FORMING BELIEFS ABOUT A CHANGING WORLD*

Fahiem Bacchus
Department of Computer Science
University of Waterloo
Waterloo, Ontario
Canada, N2L 3G1
fbacchus@logos.uwaterloo.ca

Adam J. Grove
NEC Research Institute
4 Independence Way
Princeton, NJ 08540
grove@research.nj.nec.com

Joseph Y. Halpern
IBM Research Division
Almaden Research Center, Dept. K53/802
650 Harry Road
San Jose, CA 95120–6099
halpern@almaden.ibm.com

Daphne Koller
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720
daphne@cs.berkeley.edu

April 25, 1994

Abstract

The situation calculus is a popular technique for reasoning about action and change. However, its restriction to a first-order syntax and pure deductive reasoning makes it unsuitable in many contexts. In particular, we often face uncertainty, due either to lack of knowledge or to some probabilistic aspects of the world. While attempts have been made to address aspects of this problem, most notably using nonmonotonic reasoning formalisms, the general problem of uncertainty in reasoning about action has not been fully dealt with in a logical framework. In this paper we present a theory of action that extends the situation calculus to deal with uncertainty. Our framework is based on applying the *random-worlds* approach of [BGHK94] to a situation calculus ontology, enriched to allow the expression of probabilistic action effects. Our approach is able to solve many of the problems imposed by incomplete and probabilistic knowledge within a unified framework. In particular, we obtain a *default* Markov property for chains of actions, a derivation of conditional independence from irrelevance, and a simple solution to the frame problem.

*This paper appears in *AAAI-94 (Proceedings of the Twelfth National Conference on Artificial Intelligence)*, 1994, pp. 222–229. Some of this research was performed while Daphne Koller was at Stanford University and at the IBM Almaden Research Center. Work supported in part by the Canadian Government through their NSERC and IRIS programs, by the Air Force Office of Scientific Research (AFSC) under Contract F49620-91-C-0080, and by a University of California President's Postdoctoral Fellowship.

1 Introduction

The *situation calculus* is a well-known logical technique for reasoning about action and change [MH69]. Calculi of this sort provide a useful mechanism for dealing with simple temporal phenomena, and serve as a foundation for work in planning. Nevertheless, the many restrictions inherent in the situation calculus have inspired continuing work on extending its scope.

An important source of these restrictions is that the situation calculus is simply a first-order theory. Hence, it is only able to represent "known facts" and can make only valid deductions from those facts. It is unable to represent probabilistic knowledge; it is also ill-suited for reasoning with incomplete information. These restrictions make it impractical in a world where little is definite, yet where intelligent, reasoned decisions must nevertheless be made. There has been much work extending the basic situation calculus using various nonmonotonic theories. Although interesting, these theories address only a certain limited type of uncertainty; in particular, they do not allow us to represent actions whose effects are probabilistic. This latter issue seems to be addressed almost entirely in a non-logical fashion. In particular, we are not aware of any work extending the situation calculus to deal with probabilistic information. This is perhaps understandable: until recently, it was quite common to regard approaches to reasoning based on logic as being irreconcilably distinct from those using probability. Recent work has shown that such pessimism is unjustified. In this paper we use a new theory of probabilistic reasoning called the *random-worlds method* [BGHK94] which naturally extends first-order logic. We show that this method can be successfully applied to temporal reasoning, yielding a natural and powerful extension of the situation calculus.

The outline of this paper is as follows. First, we briefly describe the situation calculus, discussing in more detail some of its problems, and some of the related work addressing these problems. We then describe our own approach. We begin by summarizing the random-worlds method. Although the application of this method to temporal reasoning is not complicated, an appropriate representation of temporal events turns out to be crucial. The solution, based on *counterfactuals*, seems to be central to many disciplines in which time and uncertainty are linked.

After these preliminaries, we turn to some of the results obtained from our approach. As we said, our goal is to go beyond deductive conclusions. Hence, our reasoning procedure assigns *degrees of belief* (probabilities) to the various possible scenarios. We show that the probabilities derived using our approach satisfy certain important desiderata. In particular, we reason correctly with both probabilistic and nondeterministic actions (the distinction between the two being clearly and naturally expressed in our language). Furthermore, we obtain a default Markov property for reasoning about sequences of actions. That is, unless we know otherwise, the outcome of an action at a state is independent of previous states. We note that the Markov property is not an externally imposed assumption, but rather is a naturally derived consequence of the semantics of our approach. Moreover, it can be overridden by information in the knowledge base. The Markov property facilitates a natural mechanism of *temporal projection*, and is a natural generalization of an intuitive mechanism of projection in deterministic domains in which we consider action effects sequentially. In general, when actions have deterministic effects (whether in fact, or only by default, which is another easily made distinction) then our approach achieves most standard desiderata.

Finally, we turn to examining one of the most famous issues that arise when reasoning about action: the *Frame Problem* and the associated *Yale Shooting Problem* (YSP) [MH69, HM87]. We show that our approach can solve the former problem, without suffering from the latter, almost automatically. Writing down a very natural expression of a frame axiom almost immediately gives the desired behavior. We state a theorem, based on the criterion of Kartha [Kar93], showing the general correctness of our approach's solution to the frame problem. We also compare our solution to one given by Baker [Bak91].

