Linköping Electronic Articles in
Computer and Information Science
Vol. 3(1999): nr 2

Useful Counterfactuals

Tom Costello
John McCarthy

Computer Science Department
Stanford University
Stanford, CA 94305

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis/1999/002/

Published on July 1, 1999 by
Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

Typeset by the authors using LaTeX
Formatted using étendu *style*

Recommended citation:
<Authors>. <Title>. Linköping Electronic Articles in Computer and Information Science, Vol. 3(1999): nr 2. http://www.ep.liu.se/ea/cis/1999/002/. *July 1, 1999.*

This URL will also contain a link to the authors' home pages.

Abstract

Counterfactual conditional sentences can be useful in artificial intelligence as they are in human affairs. In particular, they allow reasoners to learn from experiences that they did not quite have.

Our tools for making inferences from counterfactuals permit inferring sentences that are not themselves counterfactual. This is what makes them useful.

A simple class of useful counterfactuals involves a change of one component of a point in a space provided with a cartesian product structure. We call these cartesian counterfactuals. Cartesian counterfactuals can be modeled by assignment and contents functions as in program semantics. We also consider the more general tree-structured counterfactuals.

1 Introduction

> (1): "If another car had come over the hill when you passed that car, there would have been a head-on collision."

is a useful counterfactual conditional sentence. The driver will estimate how long passing took and how much time he would have had if another car had come over the hill. In this he uses his observation of the distance to the top of the hill when he got back in the right hand lane and his estimate of how long it would have taken a car suddenly coming over the hill to cover that distance.

If he regards the counterfactual (1) as true, he will be more cautious about passing in the future. If he regards it as false, it will be because he thinks he had plenty of time to complete the maneuver and get back in the right lane. If he regards it merely as a material implication with a false antecedent, he'll believe it but won't take it as a reason to change his driving habits.

We discuss such useful counterfactuals for two reasons. (1) We expect them to be useful in AI systems. (2) We have found counterfactuals inferred from experience and having behavioral consequences to admit a richer theory than counterfactuals cooked up to serve as examples of the semantics.

In general, a *counterfactual conditional sentence* (*counterfactual* for short) is a sentence

$$p \text{ implies } q,$$

in which the antecedent p is false. If *implies* is taken to be the ordinary mathematical logical implication $\supset$, then all counterfactuals are true. However, in ordinary life, it is often useful to consider certain counterfactuals, as in the above example (1), as being possibly false. Also (1) may be a useful counterfactual because it permits the driver to learn from an experience he didn't quite have.

We will use $\succ$ to represent counterfactual implication, so that (1) may be written

> Another car was coming when you passed $\succ$ there was a head-on collision.

Q. What can the driver learn if he believes (1)?

A. Whatever he could learn from "Another car came when you passed, and there was a head-on collision."

Three points.

1. We ignore the possibility that the collision might interfere with the driver's ability to learn.

2. What can be learned depends on the driver's mechanisms for learning. However, we propose that he applies the same learning

mechanism to the almost experience as he would apply to the real experience.

3. The hypothetical experience is not a full possible world, e.g. the name of the driver of the car that came over the hill isn't determined. It is a theoretical entity in an *approximate theory.* [McCarthy, 2000] treats approximate theories in some detail, but the simplest approximate theories for discussing counterfactuals are the theories of *cartesian counterfactuals* of the present article.

One straightforward inference from (1) is (2)

Another car comes when you pass under sufficiently similar circumstances $\supset$ there is a head-on collision.

Thus a counterfactual conditional is used to infer a corresponding ordinary conditional, univerally quantified over "similar circumstances". Believing the counterfactual (1) lets us make the same generalization from the counterfactual conditional that we would make had the collision occurred.

There is some resemblance between this inference and those used to infer a scientific law, but there are two important differences. (1) "similar circumstances" is often not spelled out in language. It can be just circumstances that look the same. (2) The inference is from a single experience; more is needed to infer a scientific law.

While counterfactuals are used in daily life, they have been most studied in philosophy. The main philosophical goal has been to assign meaning to these sentences. The Lewis/Stalnaker [Stalnaker, 1968] [Stalnaker and Thomason, 1970] approach is the leading one. David Lewis [Lewis, 1973] defines the truth of a counterfactual using the notion of possible worlds. He posits that possible worlds are ordered by comparative possibility. He then regards $p \succ q$ as true provided q is true in the nearest possible worlds where p is true. He gives no way of judging the closeness of worlds save that of considering the counterfactuals. Thus it seems difficult to evaluate the truth of any particular counterfactual based on other evidence.

Moreover, he Lewis/Stalnaker approach offers no way of inferring non-counterfactual sentences from counterfactuals.

The truth of a counterfactual does not just depend on the state of the world the way a direct observation does. There is always a theory—in the above example, an implicit theory of driving shared by the passenger and the driver. The theory is based on their experience and what they have learned and been taught about driving.

2 Uses of Counterfactuals

For AI purposes we ask what kinds of counterfactuals are useful to humans, as they are also likely to be the kinds useful to computer programs. Concentrating on counterfactuals that are useful in that believing them usefully affects behavior may also have some philosophical benefits, because the reasoning leading to a useful counterfactual and the useful conclusions drawn from it provide some guidance about the best kind of theory.

We begin with some examples of the uses of particular counterfactuals.

1. The counterfactual head-on collision of the introduction.

2. Two ski instructors see a pupil fall on a hill. One says "If he had bent his knees more, he would not have fallen". The other disagrees and claims "If he had put his weight on his downhill ski, he would not have fallen". This example is from [McCarthy, 1979]. Each one suggests a specific kind of lesson.

3. Our ski instructors could view the world in a different way, and both assent to the counterfactual, "if he had two more lessons he would not have fallen".

4. If Caesar had been in charge in Korea, he would have used nuclear weapons. If Caesar had been in charge in Korea, he would have used catapults. Different theoretical structures give different counterfactuals. Maybe the first suggests that if we are serious about winning, we should be more like Caesar. The second doesn't suggest anything.

5. If there had been one more book in that box you would not have been able to lift it. The lesson is not to put too many books in a box.

6. If wishes were horses, beggars would ride. We are far from proposing a way of drawing conclusions from metaphors. This very abstract counter-factual would have a very approximate theory

Each of these counterfactuals tells us something about how the world works. We can use this advice in future, if we find ourselves in a similar situation. The notion that the counterfactual is applicable on *similar* occasions is important. If we are to use a counterfactual to predict in a new occasion s, there must be some kind of test, whether or not the new occasion s is sufficiently similar to the situation that gave rise to the counterfactual. This test is given by an *approximate theory.*

The kind of approximate theories that are relevant to counterfactuals are those covering only some aspects of the world. The theory used in the car example does not include facts that determine whether or not another car can come over the hill. Likewise the skier theory does not determine if a skier bends his knees. The theories allowing inferences about Caesar in Korea respectively relate to his character as a general and the weapons he had. "If wishes were horses ...

", takes only advantage of the semantic parallel between having a wish and having a material object. Approximate theories are more thoroughly discussed in [McCarthy, 2000].

2.1 Learning

A skier might conclude that had he bent his knees on a certain slope he would not have fallen. The skier can learn from this. It is important, if we are to use counterfactuals for learning that we can recognize that they are sometimes false. In our first example we can imagine the response that "If there were another car it would have been visible in time for me to avoid it". This new counterfactual can also be true or false.

The performance of learning algorithms improves when they have more examples. Counterfactuals are one way to collect more examples than can be found by direct experience. Often it is better to imagine a data point than to experience it.

2.2 Prediction in similar circumstances

In so far as our knowledge of the world is incomplete, new sentences can tell us more about the world. Every counterfactual we are told gives us more information about how the world would be, if things were only slightly different, relative to some unstated *approximate theory*. This information can later be used if we find ourselves in a situation with only a small number of differences between it and the present, so that the *approximate theory* is applicable to both. The counterfactual

> "If there had been one more book in that box you would not have been able to lift it."

tell us that in future situations, that satisfy the unstated theory the speaker considers, boxes with more books in them will be too heavy to lift. This differs from the learning we considered earlier, as this is inferring a universal from a counterfactual, rather than using the counterfactual as an instance in a learning algorithm.

