francisco, CA: Morgan Kaufmann Publishers.

Druzdzel, M. J., and van Leijen, H. 2001. Causal reversibility in Bayesian networks. *Journal of Experimental and Theoretical Artificial Intelligence* 13(1):45–62.

Fikes, R. E., and Nilsson, N. J. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2(3-4):189–208.

Lu, T.-C., and Druzdzel, M. J. 2002. Causal models, value of intervention, and search for opportunities. In Gamez, J. A., and Salmeron, A., eds., *Proceeding of the First European Workshop on Probabilistic Graphical Models (PGM'02)*, 108–116.

Lu, T.-C. 2003. *Construction and Utilization of Mechanism-based Causal Models*. Ph.D. Dissertation, University of Pittsburgh.

Nayak, P. P. 1994. Causal approximations. *Artificial Intelligence* 70(1–2):1–58.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

Simon, H. A. 1953. Causal ordering and identifiability. In Hood, W. C., and Koopmans, T. C., eds., *Studies in Econometric Method. Cowles Commission for Research in Economics. Monograph No. 14*. New York, NY: John Wiley & Sons, Inc. chapter III, 49–74.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. Cambridge, MA: The MIT Press, second edition.

Strotz, R. H., and Wold, H. 1960. Recursive vs. nonrecursive systems: An attempt at synthesis; Part I of a triptych on causal chain systems. *Econometrica* 28(2):417–427.

Wold, H., and Jureen, L. 1953. *Demand Analysis. A Study in Econometrics*. New York: JohnWiley and Sons, Inc.