2 Preliminaries

2.1 The situation calculus

We assume some familiarity with the situation calculus and associated issues. In brief, by *situation calculus* we refer to a method of reasoning about temporal phenomena using *first-order logic* and a sorted ontology consisting of *actions* and *situations*. A situation is a "snapshot" of the world; its properties are given by predicates called *fluents*. For example, consider a simple version of the well-known Yale Shooting problem. To represent an initial situation S_0 where Fred is alive and there is an unloaded gun we can use the formula $\mathit{Alive}(S_0) \wedge \neg \mathit{Loaded}(S_0)$. The effects actions have on situation can be encoded using a *Result* function. For instance, we can write $\forall s\, (\mathit{Loaded}(s) \Rightarrow \neg \mathit{Alive}(\mathit{Result}(\mathit{Shoot}, s)))$, to assert that if a loaded gun is fired it will kill Fred.[1] We can then ask what would happen if, starting in S_0, we load the gun, wait for a moment, and then shoot: $\mathit{Alive}(\mathit{Result}(\mathit{Shoot}, \mathit{Result}(\mathit{Wait}, \mathit{Result}(\mathit{Load}, S_0))))$?

The most obvious approach for deciding whether this is true is to use first-order deduction. However, for this to work, we must provide many other facts in addition to the two above. In fact, to answer questions using deduction we would in general have to provide a complete theory of the domain, including a full specification of the initial situation and explicit formulas describing which fluents do *and do not* change whenever any action is taken. For instance, we would need to say that after a *Wait* action, a loaded gun continues to be loaded, if Fred was alive before he will be alive afterwards, and so on. The issue of stating the non-effects of actions is known as the *frame problem* [MH69]: how do we avoid having to represent the numerous axioms required to describe non-effects? We would like to omit or abbreviate these axioms somehow.

The frame problem is only one aspect of the problem of completeness; generally our knowledge will be deficient in other ways as well. For example,

- we may not know the truth value of every fluent in the initial situation.
- we may know the situation after some sequence of actions has been performed, but not know precisely which actions were taken. (This leads to one type of *explanation* problem.)
- we may not know precisely what effects an action has. This may be due to a simple lack of information, or to the fact that the action's effects are probabilistic (e.g., we might believe that there is a small chance that Fred could survive being shot). Note that even if we

[1] In general, we use upper case for constants and lower case for variables.

know the probabilities of the various action outcomes, the situation calculus's first-order language is too weak to express them.

In all such cases, it is unlikely that deductive reasoning will reach any interesting conclusions. For instance, if we leave open the logical possibility that the gun becomes unloaded while we wait, then there is nothing we can say with certainty about whether Fred lives.

Our strategy for investigating these issues is to examine a generalized notion of inference that not only reports certain conclusions (in those rare cases where our knowledge supports them), but also assigns degrees of belief (i.e., probabilities) to other conclusions. For instance, suppose KB is some knowledge base stating what we know about actions' effects, the initial situation, and so on, and we are interested in a query such as $\varphi = Alive(Result(Shoot, Result(Wait, Result(Load, S_0))))$. The next section shows how we define $\Pr(\varphi|KB)$, the degree of belief in φ (which is a number between 0 and 1) given our knowledge KB. It is entirely possible for KB to be such that $\Pr(\varphi|KB) = 0.1$, which would mean we should have high but not complete confidence that Fred would be dead after this sequence of actions. To a large extent, it is the freedom to assign intermediate probabilities (other than 0 or 1) that relieves us of traditional situation calculus' demand for complete knowledge. A related important feature is our ability to make use of statistical knowledge (for instance, an assertion that shooting only succeeds 90% of the time). Of course, the real success of our approach depends crucially on the details and behavior of the particular method we have for computing probabilities. Examining this method, and justifying its successes, is the goal of the rest of this paper.

Before continuing, we remark that the importance of the issues we have raised is well known. There have been numerous attempts to augment deductive reasoning with the ability to "jump to conclusions", i.e., *nonmonotonic* reasoning (e.g., [HM87, Kau86, Lif87]), often in an attempt to solve the frame problem. The idea of reasoning to "plausible" conclusions, rather than only the deductively certain ones, clearly shares some motivation with our decision to evaluate numeric probabilities. The connection is in fact quite deep; see [BGHK94]. However, the application of pure nonmonotonic logics to reasoning about actions has proven to be surprisingly difficult and, in any event, these approaches are not capable of dealing with probabilistic actions or with the quantitative assessment of probabilities.

There has also been work addressing the issue of probabilities in the context of actions. The propositional approaches to the problem (e.g., [Han90, DK89]) do not incorporate the full expressive power of the situation calculus. Furthermore, even those that are able to deal with abductive queries typically cannot handle explanation problems (since they do not place a prior probability distribution over the space of actions). [Ten91] achieves a first-order ontology by applying the reference-class approach of [Kyb74] to this problem. His approach, however, has a somewhat "procedural" rather than a purely logical (semantic) character. Hence, although it specifies how to do forward projection—assessing probabilities for outcomes given knowledge of an initial situation—it does not support arbitrary queries from arbitrary knowledge bases. This flexibility is important, particularly for explanation and diagnosis. Finally, none of these works subsume all the issues addressed by advocates of nonmonotonic reasoning. Our approach provides a framework for dealing with these issues in a uniform fashion.