If the approximate theory is unstated, to use this counterfactual we need to infer what theory was used. A natural default to use here is to assume that the speaker is using the same theory that you find appropriate to describe the situation.

We can apply counterfactuals in other situations because the theories on which they are based are approximate. The truth of the counterfactual only depends on certain features of the situation, and when these features re-occur, the same inference may be made. In a later section we give an example of where we can derive new facts, that do not mention counterfactuals, from a counterfactual. In our skiing domain, we show that we can derive a fact about the world (that a certain slope is a turn), from the truth of a counterfactual, ("if he had put his weight on his downhill ski, he would not have fallen").

Counterfactuals are useful for other purposes in AI. Ginsberg [Ginsberg, 1986] suggests that they are useful for planning. They also are closely related to the notion of *causality*, as discussed in [Pearl, 1988] [Geffner, 1992],[Pearl, 2000].

3 Detailed Examples

Cartesian counterfactuals involve a counterfactual sentence, a theory, a frame, and a world/point in the frame. The sentence is interpreted by the theory, relative to the frame. We move through the frame, from the current world, to a new point in the frame.

3.1 Rectangular co-ordinates

Example: 1 We consider a cartesian space of three components, x, y, z. The theory provides a co-ordinate system for the space. We are interested in the distance from a point to the origin:

$$s = \sqrt{x^2 + y^2 + z^2}$$

This sentence is our *approximate* theory.

Now consider the point $(1, 2, 1)$. This will act as our current world. We ask whether,

$$y = 3 \succ s = \sqrt{19}. \tag{2}$$

Our cartesian structure implies that x and z hold their particular values $1, 1$. Therefore we have

$$s = \sqrt{1+9+1}, \quad s = \sqrt{11}, \quad s \neq \sqrt{19}.$$

Therefore (2) is an example of an untrue counterfactual. The counterfactual "if y were 3 then s would be $\sqrt{11}$" is true, i.e. $y = 3 \succ s = \sqrt{11}$.

This example has all the essential elements of a cartesian counterfactual. We have a cartesian space, whose points are the models of the approximate theory and whose structure is given by the co-ordinates. Each of these co-ordinates in this example is the value of a variable. To find the truth value of a counterfactual we change the co-ordinate as specified by the left hand side and evaluate the right hand side.

The co-ordinate frame we chose here, using the variables x, y, z, is not the only possible frame. A transformation of those variables would give a similar theory but with different counterfactuals.. One possibility is,

$$x' \;=\; x + y, \quad y, \quad z,$$

Our initial state has the following values.

$$x' = 3, \quad y = 2, \quad z = 1,$$

Given this co-ordinate system, the counterfactual $y = 3 \succ s = \sqrt{19}$ is true, whereas it was false in the previous frame.

In this example, s was uniquely determined by the values of the co-ordinates. We can imagine that another value r, is not uniquely determined—we only know that it is smaller than s. Thus we have,

$$r \leq \sqrt{x^2 + y^2 + z^2}.$$

Again, let us choose x, y, z to have the values $(1, 2, 1)$. In this case $y = 3 \succ r \leq \sqrt{11}$ is true, as we know that $r \leq s$ and $s = \sqrt{11}$.

Furthermore, we know that $y = 3 \succ r \geq 10$ is false. However, when we cannot uniquely determine all other components in terms of the co-ordinates, some counterfactuals are indeterminate. For instance, we do not know the truth value of $y = 3 \succ r = \sqrt{6}$.

Thus different co-ordinate frame make different counterfactuals true. Thus three factors influence the truth of a counterfactual:

1. The space of possible states.
2. The co-ordinate frame on the space of possible states.
3. The current state.

The above rectangular co-ordinate system example hasn't enough structure to prefer one theory over another. However, suppose it were specified that x, y and z were the co-ordinates along the walls and the height of a point starting from the corner of a room. Then there would be some reason for preferring the x-y-z theory and its associated counterfactuals to the x'-y-z theory and its associated counterfactuals. When there is a clear reason to prefer one theory, its counterfactuals can have a somewhat objective character. These are the most useful. Even so, this would be a useful counterfactual only imbedded in a larger theory that includes some goal.

3.2 The Almost Crash—Elaborated

We elaborate the almost-crash example slightly to say "might have been a head-on collision" instead of "would have been a head-on collision". In this uncertain world, it is more realistic.

Suppose the driver estimates (triggered by the passenger's statement) that it will take 7 ± 2 seconds to complete the pass and get back in the right lane. He also estimates that it will take 9 ± 3 seconds to drop back and get in line and that if he stays in the left lane and a car comes over the hill at that instant, he collision will occur in 9 ± 5 seconds. He concludes that if a car comes over the hill at that embarrassing moment the probability of a collision is 0.2. He estimates the probability of a car coming over the hill at that instant is between 0.001 and 0.0001. He concludes that the odds, while small, are unacceptable.

We have expressed this example numerically, as might be appropriate for a robot. Formalizing this aspect of human behavior might not be so numerical.

We can suppose that his estimate of the required passing time is not an *a priori* estimate based on a theory of driving but is based on how rapidly he was overtaking the other car, i.e. is based on this experience. Thus he learns to change his driving rules from an experience of collision that he didn't quite have. Much case-based reasoning in real life is based on counterfactuals, although the theories described in [Lenz et al., 1998, Leake and Plaza, 1997] do not include counterfactual cases.

3.3 Formalization of the Skiing Examples

We now look at an axiomatization that formalizes some examples of counterfactuals in a skiing domain. We sketch a formalization of skiing, which includes both some facts about lessons, and some facts about what happens when you ski. The formalization uses the situation calculus discussed in [McCarthy and Hayes, 1969].

The analysis is designed to highlight several points.

1. Co-ordinate frames give a simple way of defining semantics for counterfactuals.

2. Different frames are appropriate for different points of view.

3. Different frames can lead to different counterfactuals.

We consider four counterfactuals. The first two are from our previously mentioned two ski instructors. Before we consider the counterfactuals, we consider the following story.

> Junior had 25 dollars. He went to a ski slope and spent his money on one cheap lesson, that cost 25 dollars. Cheap lessons teach you one skill each lesson, either bending your knees on bumps, or if you have learned that, then placing your weight on your downhill ski on turns. Expensive lessons cost twice as much, but teach two skills per lesson, in the same order. Thus, Junior learns to bend his knees. He then goes skiing, at a time picked out by situation 1, but when he comes to Slope4 he falls.
>
> Slope4 is a turn, and according to our theory, unless you place your weight on your downhill ski on turns you fall. We also believe that unless you bend your knees on a bump you fall, and that these behaviors happen only if you have learned these skills.

Our first counterfactual states that "if he had actually bent his knees in situation 1 then he would not have fallen in the next situation". This will be represented by the following counterfactual.

$$\mathrm{Actual}(\mathrm{result}(\mathrm{Bend}, \mathrm{S1})) \succ \neg \mathrm{holds}(\mathrm{Fallen}, \mathrm{Next}(\mathrm{S1})) \tag{3}$$

The other states that "if he had actually put his weight on the downhill ski in situation 1 then he would not have fallen in the next situation". In our logical formalization this will be written.

$$\mathrm{Actual}(\mathrm{result}(\mathrm{Down}, \mathrm{S1})) \succ \neg \mathrm{holds}(\mathrm{Fallen}, \mathrm{Next}(\mathrm{S1})) \tag{4}$$

The difference can be resolved by finding out whether Slope4 has a bump or is a turn. As Slope4 is a turn, the first is false and the second is true. Here, the counterfactuals are evaluated in a frame where what Actually happens is a component. Thus, we can change the fact that he bent his knees, without changing any other co-ordinate, and thus infer that he does not fall.

Both our instructors can agree that "if he had chosen more expensive lessons, then he would not have fallen".

$$\mathrm{Choice} = \mathrm{Expensive} \succ \neg \mathrm{holds}(\mathrm{Fallen}, \mathrm{Next}(\mathrm{S1})) \tag{5}$$

Here they choose to keep the number of lessons fixed, but make the lessons better.

Our hero does not assent to the above conditional. He knows that *"if he chose the expensive lessons, he would not have had any lessons"*, as he cannot afford it. Thus he believes that he would fall if he chose expensive lessons,

$$\mathrm{Choice} = \mathrm{Expensive} \succ \mathrm{holds}(\mathrm{Fallen}, \mathrm{Next}(\mathrm{S1})). \tag{6}$$

Thus he still would not have learned the requisite skills. Expensive lessons cost more, so he would not have been able to take as many lessons.

The reason that the ski instructors differ from our hero on this conditional is that they allow the amount of money to vary, while our hero allows the number of lessons to vary. They use different frames and thus get different results.

We give an axiomatization of part of this domain in Section 5, and now present a set of frames that gives the results for the first two counterfactuals.

3.4 Co-ordinate frames for Skiing

With the axiomatization of the above story given below we can now consider the first two counterfactuals. In each case we take an approximate theory, which will be some subset of the consequences of the axioms in Section 5 , and a co-ordinate frame. From the theory and the frame we judge the truth of the counterfactual. We show that choosing different approximate theories, or choosing different frames can lead to different choices.

We give the co-ordinate frame that the two instructors use. They choose what situations were actual, and what fluents held at the situation S1, and the type of Slope4 as their frame. Here *tr* is a truth value. The co-ordinates in the frame are the following terms, (two of these are propositions, and so take on the values true or false).

$$\begin{array}{l}\mathrm{Next(S1)}, \quad \mathrm{Typeslope4}, \\ \mathrm{holds(Fallen, S1)}, \quad \mathrm{holds(Skiing4, S1)}.\end{array}$$

As their core *approximate* theory they take the axioms about the effects of various ski moves, 15, the axiom relating Next and Actual, 16, and the unique names and domain closure axioms and the frame axioms 18. As the current world, they choose the values,

$$\begin{array}{ll}\mathrm{Next(S1)} = \mathrm{result(Falls, S1)} & \mathrm{Typeslope4} = \mathrm{Turn} \\ \neg\mathrm{holds(Fallen, S1)} & \mathrm{holds(Skiing4, S1)}\end{array}$$

as they agree that Junior did fall on a slope with a bump, while skiing Slope4.

With this frame the following counterfactuals are false and true, respectively.

$$\begin{array}{l}\mathrm{Actual(result(Bend, S1))} \succ \neg\mathrm{holds(Fallen, Next(S1))} \\ \mathrm{Actual(result(Down, S1))} \succ \neg\mathrm{holds(Fallen, Next(S1))}\end{array} \tag{7}$$

In contrast the counterfactual

$$\begin{array}{l}\mathrm{Actual(result(Bend, S1))} \wedge \mathrm{Typeslope4} = \mathrm{Bump} \succ \\ \quad \neg\mathrm{holds(Fallen, Next(S1))}\end{array} \tag{8}$$

is true. We prove the first two claims formally after introducing the necessary machinery in the next section.

4 Cartesian Counterfactuals via State Vectors

In the previous section we claimed that certain counterfactuals were true, given some co-ordinate frames. We now give a preliminary axiomatization in terms of state vectors that allows us to formally prove these statements.

We can define cartesian counterfactuals in terms of state vectors [McCarthy, 1962]. The value of a variable x in a state vector ξ is $c(x, \xi)$, while the state vector that is like ξ, save that x has been assigned the value v, is $a(x, v, \xi)$. We can axiomatize a and c as follows.

$$\begin{array}{l}\forall x, v.\xi.c(x, a(x, v, \xi)) = v, \\ \forall x, y, v, \xi.x \neq y \rightarrow c(y, a(x, v, \xi)) = c(y, \xi)\end{array} \tag{9}$$

The numerical example of subsection 3.1 is expressed as follows. Let ξ_0 represent the actual state of the world. We have

$$\begin{aligned} c(x, \xi_0) &= 1 \\ c(y, \xi_0) &= 2 \\ c(z, \xi_0) &= 1 \end{aligned} \tag{10}$$

We are interested in the function

$$s(\xi) = \sqrt{c(x, \xi)^2 + c(y, \xi)^2 + c(z, \xi)^2}. \tag{11}$$

The counterfactual

$$y = 3 \succ s = 7 \tag{12}$$

takes the form

$$s(a(y, 3, \xi_0)) = 7. \tag{13}$$

It is obviously false.

Notice that while cartesian counterfactuals give a meaning to "if x were 7", they do not give a meaning to "if $x + y$ were 7. This is a feature, not a bug, because in ordinary common sense, counterfactuals easily constructed from meaningful counterfactuals are often without meaning.

In the above the "variables" x, y, and z are logically constants. When we need to quantify over such variables, we need new variables with an appropriate typographical distinction.

4.1 Relating State vectors to Propositions

The above way of rephrasing a counterfactual as a statement about state-vectors is insufficient in general. It does not specify how each step is to be done in other cases. There are three steps.

1. The first is to describe the current world as the contents of a state vector. This is equation 10 above.

2. We now relativize all the other terms, those that are not co-ordinates, in the theory, making them functions of a state vector ξ. We replace the terms t that are co-ordinates by their the contents functions applied to the term, i.e. $c(t, xi)$. This gives the equation 11 above.

3. The third and final step is to interpret left hand side as a modification of the state vector. This step yields equation 13 in the above example. We then judge the relativized right hand side applied to the new state vector. That is we ask if the relativized right hand side follows from the values of the co-ordinates initially, and the relativized theory.

Our state vector does not have values for all terms, just those in our co-ordinate frame. Thus, for our numerical example, the set of co-ordinates is,

$$\mathbf{x}, \mathbf{y}, \mathbf{z}$$

We need to distinguish between the value of a constant x and its name $\mathbf{x}$. Here $\mathbf{x}$ in bold font refers to the name x, while x in plain font denotes the value of x.

Let the state vector ξ assign these the terms[1] $\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_z$. The propositions that the state vector ξ encodes are thus the three propositions,

$$x = n_x, \quad y = n_y, \quad z = n_z$$

Here we have used x, in plain font, as the proposition states that the value of $\mathbf{x}$ is the value of the term $\mathbf{n}_x$.

We assign a co-ordinate a value using the a assignment function. We can determine the value of a co-ordinate in a state vector using c, the contents function.

We call the set of co-ordinates our frame, and we assume that they are exactly the names of terms that are true of the predicate f. Rather than assume that there is only a single frame, we can make the frame an argument of c and a. This gives the axioms,

$$\begin{array}{l} \forall x, v, \xi, f. f(x) \rightarrow c(x, a(x, v, \xi, f), f) = v \\ \forall x, y, v, \xi, f. f(x) \wedge f(y) \wedge x \neq y \rightarrow \\ \quad c(y, a(x, v, \xi, f), f) = c(y, \xi, f) \end{array}$$

However, for notational convenience we assume that the frame is recoverable from the state vector, so we keep the ternary a and the binary c whenever what frame we are using is clear from the context.

Our state vector thus uniquely determines the values of the co-ordinates in the frame. From these values, we can determine a theory, the set of sentences implied by the statements that the co-ordinates have the values in the state vector.

The propositions that our state vector gives us may not be all the facts about the world. In our numerical example, the extra fact,

$$s = \sqrt{x^2 + y^2 + z^2}$$

was also the case.

We usually insist that our approximate theory is consistent with our frame. That is, the propositions true at every point in our frame are consistent with the approximate theory.

[1] We alternately suppose that the state vector assigns values to the names of constants, but for reasons of generality is is sometimes useful to be able to specify that a constant is equal to a term, rather than to a value.

To reiterate, a co-ordinate frame, f, is a tuple of names of terms. In the simplest case these are constants. A point p in a co-ordinate frame f is a tuple of terms, equal in number to the co-ordinates of f, and having the type/sort of the corresponding term in f. In the simplest case, this is a tuple of values, the denotation of the terms. The state vector that assigns the co-ordinates f the terms p is denoted $\xi(f, p)$.

The relativization of a theory A to a state vector variable ξ, written $A(\xi)$ and a set of co-ordinates X is the theory resulting from replacing each co-ordinate x by $c(\mathbf{x}, \xi)$, and replacing each function, predicate and constant k not in X by a function from ξ to the type/sort of k.