2.2 Random-worlds

We now turn to a summary of the *random-worlds method*; see [BGHK94] and the references therein for full details. We emphasize that this is a general technique for computing probabilities, given arbitrary knowledge expressed in a very rich language; it was *not* developed specifically for the problem of reasoning about action and change. As a general reasoning method, random-worlds has been shown to possess many attractive features [BGHK94], including a preference for more specific information and the ability to ignore irrelevant information. In a precise sense, it generalizes both the powerful theory of default reasoning of [GMP90] and (as shown in [GHK94]) the principle of maximum entropy [Jay78]; it can also be used to do reference class reasoning from statistics in the spirit of [Kyb74].

The two basic ideas underlying the random-worlds method are the provision of a general language for expressing statistical information, and a mechanism for probabilistic reasoning from such information.

The language we use extends full first-order logic with statistical information, as in [Bac90]), by allowing *proportion expressions* of the form $||\varphi(x)|\psi(x)||_x$. This is interpreted as denoting the proportion of domain elements satisfying φ, among those satisfying ψ.[2] (Actually, an arbitrary set of variables is allowed in the subscript.) A simple *proportion formula* has the form $||\varphi(x)|\psi(x)||_x \approx 0.6$ where "$\approx$" stands for "approximately equal." Approximate equality is required since, if we make a statement like "90% of birds can fly", we almost certainly do not intend this to mean that *exactly* 90% of birds fly. Among other things, this would imply that the number of birds is a multiple of ten! Approximate equality is also important because it allows us to capture defaults. For example, we can express "Birds typically fly" as $||Fly(x)|Bird(x)||_x \approx 1$. We omit a description of the formal semantics, noting that the main subtlety concerns the interpretation of approximate comparisons, and that the special case of ≈ 1 is related to the well-known ϵ-semantics [Pea89].

The second aspect of the method is, of course, the specific way in which degrees of belief are computed. Before reviewing these, we remark that for the purposes of most of this paper, the random-worlds method can be regarded as a black box which, given any knowledge base KB and a query φ, assesses a degree of belief (i.e., a probability) $\Pr^w_\infty(\varphi|KB)$.

Very briefly, and ignoring the subtlety of approximate equality, the method is as follows. For any domain size N, we consider all the worlds (first-order structures) of size N consistent with KB. Let $\#worlds_N(KB)$ be the number of size N worlds that satisfy KB. Appealing to the principle of indifference, we regard all such worlds as being equally plausible. It then follows that, given a domain size N, we should define $\Pr^w_N(\varphi|KB) = \frac{\#worlds_N(\varphi \wedge \mathrm{KB})}{\#worlds_N(\mathrm{KB})}$. Typically, all that is known about N is that it is "large". Thus, the *degree of belief* in φ given KB is taken to be $\lim_{N\to\infty} \Pr^w_N(\varphi|KB)$.

Applying random-worlds in a temporal context is mostly a problem of choosing an appropriate representation scheme. Here we are guided mostly by the standard ontology of situation calculus, and reason about *situations* and *actions*. Indeed, since our language includes that of first-order logic, it would be possible to use the language of standard situation calculus without change. However, we want to do more than this. In particular, we want to allow probabilistic actions and statistical knowledge. To do this, we need to allow for actions that can have several

[2] If $\psi(x)$ is identically *true*, we generally omit it.

effects (even relative to the same preconditions). For this purpose, it is useful to conceptually divide a situation into two components: the *state* and the *environment*. The state is the *visible* part of the situation; it corresponds to the truth values of the fluents. The environment is intended to stand for all aspects of the situation not determined by the fluents (such as the time, or other properties of the situation that we might not wish to express explicitly within our language).

So what is a *world* in this context? Our worlds have a three-sorted domain, consisting of states, environments, and actions. *Situations* are simply state-environment pairs. Each world provides an interpretation of the symbols in our language over this domain, in the standard manner. For the purposes of this paper, fluents are taken to be unary predicates over the set of states.[3] Actions map situations to new situations via a *Result* function; hence, each world also provides, via the denotation of *Result*, a complete specification of the effect of an action on every situation.

Each state in the world's domain can be viewed as a truth assignment to the fluents. If we have k fluents in the language, say $P_1, \ldots, P_k$, we require that there be at most one state for each of the 2^k possible truth values of the fluents.[4] We do this by adding the following formula to the KB:

$$\forall v, v'((P_1(v) \equiv P_1(v') \wedge \cdots \wedge P_k(v) \equiv P_k(v')) \Rightarrow v = v').$$

Because the set of states is bounded, when we take the domain size to infinity (as is required by random worlds), it is the set of actions and the set of possible environments that grow unboundedly.

As stated above, action effects are represented using a *Result* function that maps an action and a situation to a situation. In order to formally define, within first-order logic, a function whose range consists of pairs of domain elements, we actually define two functions—$Result_1$ and $Result_2$—that map actions and situations to states and environments respectively. We occasionally abuse notation and use *Result* directly in our formulas. Note that the mapping from an action and a situation to a situation is still a deterministic one. However, *Result* is not necessarily deterministic when we only look at states. Two situations can agree completely in terms of what we say about them (their state), and nevertheless an action may have different outcomes.