Definition: 1 *Let $\mathbf{f}_i$ be a term in a co-ordinate frame f, and let n be a term, and let $p = p_1, \ldots, p_n$ be a point in f, the current world, and let A be an approximate theory.*

A counterfactual sentence $\mathbf{f}_i = \mathbf{n} \succ \psi$, is true in $\langle A, f, p\rangle$, if and only if,

$$\bigwedge_{j \leq m} c(\mathbf{f}_j, \xi_0) = p_j, A(\xi_0) \models \psi(a(\mathbf{f}_i, \mathbf{n}, \xi_0)).$$

We now consider our first example. We define our frame as follows.

$$\begin{aligned} &\forall var. f_0(var) \equiv (var = \mathbf{x}) \vee (var = \mathbf{y}) \\ &\quad \vee (var = \mathbf{z}) \end{aligned}$$

It is important to note that $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are names of constants, not variables themselves.

We now define our initial state vector ξ_0.

$$\begin{aligned} &c(\mathbf{x}, \xi_0, f_0) = 1,\ c(\mathbf{y}, \xi_0, f_0) = 2, \\ &c(\mathbf{z}, \xi_0, f_0) = 1 \end{aligned}$$

We need unique names axioms for the names of terms. We use natural numbers as their own names, for simplicity.

We then have that

$$c(\mathbf{x},\ a(\mathbf{y}, 3, \xi_0, f_0), f_0) = 1.$$

We can derive this from the second property of a and c, and the fact that $\mathbf{x} \neq \mathbf{y}$.

We have $x = 1, y = 3, z = 1$ $s = \sqrt{x^2 + y^2 + z^2} \models s = \sqrt{11}$.

We have this set of sentences as our theory, as these are the formulas that result from the terms x, y etc. to their values, 1, etc. in the state vector.

We now prove that the first two counterfactuals that we considered in our skiing story are true and false respectively. To show this we need to introduce an axiomatization of skiing, which we do in the nest section. The axiomatization is only needed for proving the later theorems, and may be skipped by the un-interested reader.

5 Axiomatization

This section presents a formalization of a skiing domain in the situation calculus. The major novelty of this domain is that it allows some reasoning about the actions of an agent. Normally, in the situation calculus all sequences of actions are considered. Here we look at what sequences would happen given some facts about our agent's beliefs and some rules about how he acts.

We include the notion of an actual time line picked out of the possible time lines in the situation calculus. This follows Reiter and Pinto [Pinto and Reiter, 1993]. We enrich this with a Next partial function, that picks out the next actual situation.

We have five sorts, the usual situation, fluent and action sorts, plus sorts for terrain types and skills.

We have five actions Skislope4, Falls and Bend, Down, and Wait. We have fluents, Fallen, Skiing4, Learned(sk) for every skill sk. Our two skills are Bendknees and Downhillski.

We have that after his one lesson our hero knows how to bend his knees.

$$\begin{array}{l}\mathrm{holds}(\mathrm{Learned}(\mathrm{Bendknees}), \mathrm{S0})\\ \neg\mathrm{holds}(\mathrm{Learned}(\mathrm{Downhillski}), \mathrm{S0})\\ \neg\mathrm{holds}(\mathrm{Fallen}, \mathrm{S0})\\ \neg\mathrm{holds}(\mathrm{Skiing4}, \mathrm{S0})\end{array} \tag{14}$$

We now have some facts about what the actual time line is, given the state of the world, and the knowledge of our hero. If our hero sees a bump, and has learned to bend his knees, he will actually do so, otherwise he will actually fall. If he sees a turn, then he will put his weight on his downhill ski if he has learned to do that, otherwise he will actually fall. We also add that if he does the wrong thing, then he will fall, and add the results of falling etc.

$$\begin{array}{l}\forall s.\mathrm{Typeslope4} = \mathrm{Bump}\\ \wedge\mathrm{holds}(\mathrm{Learned}(\mathrm{Bendknees}), s) \rightarrow\\ \quad \mathrm{Actual}(\mathrm{result}(\mathrm{Bend}, \mathrm{result}(\mathrm{Skislope4}, s)))\end{array} \tag{15}$$

$$\begin{array}{l}\forall s.\mathrm{Typeslope4} = \mathrm{Bump}\\ \wedge\neg\mathrm{holds}(\mathrm{Learned}(\mathrm{Bendknees}), s) \rightarrow\\ \quad \mathrm{Actual}(\mathrm{result}(\mathrm{Falls}, \mathrm{result}(\mathrm{Skislope4}, s)))\end{array}$$

$$\begin{array}{l}\forall s.\mathrm{Typeslope4} = \mathrm{Turn}\\ \wedge\mathrm{holds}(\mathrm{Learned}(\mathrm{Downhillski}), s) \rightarrow\\ \quad \mathrm{Actual}(\mathrm{result}(\mathrm{Down}, \mathrm{result}(\mathrm{Skislope4}, s)))\end{array}$$

$$\begin{array}{l}\forall s.\mathrm{Typeslope4} = \mathrm{Turn}\wedge\\ \neg\mathrm{holds}(\mathrm{Learned}(\mathrm{Downhillski}), s) \rightarrow\\ \quad \mathrm{Actual}(\mathrm{result}(\mathrm{Falls}, \mathrm{result}(\mathrm{Skislope4}, s)))\end{array}$$

$$\forall s.\text{Typeslope4} = \text{Turn} \wedge \text{holds}(\text{Skiing4}, s) \rightarrow \\ \text{holds}(\text{Fallen}, \text{result}(\text{Bend}, s))$$

$$\forall s.\text{Typeslope4} = \text{Bump} \wedge \text{holds}(\text{Skiing4}, s) \rightarrow \\ \text{holds}(\text{Fallen}, \text{result}(\text{Down}, s))$$

$$\begin{aligned}
&\forall s.\text{holds}(\text{Fallen}, \text{result}(\text{Falls}, s))\\
&S1 = \text{result}(\text{Skislope4}, \text{S0})\\
&\text{Actual}(\text{S0})\\
&\text{Actual}(\text{result}(\text{Skislope4}, \text{S0}))\\
&\forall as.s \neq \text{S0} \wedge s \neq S1 \wedge \neg\exists a.s = r(a, S1) \rightarrow\\
&\text{Actual}(s) \equiv \text{Actual}(\text{result}(\text{Wait}, s))
\end{aligned}$$

These axioms are enough to pick out an actual timeline, of skiing Slope4, then carrying out an action that depends on the terrain of Slope4 and Junior's knowledge, and then waiting for ever.

We now state that after every actual situation there is a unique actual situation, that picked out by Next. We take this to be the definition of a partial function Next, so that sentences containing Next are notations for sentences containing Actual.

$$\forall s.\text{result}(a, s) = \text{Next}(s) \overset{def}{\equiv} \text{Actual}(\text{result}(a, s)) \tag{16}$$

We also add that Slope4 is a turn.

$$\text{Typeslope4} = \text{Turn} \tag{17}$$

We also need to add unique names, domain closure, and frame axioms.

We have as our unique names axioms, fully unique names, except for the function Actual, the situation constant S1 and the constant Typeslope4.

We now add our frame axioms for each fluent.

$$\forall a, s.\text{holds}(\text{Skiing4}, \text{result}(a, s)) \equiv a = \text{Skislope4}$$

$$\begin{aligned}
&\forall s, a.a \neq \text{Falls} \wedge (a = \text{Bend} \rightarrow\\
&\quad \text{Typeslope4} \neq \text{Turn} \vee \neg\text{holds}(\text{Skiing4}, s))\wedge\\
&(a = \text{Down} \rightarrow \text{Typeslope4} \neq \text{Bump} \vee\\
&\quad \neg\text{holds}(\text{Skiing4}, s))\\
&\quad \rightarrow (\text{holds}(\text{Fallen}, s) \equiv \text{holds}(\text{Fallen}, \text{result}(a, s)))
\end{aligned} \tag{18}$$

$$\forall s, f.f \neq \text{Skiing4} \rightarrow \\ \text{holds}(f, s) \equiv \text{holds}(f, \text{result}(\text{Wait}, s))$$

$$\forall a, s, sk.\text{holds}(\text{Learned}(sk), s) \equiv \\ \text{holds}(\text{Learned}(sk), \text{result}(a, s))$$

These are enough frame axioms to generate explanation closure axioms for every fluent.

6 Deriving Counterfactuals

Theorem: 1 *The counterfactual*

$$\mathrm{Actual}(\mathrm{result}(\mathrm{Bend},\mathrm{S1})) \succ \mathrm{holds}(\mathrm{Fallen},\mathrm{Next}(\mathrm{S1}))$$

is true, in the theory A axiomatized by the axioms about the effects of various ski moves, 15, the axiom relating Next *and* Actual*, 16, and the unique names axioms and the frame axioms 18, with the following frame,*

$$\begin{array}{l} \mathbf{Next}(\mathbf{S1}), \quad \mathbf{Typeslope4}, \\ \mathbf{holds}(\mathbf{Skiing4},\mathbf{S1}), \mathbf{holds}(\mathbf{Fallen},\mathbf{S1}), \end{array}$$

and the following current world,

$$\begin{array}{ll} \mathrm{Next}(\mathrm{S1}) = \mathrm{result}(\mathrm{Falls},\mathrm{S1}) & \mathrm{Typeslope4} = \mathrm{Turn} \\ \neg\mathrm{holds}(\mathrm{Fallen},\mathrm{S1}) & \mathrm{holds}(\mathrm{Skiing4},\mathrm{S1}) \end{array}$$

Proof: We use the first property of our assignment functions to derive that, in our new world, Actual(result(Bend, S1)).

Actual(result(Bend, S1)) is a shorthand for the result of equating a term in the frame, **Next**(**S1**), to the term **result**(**Bend**, **S1**).

We use the second property of our assignment function a to derive that holds(Skiing4, S1). This follows as **holds**(**Skiing4**, **S1**) is in the frame, and holds(Skiing4, S1) is true in our current world, so we have that its value persists. We also show that Typeslope4 = Turn in the same way.

We now use the effect axiom to derive that

$$\mathrm{holds}(\mathrm{Fallen},\mathrm{result}(\mathrm{Bend},\mathrm{S1})).$$

Finally the definition of Next, which is in the core theory gives us

$$\mathrm{holds}(\mathrm{Fallen},\mathrm{Next}(\mathrm{S1})).$$

∎

Theorem: 2 *The counterfactual*

$$\mathrm{Actual}(\mathrm{result}(\mathrm{Down},\mathrm{S1})) \succ \neg\mathrm{holds}(\mathrm{Fallen},\mathrm{Next}(\mathrm{S1}))$$

is true, in the theory A axiomatized above, with the frame and current world specified above.

Proof: We use the first property of our assignment functions to derive that in our new world, Actual(result(Down, S1)) as above.

We use the second property of our assignment functions to derive that holds(Skiing4, S1) and Typeslope4 = Turn as we did previously.

We now use the frame axiom 18 to derive that

$$\neg\mathrm{holds}(\mathrm{Fallen},\mathrm{result}(\mathrm{Down},\mathrm{S1})),$$

using the given fact that holds(Skiing4, S1) and Typeslope4 = Turn. Finally the definition of Next gives us

$$\neg\mathrm{holds}(\mathrm{Fallen},\mathrm{Next}(\mathrm{S1})).$$

∎

7 Deriving Facts from Counterfactuals

In this section we show that we can derive facts that do not contain counterfactual implications from counterfactuals.

We take as our theory, the axioms in Section 5, save for the last axiom Typeslope4 = Turn. From the counterfactual,

$$\mathrm{Actual}(\mathrm{result}(\mathrm{Down}, \mathrm{S1})) \succ \neg \mathrm{holds}(\mathrm{Fallen}, \mathrm{Next}(\mathrm{S1}))$$

we derive that, Typeslope4 = Turn.

Theorem: 3 *Let the theory A be axiomatized by the axioms in Section 5 save 17 and the counterfactual*

$$\mathrm{Actual}(\mathrm{result}(\mathrm{Down}, \mathrm{S1})) \succ \neg \mathrm{holds}(\mathrm{Fallen}, \mathrm{Next}(\mathrm{S1})),$$

judged relative to the approximate theory B axiomatized by the axioms about the effects of various ski moves, 15, the axiom relating Next *and* Actual *16, and the unique names axioms and the frame axioms 18. Let $f(var)$ be defined as follows:*

$$\forall var. f(var) \equiv \begin{array}{l} var = \mathbf{Next}(\mathbf{S1}) \vee \\ var = \mathbf{holds}(\mathbf{Fallen}, \mathbf{S1}) \\ var = \mathbf{holds}(\mathbf{Skiing4}, \mathbf{S1}) \end{array}$$

Then, $A \models \mathrm{Typeslope4} = \mathrm{Turn}$.

Proof: We first note that in A, the co-ordinates have the values, Falls, t, t. We expand the definition of a counterfactual being true, namely,

$$\begin{array}{l} c(\mathbf{Next}(\mathbf{S1}), \xi_0) = \mathrm{result}(\mathrm{Falls}, \mathrm{S1}), \\ c(\mathbf{holds}(\mathbf{Fallen}, \mathbf{S1}), \xi_0) = t, \\ c(\mathbf{holds}(\mathbf{Skiing4}, \mathbf{S1}), \xi_0) = t, \\ B(\xi_0) \models \\ \neg \mathrm{holds}(\mathrm{Fallen}, c(\mathbf{Next}(\mathbf{S1}), a(\mathbf{Next}(\mathbf{S1}), \mathrm{result}(\mathrm{Down}, \mathrm{S1}), \xi_0))). \end{array}$$

to get, by replacing the function c everywhere its value at that argument,

$$\begin{array}{l} \mathrm{Next}(\mathrm{S1}) = \mathrm{result}(\mathrm{Down}, \mathrm{S1}), \neg \mathrm{holds}(\mathrm{Fallen}, \mathrm{S1}), \\ \mathrm{holds}(\mathrm{Skiing4}, \mathrm{S1}), B \models \\ \quad \neg \mathrm{holds}(\mathrm{Fallen}, \mathrm{result}(\mathrm{Down}, \mathrm{S1})) \end{array}$$

This is a meta-theoretic statement, so we translate this into second order logic, by quantifying over predicates, functions and objects in the standard way[2].

$$\begin{array}{l} \forall \bar{Z}. \mathrm{Next}'(\mathrm{S1}') = \mathrm{result}'(\mathrm{Down}', \mathrm{S1}') \wedge \\ \quad \neg \mathrm{holds}'(\mathrm{Fallen}', \mathrm{S1}') \wedge \\ \mathrm{holds}'(\mathrm{Skiing4}', \mathrm{S1}') \wedge B[\bar{Z}/\bar{Z}'] \rightarrow \\ \neg \mathrm{holds}'(\mathrm{Fallen}', \mathrm{Next}'(\mathrm{S1}')), \end{array}$$

[2] We do not need to quantify over the domain, as in this case we have enough axioms to ensure that the domains of each of our sorts are fixed.

where $\bar{Z}$ is the tuple of all constant symbols in the language, and $\bar{Z}'$ is a tuple of variables, equal in type and arity to the constants in $\bar{Z}$, we write variables in Z' as Next' etc.

This second order sentence captures the truth of the counterfactual. We can conjoin this with the other axioms in the theory A, and ask whether it implies, $\mathrm{Typeslope4} = \mathrm{Turn}$ in second order logic.