As promised, this new ontology allows us to express nondeterministic and probabilistic actions, as well as the deterministic actions of the standard situation calculus. For example, consider a simple variant of the Yale Shooting Problem (YSP), where we have only two fluents, *Loaded* and *Alive*, and three actions, *Wait*, *Load*, and *Shoot*. Each world will therefore have (at most) four states, corresponding to the four possible truth assignments to *Loaded* and *Alive*. We assume, for simplicity, that we have constants denoting these states: $V_{AL}, V_{A\bar{L}}, V_{\bar{A}L}, V_{\bar{A}\bar{L}}$. Each world will also have domain elements corresponding to the three named actions, and possibly

[3] We observe that we can easily extend our ontology to allow complex fluents (e.g., *On(A,B)* in the blocks world), and/or reified fluents.

[4] This restriction was also used by Baker [Bak91] in his solution to the frame problem. It does not postulate the existence of a state for *all* possible assignments of truth values, and hence allows a correct treatment of ramifications. Baker then uses circumscription to ensure that there is exactly one state for each assignment of truth values *consistent with the* KB. In our framework, the combinatorial properties of random-worlds guarantee that this latter fact will hold in almost all worlds.

to other (unnamed) actions. The remaining domain elements correspond to different possible environments. The fluents are unary predicates over the states, and $Result_1$ is a function that takes a triple—an action, a state, and an environment—and returns a new state.[5] In the KB we can specify different constraints on $Result_1$. For example,

$$\forall v\ (Loaded(v) \Rightarrow ||\neg Alive(Result_1(Shoot, v, e))||_e \approx 0.9), \tag{1}$$

asserts that the *Shoot* action has probabilistic effects; it says that 90% of shootings (in a state where the gun is loaded) result in a state in which Fred is dead. On the other hand,

$$\forall v, e\ (Loaded(v) \Rightarrow \neg Alive(Result_1(Shoot, v, e))), \tag{2}$$

asserts that *Shoot* has the deterministic effect of killing Fred when executed in any state where the gun is loaded.

We might not know what happens is the gun is not loaded: Fred might still die of the shock. In such cases, we can simply leave this unspecified. Later in the paper, we discuss the different ways in which our language allows us to specify the effects of actions, and the conclusions these entail.

2.3 Counterfactuals

While our basic ontology seems natural, there are other possible representations. However, it turns out that the use of a *Result* function is crucial. Although the use of *Result* is quite standard in situation calculus, it is important to realize that its denotation in each world tells us the outcome of each action in all situations, including those situations that never actually occur. That is, in each world *Result* provides *counterfactual* information.

This can best be understood using an example. Consider the YSP example, where for simplicity we ignore environments and consider only a single action—*Shoot*—which is always taken at the initial state. We know that Fred is alive at the initial state, but nothing about the state of the gun—it could be loaded or not. Assume that, rather than having a *Result* function, we choose to have each world simple denote a single run (history) for this experiment. In this new ontology, we could use a constant V_0 denoting the initial state and another constant V_1 denoting the second state; each of these will necessarily be equal to one of the four states described above. In order to assert that shooting a loaded gun kills Fred, we would state that $Loaded(V_0) \Rightarrow \neg Alive(V_1)$. Furthermore, assume that after being shot the gun is no longer loaded. It is easy to see that there are essentially three possible worlds (up to renaming of states): if $Loaded(V_0)$ (so that $V_0 = V_{AL}$), then necessarily $V_1 = V_{\bar{A}\bar{L}}$, and if $\neg Loaded(V_0)$ then either $V_1 = V_{\bar{A}\bar{L}}$ or $V_1 = V_{A\bar{L}}$. The random-worlds method, used with this new ontology, would give a degree of belief of $\frac{1}{3}$ to the gun being loaded at V_0, simply because *Shoot* has more possible outcomes if the gun is unloaded. Yet intuitively, since we know nothing about the initial status of the gun, the correct degree of belief for $Loaded(V_0)$ is $\frac{1}{2}$. This is the answer we get by using the ontology of situation calculus with the *Result* function. In this case, the different worlds correspond to the different denotations of *Result* and V_0. Assuming that no action can revive

[5] Similarly, the $Result_2$ function returns a new environment, but there is usually no need for the user to provide information about this function.

Fred once he dies, there are only two possible denotations for *Result*: $Result(Shoot, V_{A\bar{L}})$ is either $V_{\bar{A}\bar{L}}$ or $V_{A\bar{L}}$, while $Result(Shoot, V) = V_{\bar{A}\bar{L}}$ if $V \neq V_{A\bar{L}}$. Furthermore, V_0 is either V_{AL} or $V_{A\bar{L}}$. Hence, there are four possible worlds. In exactly two of these, we have that $Loaded(V_0)$. The key idea here is that, because our language includes *Result*, each world must specify not only the outcome of shooting a loaded gun, but also the outcome of shooting *had the gun been unloaded*. Once this counterfactual information is taken into account, we get the answers we expect.

We stress that the *KB* does not need to include any special information because of our use of counterfactuals. As is standard in the situation calculus, we put into the *KB* exactly what we know about the *Result* function (for example, that shooting a loaded gun necessarily kills Fred). The *KB* admits a set of satisfying worlds, and in each of these worlds *Result* will have some counterfactual behavior. The random worlds method takes care of the rest by counting among these alternate behaviors.

The example above and the results below show that random worlds works well with an ontology that has implicit counterfactual information (like the situation calculus and its *Result* function). On the other hand, with other ontologies (such as the language used above that simply records what actually happens and nothing more) the combinatorics lead to unintuitive answers. Hence, it might seem that counterfactual ontologies are simply a technical requirement of random worlds. However, the issue of counterfactuals seems to arise over and over again in attempts to understand temporal and causal information. They have been used in both philosophy and statistics to give semantics to causal rules [Rub74]. In game theory [Sta94] the importance of counterfactuals (or strategies) has long been recognized. Baker's approach [Bak91] to the frame problem is, in fact, also based on the use of counterfactuals.

We have already mentioned that random-worlds subsumes the principle of maximum entropy. It has been argued [Pea88] that maximum entropy (and hence random-worlds) cannot deal appropriate with causal information. In fact, our example above is closely related, in a technical sense, to the problematic examples described by Pearl. But once again, an appropriate representation of causal rules using counterfactuals solves the problem [Hun89]. In fact, counterfactuals have been used recently to provide a formulation of Bayesian networks based on deterministic functions [Pea93]. All these applications of counterfactuals turn out to be closely linked to our own, even though none consider the random-worlds method. The ontology of this paper is, in some sense, the convergence of these technically diverse, but philosophically linked, frameworks. As our results suggest, the generality of the random-worlds approach may allow us to draw these lines of research together, and so expose the common core.

3 Results

As a minimal requirement, we would like our approach to be compatible with standard deductive reasoning, whenever the latter is appropriate. As shown in [BGHK94], this desideratum is automatically satisfied by random worlds:

Proposition 3.1: *If φ is a logical consequence of a knowledge base* KB*, then* $\Pr^w_\infty(\varphi|\mathrm{KB}) = 1$.

Hence, our approach supports all the conclusions that can be derived using ordinary situation calculus. However, as we now show, it can deal with much more.

An important concept in reasoning about change is the idea of a *state transition*. In our context, a state transition takes us from one situation to the next via the *Result* function. Since we can only observe the state component of a situation, we are particularly interested in the probability that an action takes us from a situation $(V, \cdot)$ to another $(V', \cdot)$ (where the specific identity of the environment is irrelevant). We are in fact interested in the *transition probability* $\Pr_\infty^w(\mathit{Result}(A, V, E) = V' | \mathit{KB})$. As we show later on in this section, these transition probabilities can often be used to compute the cumulative effects of sequences of actions.

We can use the properties of random worlds to derive transition probabilities from our action descriptions. Consider a particular state V and action A. There are many ways in which we can express knowledge relevant to associated transition probabilities. One general scheme uses assertions of the form

$$\forall e\, (\varphi(\mathit{Result}_1(A, V, e))), \tag{3}$$

where φ is a Boolean combination of fluents. Assertion (3) says that φ is true of all states that can result from taking A at state V. In general, when KB entails such a statement, then Proposition 3.1 can be used to show that our degree of belief in $\varphi(\mathit{Result}_1(A, V, E)) = 1$. For example, if KB consists of (2) only, then $\Pr_\infty^w(\mathit{Alive}(\mathit{Result}_1(\mathit{Shoot}, V_{AL}, E))|\mathit{KB}) = 0$, as expected (here, φ is Alive).

Assertion (2) describes a deterministic effect. However, even for nonprobabilistic statements such as (3), our approach can go far beyond deductive reasoning. For instance, we might not always know the full outcome of every action in every state. A *Load* action might result in, say, between one and six bullets being placed in the gun. If we have no other information, our approach would assign a degree of belief of $\frac{1}{6}$ to each of the possibilities. In general, we can formalize and prove the following result (where, as in our remaining results, E is a constant over environments not appearing anywhere in KB):

Proposition 3.2: *Suppose* KB *contains (3), but no additional information about the effects of A in V. Then, $\Pr_\infty^w(\mathit{Result}_1(A, V, E) = V'|\mathrm{KB}) = \frac{1}{m}$, where m is the number of states satisfying φ, and V' is one of these states.*

We note that we can prove a similar result in the case where our ignorance is due to incomplete information about the initial state (as illustrated in the previous section).

As we discussed, our language can also express information about probabilistic actions (where we have statistical knowledge about the action's outcomes). Our theory also derives many of the conclusions we would expect. For example, if KB contains (1), then we would conclude $\Pr_\infty^w(\neg \mathit{Alive}(\mathit{Result}_1(\mathit{Shoot}, V, E))|\mathit{KB} \wedge \mathit{Loaded}(V)) = 0.9$. In general, the *direct inference* property exhibited by random worlds allows us to prove the following:

Proposition 3.3: *If* KB *entails $||\varphi(\mathit{Result}(A, V, e))||_e \approx \alpha$, then $\Pr_\infty^w(\varphi(\mathit{Result}(A, V, E))|\mathrm{KB}) = \alpha$.*

Nondeterminism due to ignorance on the one hand, and probabilistic actions on the other, are similar in that they both lead to intermediate degrees of belief between 0 and 1. Nevertheless, there is an important conceptual difference between the two cases, and we consider it a significant feature of our approach that it can capture and reason about both.

Given our statistical interpretation of defaults, the ability to make statistical statements about the outcomes of actions also allows us to express a *default assumption* of determinism. For instance, $\forall v\,(\mathit{Loaded}(v) \Rightarrow ||\neg \mathit{Alive}(\mathit{Result}_1(\mathit{Shoot}, v, e))||_e \approx 1)$ states that shooting a loaded gun *almost* surely kills Fred. Even though a default resembles a deterministic rule in many ways, the distinction can be important. We would prefer to explain an unusual occurrence by finding a violated default, rather than by postulating the invalidity of a law of nature (which would result in inconsistent beliefs). For example, if, after the shooting, we observe Fred walking away, then our approach would conclude that Fred survived the shooting, rather than that he is a zombie. This distinction between certain outcomes and default outcomes is also easily made in our framework.