It does, as can be seen from instantiating the universally quantified variables, with their corresponding constants, except for Actual' and Next' which are instantiated by the predicate P true of the smallest set of situations satisfying the formulas below, and the partial function, $\mathrm{Next}''(s) = \mathrm{result}(a, s) \equiv P(\mathrm{result}(a, s))$, defined from P,

$$
\begin{aligned}
&P(\mathrm{S0}) \wedge P(\mathrm{result}(\mathrm{Skislope4}, \mathrm{S0}))\wedge \\
&P(\mathrm{result}(\mathrm{Down}, \mathrm{result}(\mathrm{Skislope4}, \mathrm{S0}))) \\
&\wedge \forall s. s \neq \mathrm{S0} \wedge s \neq \mathrm{result}(\mathrm{Skislope4}, \mathrm{S0})\wedge \\
&\quad s \neq (\mathrm{result}(\mathrm{Down}, \mathrm{result}(\mathrm{Skislope4}, \mathrm{S0})) \wedge P(s) \rightarrow \\
&P(\mathrm{result}(\mathrm{Wait}, s))
\end{aligned}
$$

The result of instantiating these terms, and noting that the left hand side is derivable from the theory A gives the sentence.

$$\neg \mathrm{holds}(\mathrm{Fallen}, \mathrm{result}(\mathrm{Down}, \mathrm{S1}))$$

However, we also have $\mathrm{holds}(\mathrm{Skiing4}, s1)$, and the axiom,

$$
\begin{aligned}
&\forall s. \mathrm{Typeslope4} = \mathrm{Bump} \wedge \mathrm{holds}(\mathrm{Skiing4}, s) \rightarrow \\
&\mathrm{holds}(\mathrm{Fallen}, \mathrm{result}(\mathrm{Down}, s))
\end{aligned}
$$

Instantiating this with S1 and simplifying gives $\mathrm{Typeslope4} \neq \mathrm{Bump}$. As we know by domain closure that

$$\mathrm{Typeslope4} = \mathrm{Bump} \vee \mathrm{Typeslope4} = \mathrm{Turn},$$

we have

$$\mathrm{Typeslope4} = \mathrm{Turn},$$

as required. ∎

8 Bayesian Networks

This approach can be seen to be similar to modeling systems with structured equations [Simon, 1953] [Druzdel and Simon, 1994], or Bayesian networks [Balke and Pearl, 1994b] [Balke and Pearl, 1994a] [Pearl, 1995]. Rather than have equations that give the value of a variable, we have arbitrary propositions. The dependency relationships are captured by the frame, rather than links. In our model, exogenous variables are in the frame, as are the functions that give the value of the other variables. Updating the Bayes net can then be seen to be updating the frame.

One major difference between our approach and structural equation models or Bayesian networks is that we consider arbitrary propositions, and consider these relative to a background approximate theory. This approximate theory can be rich, that is, not completely describable. The other major difference is that Bayesian networks focus on the probability distribution of certain variables, rather than on facts in general.

We now briefly sketch Galles's and Pearl's formalization of causal models, their generalization of structural equations.

Definition: 2 *A causal model is a triple $M = \langle U, V, F\rangle$, where U is a set of variables, called exogenous variables, V is a set of variables, disjoint from U, called endogenous variables, and F is a set of functions f_i, from $U \cup (V/v_i)$ to v_i for each v_i in V. We shall sometimes represent a function f_i by a defining equation for v_i.*

If X is a tuple of variables in V, and x a tuple of values for the elements of X, we write $X = x$, for the set of equations that result in equating the elements of x with the corresponding value in x.

A sub-model M_X of M is the causal model,

$$M_x = \langle U, V, F_X\rangle$$

where

$$F_X = \{f_i : V_i \notin X\} \cup \{X = x\}$$

If Y is a variable in V, the potential response of Y to equating X to x is the solution for Y of the equations F_X.

The counterfactual $X = x \succ Y = y$ is true if y is the potential response of Y to equating X to x.

We now show how to represent causal model as a cartesian frame.

Definition: 3 *Let $M = \langle U, V, F\rangle$, be a causal model where $U = \{u_1, \ldots, u_m\}$, $V = \langle v_1, \ldots, v_n\rangle$, and*

$$F = \{v_i = f_i(u_1, \ldots, u_m, v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n) | i \leq n\}.$$

The cartesian frame for M is in the equational language, with variables, $u_i | i \leq m$ and functions, v_i of $n + m - 1$ arguments, and functions.

The cartesian frame M^F is the set of function terms v_i, and the current world M^W is the tuple,

$$\begin{array}{l} f_1(u_1, \ldots, u_m, v_2, \ldots, v_n), \\ \ldots, f_i(u_1, \ldots, u_m, v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n), \\ \ldots f_1(u_1, \ldots, u_m, v_1, \ldots, v_{n-1}). \end{array}$$

We now show that $X = x \succ Y = y$ is true in a causal model M if and only if $X = x \succ Y = y$ is true in the cartesian frame M^F, with current world M^W.

Theorem: 4 *$X = x \succ Y = y$ is true in a causal model M if and only if $\bigwedge X = x \succ Y = y$ is true in the cartesian frame M^F, with current world M^W.*

Proof: We prove this by showing that the equations in sub-model M_X are exactly the formulas encoded by the result of updating the state vector $\xi(M^F, M^W)$ by each element of $X = x$.

The sub-model M_X is the causal model,

$$M_x = \langle U, V, F_X \rangle$$

where

$$F_X = \{f_i : v_i \notin X\} \cup \{X = x\}$$

The formulas encoded by $a(v_i, z, \xi(M^F, M^W))$ are

$$\{f_i : v_i \neq z\} \cup \{v_i = z\}.$$

The formulas encoded by the sequence of assignments is

$$\{f_i : v_i \notin X\} \cup \{X = x\}.$$

To show this, it is enough to realize that no v_i can appear twice in X, and thus every distinct update updates a distinct variable.

Thus the equations in sub-model M_X is exactly the formulas encoded. Finally, the conditions for y being a potential response of Y to equating X to x are exactly the conditions for $Y = y$ following from the equations,

$$\{f_i : v_i \notin X\} \cup \{X = x\}.$$

∎

9 An Application of Cartesian Counterfactuals

Cartesian counterfactuals have been used to answer challenge questions in the U.S. DARPA (Defense Advanced Projects Agency) High Performance Knowledge Bases (HPKB) program. In this section briefly explain what this project does, and then we some examples of how cartesian counterfactuals are used.

The HPKB program is developing technology to build large (10 to 100 thousand axioms), reusable knowledge bases that can answer questions on a specific topic. One branch of the project considers Crisis Analysis. That is, it attempts to answer the kinds of questions intelligence analysts would ask when a crisis erupted somewhere in the world.

The HPKB project is competitive. The research groups are dividing into two groups, and build competing systems, which are tested yearly on their ability to answer newly posed questions about a chosen topic. These questions range from the (seemingly) simple, "What is the difference between sarin and anthrax?", to quite involved counterfactuals.

In order to even meaningfully state the counterfactuals, we need to give some background information. Here we give an extract from a scenario, modified by string substitution to be set in the world of the *Lord of the Rings*.

> Mordor supplies a hobbit terrorist group, the Hobbit Liberation Army (HLA), with money, explosives, small arms, and advisors and offers the use of a training facility near Gondor. Mordor arranges to acquire weapons-grade anthrax from a Ranger Mafia cell. Mordor contacts an Elvish biological weapons expert, known only as Feanor, who agrees to supply Mordor with equipment and weapons-engineering skills to complete the weaponization of several small biological devices.
>
> ⋮
>
> The orcs of Mordor, aided by the nefarious Elvish biological weapons expert, successfully build a small, hobbit-portable anthrax sprayer. Separately, arrangements are made with the HLA to begin training exercises in Mordor at the western terrorist training camp.
>
> ⋮

Although entirely hypothetical, the scenario is very detailed, running to 15 pages.

Given this scenario, (called the Y1 Phase II LOTR Scenario) the following counterfactual was posed.

> How would the Y1 Phase II Lord of the Rings Scenario be affected if BW experts of Mordor did not provide advanced technology and scientific expertise aid to a terrorist group?

This question can be stated in KIF (knowledge interchange format) as follows.

```
(and
    (not (occurs-in ?event (scenario-minus
                               ?event1 Y2-scenario)))
    (occurs-in ?event (day ?n  Y2-Scenario ))
    (supply ?event1)
    (object-acted-on ?event1 ?aid)
    (indirect-object-to ?event1 ?group)
    (terrorist-group ?group)
    (expertise ?aid)
    (actor ?event1 ?experts)
```

```
    (citizens-of ?expert Mordor)
    (expert ?experts)
    (information-about ?aid advanced-technology)
    (actor ?event ?actor)
)
```

In the system built by the one team, to which the Formal Reasoning Group is associated, this query is sent to a theorem prover (ATP), which extracts binding from a knowledge base for the variables (the names preceded by a question mark). The binding are then translated into a natural language answer, and the proof into a natural language explanation of the answer. We will not discuss these parts further, as they do not bear on the use of cartesian counterfactuals.

The system find the correct answer, namely Mordor would not have been able to weaponize the anthrax, by reasoning that had they not received the elvish aid, then they would not have been able to weaponize the material, as biological weapon engineering skills are necessary for this task. The reasoning is about 40 resolution steps, and involves 7 rules and 30 ground facts.