In general, we may have many pieces of information describing the behavior of a given action at a given state. For example, consider the YSP with an additional fluent *Noisy*, where our *KB* contains (1) and

$$\forall v\,(\mathit{Loaded}(v) \Rightarrow ||\mathit{Noisy}(\mathit{Result}_1(\mathit{Shoot}, v, e))||_e \approx 0.8).$$

Given all this information, we would like to compute the probability that shooting the gun in a state V where $\mathit{Alive}(V) \wedge \mathit{Loaded}(V)$ results in the state V_{ALN} (where N stands for *Noisy*). Unless we know otherwise, it seems intuitive to assume that Fred's health in the resulting state should be independent of the noise produced; that is, the answer should be $0.1 \times 0.8 = 0.08$. This is, in fact, the answer produced by our approach. This is an instance of a general result, asserting that transition probabilities can often be computed using *maximum entropy*. While, we do not have the space to fully describe the general result, we note that it entails a *default assumption of independence*. That is, unless we have reason to believe that *Alive* and *Noisy* are correlated, our approach will assume that they are not. We stress that this is only a default. We might know that *Alive* and *Noisy* are negatively correlated (perhaps because lack of noise is sometimes caused by a misfiring gun). In this case we can easily add to the *KB*, for example, that $\forall v\,(\mathit{Loaded}(v) \Rightarrow ||\mathit{Noisy}(\mathit{Result}_1(\mathit{Shoot}, V, e)) \wedge \mathit{Alive}(\mathit{Result}_1(\mathit{Shoot}, V, e))||_e \approx 0.05)$. The resulting *KB* is not inconsistent; the default assumption of independence is dropped automatically.

We now turn to the problem of reasoning about the effects of a sequence of actions. The *Markov* assumption, which is built into most systems that reason about probabilistic actions [Han90, DK89], asserts that the effects of an action depend only on the state in which it is taken. As the following result demonstrates, our approach *derives* this principle from the basic semantics. We note that the Markov assumption is only a default assumption in our framework; it fails if the *KB* contains assertions implying otherwise. Formally, it requires that our information about *Result* be expressed solely in terms of *transition proportions*, i.e., proportion expressions of the form $||\varphi(\mathit{Result}_1(A, V, e))||_e$, where φ is a Boolean combination of fluents. Hence, if our *KB* contains information about $||\mathit{Result}(A_1, \mathit{Result}(A_2, V, e))||_e$, the Markov property might no longer hold.

Proposition 3.4: *Suppose that the only occurrence of Result in* KB *is in the context of transition proportions, and that E and E' do not appear in* KB. *Then*

$$\Pr^w_\infty(\mathit{Result}(A_1, V, E) = (V', E') \wedge \mathit{Result}_1(A_2, V', E') = V'' \mid \mathrm{KB}) = \\ \Pr^w_\infty(\mathit{Result}_1(A_1, V, E) = V' | \mathrm{KB}) \;\times\; \Pr^w_\infty(\mathit{Result}_1(A_2, V', E') = V'' | \mathrm{KB}).$$

Of course, it follows from the proposition that to compute $\Pr_\infty^w(\mathit{Result}(A_2, \mathit{Result}(A_1, V, E)) = V''$, we just sum over all intermediate states. This result generalizes to arbitrary sequences of actions in the obvious way.

4 The Frame Problem

Perhaps the best single illustration of the power of our approach in the context of the situation-calculus is its ability to deal simply and naturally with the frame problem. Many people have an intuition about the frame problem which is, roughly speaking, that "fluents tend not to change value very often". This suggests that if we could formalize this general principle (that change is unusual), it could serve as a substitute for the many explicit frame axioms that would otherwise be needed. However, as shown in [HM87], the most obvious formulations of this idea in standard nonmonotonic logics often fail. Suppose we use a formalism that, in some way, tries to minimize the number of changes in the world. In the YSP, after waiting and then shooting we expect there to be *some* change: we expect Fred to die. But there is another model which seems to have the "same amount" of change: the gun miraculously becomes unloaded as we wait, and thus Fred does not die. This seems to be the wrong model, but it turns out to be difficult capture this intuition formally. Subsequent to Hanks and McDermott's paper, there was much research in this area before adequate solutions were found.

How does our approach fare? It turns out that we can use our statistical language to directly translate the intuition we have about frame axioms, and the result gives us exactly the answers we expect in such cases as the YSP. We formalize the statement of minimal change for a fluent P by asserting that it changes in very few circumstances; that is, any action applied in any situation is unlikely to change P: $||P(\mathit{Result}_1(a, v, e)) \not\equiv P(v)||_{(a,v,e)} \approx 0$. Of course, the statistical chance of such frame violations cannot be exactly zero, because some actions do cause change in the world. However, the "approximately equals" connective allows for this. Roughly speaking, the above axiom, an instance of which can be added for each fluent P for which we think the frame assumption applies, will cause us to have degree of belief 0 in a fluent changing value unless we have explicit knowledge to the contrary.[6]

There is one minor subtlety. Recall that in the random-worlds approach, we consider the limit as the domain tends to infinite size. As we observed, since the number of states is bounded, this means that the number of environments and actions must grow without bound. This does not necessarily mean that the number of actions grows without bound. However, in the presence of the frame axioms (as given above), we need this stronger assumption. This need is quite easy to explain. If the only action is *Shoot*, then half the triples (a, v, e) (those where *Loaded* is true in v) would lead to a change in the fluent *Alive*. In this framework, it would be inconsistent to simultaneously suppose that there is only one way of changing the world (i.e., *Shoot*) and also that every fluent (and in particular, *Alive*) hardly ever changes. Making the quite reasonable assumption that there are many other ways of effecting change in the world (i.e., many other actions in the domain), even though we may say nothing about them, removes the contradiction.

[6] Note that having degree of belief 0 does not mean that we believe something to be impossible, but only extremely unlikely. Hence, this representation does allow for unexpected change, a useful feature in explanation problems.

Given this, if we add frame axioms as given above we get precisely the results we want. If we try to predict forward from one state to the next, we conclude (with degree of belief 1) that nothing changes except those fluents that the action is known to affect. If we consider a sequence of actions, we can predict the outcome by applying this rule for the first action with respect to the initial state, then applying the second action to the state just obtained, and so on. This is essentially a consequence of Proposition 3.4, combined with the properties of our frame axiom. In the YSP, for example, the *Load* action will cause the gun to be loaded, but will change nothing else. *Wait* will then leave the state completely unchanged. Finally, because the gun will still be loaded, performing *Shoot* will kill Fred as expected.

The idea of a formal theory being faithful to this intuitive semantics (essentially, that in which we consider actions one at a time, assuming minimal change at each step) has recently been formalized by Kartha [Kar93]. Roughly speaking, he showed that a simple procedural language $\mathcal{A}$ [GL92] can be embedded into three approaches for dealing with the frame problem [Bak91, Ped89, Rei91], so that the answers prescribed by $\mathcal{A}$'s semantics (which are the intuitively "right" answers) are also obtained by these formalisms. The following result shows that we also pass Kartha's test. Specifically:

Proposition 4.1: *There is a sound and complete embedding of $\mathcal{A}$ into our language in which the frame axioms appear in the above form.*

Thus, the random-worlds approach succeeds in solving the frame problem as well as the above approaches, at least in this respect. However, as we mentioned above, our approach is significantly more expressive, in that it can deal with quantitative information in a way that none of these other approaches can. Furthermore, our approach does not have difficulty with state constraints (i.e., ramifications), a problem encountered by a number of other solutions to the frame problem (e.g., those of Reiter and Pednault).

Why does the random-worlds method work so easily? There are two reasons. First, the ability to say that proportions are very small lets us express, in a natural way *within our language*, the belief that frame violations are rare. Alternative approaches to the problem tend to use powerful minimization techniques, such as circumscription, to encode this. But much more important is our use of an ontology that includes counterfactuals. This turns out to be crucial in avoiding the YSP. Even if the gun does in fact become unloaded somehow, we do not escape the fact that shooting with a loaded gun *would have* killed Fred. Baker and Ginsberg's [BG89] solution to the frame problem (based on circumscription) relies on a similar notion of counterfactual situations. But while the solutions are related, they are not identical: for instance, we do not suffer from the problem concerning extraneous fluents that Baker [Bak89] mentions.[7]

Some solutions to the YSP work by augmenting a principle of minimal change with a requirement that we should prefer models in which change occurs as late as possible (e.g., [Kau86, Sho88]). This solves the original YSP because the model in which Fred dies violates the frame axiom (that Fred should remain alive) later than the model in which the gun miraculously becomes unloaded. However, it has been observed that such theories fail on certain explanation

[7]We also note that Baker and Ginsberg's solution was constructed especially to deal with the problem of minimizing frame violations. Our solution to the frame problem and the YSP arises naturally and almost directly from our general approach.

problems, such as Kautz's [Kau86] stolen car example. Our approach deals well with explanation problems. In Kautz's example, we park our car in the morning only to find when we return in the evening that it has been stolen. Theories that delay change lead to the conclusion that the car was stolen just prior to our return. A more reasonable answer is to be indifferent about exactly when the car was stolen. Our approach assigns equal probability to the car being stolen over each time period of our absence. That is, if KB axiomatizes the domain in the natural way, and the only action that makes a car disappear from the parking lot is the $StealCar$ action, then we would conclude that:

$$\Pr_\infty^w(A_i = StealCar | KB \wedge \neg Parked(Result_1(A_\ell, Result(\cdots \\ Result(A_1, Result(ParkCar, V_0, E)) \cdots)))) = \tfrac{1}{\ell}.$$

5 Conclusion

As shown in [BGHK94], the random-worlds approach provides a general framework for probabilistic and default first-order reasoning. The key to adapting random worlds to the domain of causal and temporal reasoning lies in the use of counterfactual ontologies to represent causal information. Our results show that the combination of random worlds and counterfactuals can be used to address many of the important issues in this domain. The ease with which the general random-worlds technique can be applied to yet another important domain, and its success in dealing with the core problems encountered by other approaches, shows its versatility and broad applicability as a general framework for inductive reasoning.