In the case of this counterfactual, the frame is the facts true at the beginning of the scenario, and the events that occur before the elvish aid arrives. The events that occur later are not in the frame, as they are not independent of the earlier events.

Counterfactuals are included in the HPKB project as they are a very natural way to explore an evolving situation. When thinking about the present, past cases, and counterfactual variants of them are often employed.

10 Theories admitting counterfactuals

As stated earlier, different theories admit different counterfactuals. Our goal is to make this more definite for common sense theories, but theories involving differential equations (and difference equations) admit particular kinds of counterfactuals in informative ways. We discuss them first and then move on to common sense theories for which counterfactuals are more important.

10.1 Theories given by differential equations

Theories given by differential equations give some clearcut examples.

The solutions are determined by boundary conditions. If the theory includes both the differential equations and the boundary conditions, the most obvious counterfactual is to keep the differential equations and change the boundary conditions to a different set of admissible boundary conditions.

A simple example is provided by the equations of celestial mechanics regarding the planets as point masses. We can use the equations to predict the future of the system from the present but with Mars having a different position and velocity than it has at present. We can also solve the equations supposing a sudden change in the mass of Mars in this theory.

A more complex theory in which Mars has a distribution of mass, which might be necessary to predict the future positions of Deimos and Phobos (the moons of Mars), would not consider a sudden change of the mass of Mars that didn't specify how the new mass was distributed as a definite alternate initial state.

Other counterfactuals are not meaningful in celestial mechanics, e.g. what if Mars had a circular orbit. This isn't meaningful, because even if Mars started along a circle, you would have to make arbitrary assumptions in order to keep it there and might end up violating the law of conservation of momentum.

However, the differential equations don't need to involve time to admit counterfactuals. A theory that gives the distribution of potential in a region as a function of the distribution on the boundary can also consider altered values on the boundary.

The same considerations that apply to differential equations apply to difference equations.

What if the deuteron had a mass one percent larger than it does? For chemistry and the theory of atomic spectra this is a reasonable counterfactual. For example, its effect on spectra and the rate of chemical reactions could be predicted. However, this is a meaningless counterfactual for nuclear physics, because it doesn't say whether the extra mass is in the proton or the neutron or somewhere else. Quantum mechanics doesn't tolerate giving a proton or neutron a different mass in a particular atom, because it violates the notion of identical particles required to apply the anti-symmetry rule for wave functions.

Thus we see that counterfactuals reasonable at one level of theory may be unreasonable at a more fundamental level.

Such counterfactuals in physics are appropriately handled informally — as long as the physicists are people. Robotic common sense reasoning requires a formal treatment.

10.2 Common sense theories

Situation calculus theories are somewhat similar to celestial mechanics in the kinds of counterfactuals they admit. However, different situation calculus formalizations of the same phenomenon may admit different counterfactuals.

Consider two similar theories, the Yale Shooting Problem with the additional statement what walking implies being alive, and the Yale Shooting Problem with the additional rule that shooting causes someone to stop walking (when the gun is loaded). These theories are equivalent when we add induction, as the domain constraint can be derived by an induction.

We now consider what might happen if the gun was a pellet gun.

Pellet guns wounds, though painful are not fatal. Thus, we remove the effect axiom that states shooting kills. In one case, we are left with the effect axiom that shooting stops someone walking, even though it does not kill them—a reasonable conclusion. In the other, we are left with no effects, as the second effect axiom was a consequence of the domain constraint and the axiom that shooting kills.

Thus how we axiomatize our theory can alter the truth of counterfactuals.

11 Non-cartesian counterfactuals

So far we have considered cartesian counterfactuals. In this section we consider how we can go beyond this basic case. The essential restriction that characterizes cartesian counterfactuals is that every point in the product space is a meaningful state of affairs. Thus, if we have a frame with co-ordinates $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$, we can choose any values for $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$, say n_x, n_y, and n_z, and the theory that we get when we add $\mathbf{x} = n_x, \mathbf{y} = n_y, \mathbf{z} = n_z$ to our approximate theory correctly predicts the truth of counterfactuals with premises asserting the values of $\mathbf{x}$, etc.

Sometimes there are assignments of co-ordinates that are not meaningful. This can be because the co-ordinates having those values breaks some rule, or is impossible. In other cases, though those values of the co-ordinates are not excluded by our approximate theory, it may be that our approximate theory does not correctly predict the outcome for those values. Finally, so co-ordinates are not meaningful given when other co-ordinates have certain values. For instance, a choice of whether to bend ones knees or not does not make sense unless you are skiing.

This is made clearer by some examples.

The first example shows a case where some values of the co-ordinates are excluded by the approximate theory. In this theory x is the height of the bottom of a spring, and y the height of the top of the spring, s is then the height of the spring, and f the force outwards on each endpoint. l is the length of the spring at rest, and e the elasticity. In this case, Hooke's Law tells us *"ut tensio sic vis"*, i.e. the force is proportional to the extension of the spring, where e is the proportion, and further, it is in the opposite direction.

Consider the following approximate theory,

$$
\begin{aligned}
&x > y \\
&s = x - y \\
&f = -e * (s - l) \\
&l = 2 \\
&e = 1
\end{aligned}
$$

In the current world $x = 3, y = 5$, so the spring is at its nominal length.

Then, if we have $\mathbf{x}, \mathbf{y}$ as our co-ordinate frame, we have that if $x = 1 \succ s = 4$, further we have that $x = 1 \succ f = -2$, that is, the force on the bottom of the string is 2 units upwards.

What would happen if we moved the bottom of the string to position 6. In the real world, this would be either impossible, (as we would try to push something through itself), or possibly we might end up with a reversed situation. In either case, the approximate theory is no longer the case. We can tell this because the theory makes the following counterfactual true $x = 6 \succ \perp$.

However, sometimes we will go outside the range of meaningfulness of our approximate theory without reaching a contradiction inside it. Consider our spring theory above. Suppose that the spring breaks[3] when it is stretched to length 6.

Our theory tells us that $y = 10 > f = -10$. This is incorrect, but we cannot tell this from our approximate theory alone. In the previous case we could detect a problem because of inconsistency. In this case, we cannot detect the problem in that way.

In both these cases, out theory was cartesian in a certain region, but outside that region, that simple structure broke down, and the cartesian structure no longer provided meaningful answers.

12 Tree Structures of Possibilities

To go beyond cartesian counterfactuals, it is useful to consider trees of possibilities. A node in the tree corresponds to fixing some of the aspects of the entity being considered and regarding the others as variable.

To fix the ideas, let the entity be the world and its history—but there are others.

We consider trees with finite branching and finite depth. We may regard the leaves as possible worlds.

A theory of counterfactuals based on the trees can have more structure than theories based solely on the possible worlds.

Returning to the adventures of Junior, we consider the node A from which the possibilities branch according to whether Junior went surfing, went skiing or stayed home. If Junior went skiing, it makes sense to ask whether he took cheap skiing lessons, expensive skiing lessons or none. These are edges leading from the node leading from the edge in which he went skiing. If he went surfing, there are no edges leading from that node corresponding to the different skiing lessons. [Therefore, we don't have a cartesian product structure, but we could force one by putting edges for the kinds of ski lessons leading from the node in which he went surfing. We don't do that.]

[3] Steel springs will tend to stretch permanently when stretched or deformed beyond a certain point, but other materials, such as glass will break almost immediately outside their elastic region, especially if pulled quickly.

Now let us suppose that Junior had lunch on the given day, regardless of whether he skied or surfed. If he went skiing or surfing, we suppose that at lunch that he may have had either a hamburger or a hot dog or a pizza. If he stayed home he had chicken soup. In fact suppose that he had a hamburger. Now consider the counterfactual, *"If Junior had had a hot dog, he would have had indigestion."* This counterfactual may be stated either about the skiing node or about the surfing node. Let's put in another branch on whether Junior telephoned Deborah or Sheila or neither.

We can imagine the tree of possibilities to have been constructed in two ways. in one the split on what Junior had for lunch precedes in the tree the split on whom he telephoned. We say "precedes in the tree", because we are not concerned with the temporal order of these events, ignoring the fact that if he telephoned Deborah he had to do it before lunch whereas telephoning Sheila would have been done after lunch.

Let the variable x have the value 1, 2 or 3 according to whether Junior went skiing, surfing or stayed home. Let y have one of these values according to whether he had expensive lessons, cheap lessons or none. Let z have values according to what he had for lunch and w have values according to whom he telephoned.

We can label the edges of the tree. The edges from node A are labeled, $x = 1$, $x = 2$, and $x = 3$. We can use a notation for this reminiscent of the notation used for state vectors and write $a(x, 1, A)$ or more explicitly $a(\text{"Junior went skiing"}, A)$. However, in the tree case, we don't have all the nice properties of the state vector case. The edges leading from the node $a(x, 1, A)$ are labelled $y = 1$, $y = 2$ and $y = 3$. However, the edges leading from $a(x, 2, A)$ and $a(x, 3, A)$ have no y-labels nor do any of their successors in the tree, i.e. $a(y, 1, a(x, 2, A))$ is undefined. If Junior didn't go skiing, there is no branch according to what kind of lessons he took.

Here are the trees. Tree 1 puts skiing, surfing and home on the edges leading from A, and Tree 2 puts whom he telephoned on these edges.

The z and w situations are simpler. Provided the value of x is 1 or 2, the tree may be arranged to put the edges labelled with z before or after the edges labelled with w. We can write, for example,

$$\begin{aligned}&(\forall \alpha\beta\gamma)(\gamma \neq 1\\ &\rightarrow a(w, \beta, a(z, \alpha, a(x, \gamma, A))) = a(w, \beta, a(z, \alpha, a(x, \gamma, A))))\end{aligned} \tag{19}$$

This is a *local cartesian product* case. We can say that the variables z and w commute. The relation between z and x is more complex, because telephoning and having hot dogs for lunch arise only when x is 1 or 2. Also when $x = 1$, z and w may jointly commute with y.

We say that a pair of co-ordinates x_1, x_2 is *locally cartesian* under a condition ϕ, when

$$\phi \rightarrow \forall v_1, v2.a(x_1, v_1, a(x_2, v_2, A)) = a(x_1, v_1, a(x_2, v_2, A))$$

This notion make sense for sets of co-ordinates. A set of co-ordinates are *locally cartesian* under a condition ϕ, when all orders of assignment agree.

We can now write

$$\begin{array}{l}(\forall \alpha)(\alpha = 1 \vee \alpha = 2) \rightarrow \\ Eats(\text{HotDog}, a(x, \alpha, A)) \\ \succ Holds(\text{Indigestion}, Next(a(x, \alpha, A)))\end{array} \tag{20}$$

The branching tree of propositional possibilities is isomorphic to the structure of compound conditional expressions discussed in [McCarthy, 1963].

13 Conclusions

We have shown the usefulness of counterfactuals in theories with a cartesian product structure, showing how to infer such counterfactuals and how to make inferences from them. We also consider more general tree-structured counterfactuals.

We do not claim there is a single set of true counterfactuals. Rather, a counterfactual can only be judged relative to a background theory. In the cartesian case, the theory involves a choice of a "co-ordinate frame".

These counterfactuals give information about how the world behaves, so that in future situations the reasoner can better predict what will happen.

To be useful, a counterfactual needs to be imbedded in a theory that includes goals or a notion of utility.

The theories inhabited by counterfactuals are usually approximate theories of the world and sometimes involve concepts and objects that are not fully defined. Approximate theories and approximate objects and their relationships are discussed in [McCarthy, 2000].

Acknowledgments

A preliminary version of this article was accepted for the 1998 NCAI but was withdrawn in disagreement with AAAI's policy of requiring assignment of copyright. This research has been partly supported by ARPA contract no. USC 621915, the ARPA/Rome Laboratory planning initiative under grant (ONR) N00014-94-1-0775 and ARPA/AFOSR under (AFOSR) grant # F49620-97-1-0207.

References

[Balke and Pearl, 1994a] Balke, A. and Pearl, J. (1994a). Counterfactuals and Policy Analysis in Structural Models. In Besnard, P. and Hanks, S., editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. Morgan Kaufmann.

[Balke and Pearl, 1994b] Balke, A. and Pearl, J. (1994b). Counterfactuals Probabilities: Computational Methods, Bounds, and Applications. In de Mantaras, R. L. and Poole, D., editors, *Uncertainty in Artificial Intelligence 10*, pages 46–54. Morgan Kaufmann, San Mateo, California.

[Druzdel and Simon, 1994] Druzdel, M. J. and Simon, H. A. (1994). A structured probabilistic representation of action. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-94)*, pages 154–161.

[Geffner, 1992] Geffner, H. (1992). *Default Reasoning: Causal and Conditional Theories*. ACM Doctoral dissertation award: 1990. MIT Press.

[Ginsberg, 1986] Ginsberg, M. L. (1986). Counterfactuals. *Artificial Intelligence*, 30.

[Leake and Plaza, 1997] Leake, D. and Plaza, E., editors (1997). *Case-Based Reasoning Research and Development, Proceedings of the 2nd International Conference on Case-Based Reasoning (ICCBR-97)*, Lecture Notes in Artificial Intelligence, Berlin. Springer-Verlag.

[Lenz et al., 1998] Lenz, M., Bartsch-Sporl, B., H-D., B., and Wess, S., editors (1998). *Case-Based Reasoning Technology: from foundations to applications*. Number 1400 in Lecture Notes In AI. Springer-Verlag.

[Lewis, 1973] Lewis, D. (1973). *Counterfactuals*. Harvard University Press.

[McCarthy, 1962] McCarthy, J. (1962). Towards a mathematical science of computation. In *Information Processing '62*, pages 21–28. North-Holland. Proceedings of 1962 IFIP Congress.

[McCarthy, 1963] McCarthy, J. (1963). A Basis for a Mathematical Theory of Computation[4]. In Braffort, P. and Hirschberg, D., editors, *Computer Programming and Formal Systems*, pages 33–70. North-Holland, Amsterdam.

[McCarthy, 1979] McCarthy, J. (1979). Ascribing mental qualities to machines[5]. In Ringle, M., editor, *Philosophical Perspectives in Artificial Intelligence*. Harvester Press. Reprinted in [McCarthy, 1990].

[McCarthy, 1990] McCarthy, J. (1990). *Formalizing Common Sense: Papers by John McCarthy*. Ablex Publishing Corporation.

[4]http://www-formal.stanford.edu/jmc/basis.html

[5]http://www-formal.stanford.edu/jmc/ascribing.html

[McCarthy, 2000] McCarthy, J. (2000). Approximate objects and approximate theories[6]. In Cohn, A. G., Giunchiglia, F., and Selman, B., editors, *KR2000: Principles of Knowledge Representation and Reasoning,Proceedings of the Seventh International conference*, pages 519–526. Morgan-Kaufman.

[McCarthy and Hayes, 1969] McCarthy, J. and Hayes, P. J. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence[7]. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press.

[Pearl, 1988] Pearl, J. (1988). Embracing causality in formal reasoning. *Artificial Intelligence*, 35:259–271.

[Pearl, 1995] Pearl, J. (1995). Causation, Action and Counterfactuals. In Gammerman, A., editor, *Computational Learning and Probabilistic Reasoning*, pages 235–255. John Wiley and Sons, New York.

[Pearl, 2000] Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.

[Pinto and Reiter, 1993] Pinto, J. and Reiter, R. (1993). Adding a time line to the situation calculus. In *Second Symposium on Logical Formalizations of Commonsense Reasoning*, pages 172–177.

[Simon, 1953] Simon, H. A. (1953). Causal ordering and identifiability. In *Studies in Econometric Method.*, number 14 in Cowles Commission for Research in Economics., monograph III. John Wiley and Sons, Inc., New York.

[Stalnaker, 1968] Stalnaker, R. (1968). A theory of conditionals. In Rescher, N., editor, *Studies in logical theory*, number 2 in American philosophical quarterly monograph series. Blackwell, Oxford. Also appears in *Ifs*, (ed., by W. Harper, R. C. Stalnaker and G. Pearce), Reidel, Dordrecht, 1981.

[Stalnaker and Thomason, 1970] Stalnaker, R. and Thomason, R. H. (1970). A semantic analysis of conditional logic. *Theoria*, 36:23–42.

[6]http://www.formal.stanford.edu/jmc/approximate.html
[7]http://www-formal.stanford.edu/jmc/mcchay69.html