There is, however, one important issue which this approach fails to handle appropriately: the *qualification problem*. The reasons for this failure are subtle, and cannot be explained within the space limitations. However, as we discuss in the full paper, the problem is closely related to the fact that random worlds does not learn statistics from samples. This aspect of random-worlds was discussed in [BGHK95], where we also presented an alternative method to computing degrees of belief, the *random-propensities* approach, that does support learning. In future work, we hope to apply this alternative approach to the ontology described in this framework. We have reason to hope that this approach will maintain the desirable properties described in this framework, and will also deal with the qualification problem.

References

[Bac90] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge.* MIT Press, Cambridge, Mass., 1990.

[Bak89] A. Baker. A simple solution to the Yale shooting problem. In R. J. Brachman, H. J. Levesque, and R. Reiter, editors, *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pages 11–20, San Francisco, 1989. Morgan Kaufmann.

[Bak91] A. Baker. Nonmonotonic reasoning in the framework of the situation calculus. *Artificial Intelligence*, 49:5–23, 1991.

[BG89] A. Baker and M. L. Ginsberg. Temporal projection and explanation. In *Proc. Eleventh International Joint Conference on Artificial Intelligence (IJCAI '89)*, pages 906–911, 1989.

[BGHK94] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistical knowledge bases to degrees of belief. Technical Report 9855, IBM, 1994. To appear, *Artificial Intelligence*. Available by anonymous ftp from logos.uwaterloo.ca/pub/bacchus or via WWW at http://logos.uwaterloo.ca. A preliminary version of this work appeared in *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, 1993, pages 563–569.

[BGHK95] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. Reasoning with noisy sensors in the situation calculus. In *Proc. Fourteenth International Joint Conference on Artificial Intelligence (IJCAI '95)*, pages 1933–1940, 1995.

[DK89] T. Dean and K. Kanazawa. Persistence and probabilistic projection. *IEEE Tran. on Systems, Man and Cybernetics*, 19(2):574–85, 1989.

[GHK94] A. J. Grove, J. Y. Halpern, and D. Koller. Random worlds and maximum entropy. *Journal of A.I. Research*, 2:33–88, 1994.

[GL92] M. Gelfond and V. Lifschitz. Representing actions in extended logic programming. In K. Apt, editor, *Logic Programming: Proc. Tenth Conference on Logic Programming*, pages 559–573, 1992.

[GMP90] M. Goldszmidt, P. Morris, and J. Pearl. A maximum entropy approach to nonmonotonic reasoning. In *Proc. National Conference on Artificial Intelligence (AAAI '90)*, pages 646–652, 1990.

[Han90] S. J. Hanks. *Projecting Plans for Uncertain Worlds*. PhD thesis, Yale University, 1990.

[HM87] S. Hanks and S. McDermott. Nonmonotonic logic and temporal projection. *Artificial Intelligence*, 33(3):379–412, 1987.

[Hun89] D. Hunter. Causality and maximum entropy updating. *International Journal of Approximate Reasoning*, 3(1):379–406, 1989.

[Jay78] E. T. Jaynes. Where do we stand on maximum entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118. MIT Press, Cambridge, Mass., 1978.

[Kar93] G. Kartha. Soundness and completeness theorems for three formalizations of action. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, pages 724–729, 1993.

[Kau86] H. Kautz. A logic of persistence. In *Proc. National Conference on Artificial Intelligence (AAAI '86)*, pages 401–405, 1986.

[Kyb74] H. E. Kyburg, Jr. *The Logical Foundations of Statistical Inference.* Reidel, Dordrecht, Netherlands, 1974.

[Lif87] V. Lifschitz. Formal theories of action: Preliminary report. In *Proc. National Conference on Artificial Intelligence (AAAI '87)*, pages 966–972. 1987. Reprinted in *The Frame Problem in Artificial Intelligence*, F. Brown (ed.), Morgan Kaufmann, San Francisco, pp. 121–127.

[MH69] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In D. Michie, editor, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, Edinburgh, 1969.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Francisco, Calif., 1988.

[Pea89] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: a survey. In R. J. Brachman, H. J. Levesque, and R. Reiter, editors, *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pages 505–516, 1989. Reprinted in *Readings in Uncertain Reasoning*, G. Shafer and J. Pearl (eds.), Morgan Kaufmann, San Francisco, Calif., 1990, pp. 699–710.

[Pea93] J. Pearl. Aspects of graphical models connected with causality. In *49th Session of the International Statistics Institute*, 1993.

[Ped89] E. Pednault. ADL: Exploring the middle ground between STRIPS and the situation calculus. In R. J. Brachman, H. J. Levesque, and R. Reiter, editors, *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pages 324–332, San Francisco, 1989. Morgan Kaufmann.

[Rei91] R. Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In V. Lifschitz, editor, *Artificial Intelligence and Mathematical Theory of Computation*, pages 359–380. Academic Press, New York, 1991.

[Rub74] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. In *Journal of Educational Psychology*, volume 66, pages 688–701, 1974.

[Sho88] Y. Shoham. Chronological ignorance: experiments in nonmonotonic temporal reasoning. *Artificial Intelligence*, 36:271–331, 1988.

[Sta94] R. C. Stalnaker. Knowledge, belief and counterfactual reasoning in games. In C. Bicchieri and B. Skyrms, editors, *Proceedings of the Second Castiglioncello Conference.* Cambridge University Press, 1994. To appear.

[Ten91] J. D. Tenenberg. Abandoning the completeness assumptions: A statistical approach to the frame problem. *International Journal of Expert Systems*, 3(4):383–408, 1991.