Abandoning the Completeness Assumptions: A Statistical Approach to the Frame Problem

Josh D. Tenenberg*
Computer Science
University of Rochester
Rochester, New York
U.S.A., 14627
josh@cs.rochester.edu

1 Introduction

The purpose of this paper is twofold. The first is to challenge the assumptions regarding knowledge completeness that planning agents are often taken as possessing, and the solutions to the frame problem that depend upon these assumptions. The second is to present preliminary ideas regarding an approach to the frame problem that provides a more explicit formulation of an agent's uncertainty using statistically derived probabilities.

The frame problem, inferring what does not change as a result of an agent's actions, is an instance of the more general problem of predicting

*This work was supported in part by the Air Force Systems Command, Rome Air Development Center, Griffiss Air Force Base, New York 13441-5700, and the Air Force Office of Scientific Research, Bolling AFB, DC 20332 under Contract Number F30602-85-C-0008 which supports the Northeast Artificial Intelligence Consortium (NAIC), in part by ONR/DARPA research contract #N00014-80-C-0197, and in part by U.S. Army Communication-Electronics Command grant #DAAB10-87-D-022.

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*

2. REPORT DATE
1990

3. REPORT TYPE AND DATES COVERED
Unknown

4. TITLE AND SUBTITLE
Abandoning the Completeness Assumptions: A Statistical Approach to the Frame Problem

5. FUNDING NUMBERS
DAAB10-86-C-0567

6. AUTHOR(S)
Josh D. Tenenberg

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Rochester
Department of Computer Science
Rochester, NY 14627

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army CECOM Signals Warfare Directorate
Vint HIll Farms Station
Warrenton, VA 22186-5100

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
92-TRF-0014

1. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Statement A; Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The purpose of this paper is twofold. The first is to challenge the assumptions regarding knowledge completeness that planning agents are oftentaken as possessing, and the solutions to the frame problem that depend upon these assumptions. The second is to present preliminary ideas regarding and approach to the frame problem that provides a more explicit formulation of an agent's uncertainty using statistically derived probabilities.

1. SUBJECT TERMS
Artificial Intelligence, Data Fusion, Frame Problem
Statistics, Completeness Assumptions

15. NUMBER OF PAGES
34

18. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UL

what is true of the future, based upon what is currently known (the *prediction problem*). Previous attempts at solving the frame problem fall into two broad, yet distinct, classes, neither of which is extensible to the prediction problem. In one case, the agent's lack of knowledge of change is warrant for inferring that change has *not* occured. Unfortunately, this approach has promise only in worlds in which few changes occur beyond the agent's purview. Otherwise, the correspondence between the agent's beliefs and the world quickly diverges. In the other case, inferring non-occurence of change is sanctioned *only* when it is known that every action which can cause change did not occur. But for incompletely known worlds, few assertions can meet this stringent criterion. Hence, the agent's knowledge of what is currently true quickly dissipates.

I propose[1] an alternative approach that seeks a middle ground between the permissiveness of the first class, and the caution of the second. The central idea is to consider that actions achieve goals relative to a context with some *likelihood* of success. Due to errors in representation, in measurement, and in motor control, actions are not guaranteed to have their intended effect.

Stated coarsely, one has knowledge of the form

$$\%(\mathit{inRoom}(s+1))|\mathit{inRoom}(s) \wedge \mathit{pushOut}(s)) \in [.9, .95].$$

That is, the proportion of times that an effect holds (e.g., that an object is in the room) in a given state when in the previous state, a set of preconditions holds (the object was just in the room) and an action occurs (the agent attempts to push the object out of the room) is in the interval [.9, .95]. From this general statistic over sets of previous times in which an object was in the room, one can induce a probability about a *particular* time in which the object is predicted to be in the room. This rule, then, encodes knowledge that has typically been encapsulated in frame axioms. Further, this above statistic also provides an approach for attacking the *qualification* problem. Namely, that all qualifications not explicitly mentioned (such as that the object might be glued to the floor, that the object might be slippery, etc.) are summarized by the fact that these qualifications only defeat the success of the action some small proportion of the time (e.g., in the above case, between 5% and 10% of the time).

[1] Much of this work was done in collaboration with Jay Weber, Lockheed.

The novelty of this approach is its use of *statistically based* probabilities, an application of Kyburg's theory of probability, as described in [Kyburg, 1974]. Kyburg's theory provides not only a means for relating probabilities to statistics, but also preferences rules for mediating between conflicting statistics.

Although this work is still in its preliminary stages, this general approach shows promise in providing

1. a uniform framework for the frame and the qualification problems,
2. extensibility to the prediction problem,
3. a means for an agent's knowledge to slowly obsolesce, rather than being either eternally perpetuated, or prematurely discarded,
4. sanction for inferring the truth of propositions whose truth was not necessarily known in the *previous* instant,
5. the ability to encode knowledge of the form that certain kinds of events occur with given frequencies.

The following section examines the frame problem and the previous two broad solution classes. Each of these classes, and their concommitant shortcomings are examined in some detail. Section 5 discusses the probabilistic approach.

2 The Frame Problem

The first clear formalization of action reasoning in artificial intelligence [McCarthy and Hayes, 1981], the ***situation calculus***, involved using sets of timestamped sentences to represent states of the world. Each time stamp denotes a ***situation***, a snapshot of the world. Typical sentences are of the form "P is true in situation S." These sentences are viewed as encoding beliefs that an agent might have about its world, which it could use to find plans of action to satisfy its goals. Expressions such as "A is on B", which in static theories would be truth-functional propositions, are called ***fluents*** in the situation calculus, and denote functions from situations to truth values. A fluent and a situation ("A is on B in S") is called a *proposition*, and evaluates to either true or false.

Representing actions amounts to writing down "laws of motion" [Hayes. 1981] of the form "If fluent P holds in situation S, then fluent Q will hold in the situation resulting from applying action A." For instance, one such law of motion is "if there is nothing on the box now, and I am next to the box, then I will be on the box after I climb onto it." That is, these axioms relate facts believed true at one time to facts believed to become true at the next time as a result of the agent's action. Leaving aside the problem of whether we can provide sufficient antecedent conditions, (the *qualification problem*), these laws of motion underspecify what is true in the resulting state, since no axioms sanction inferring the truth of any propositions other than Q (and perhaps those things implied by Q). That is, from the law of motion above I can neither infer that following the climb action A has remained the same color, or that the box is still in the same room.

Providing a representation that sanctions these inferences is deceivingly difficult. One could, of course, "list systematically all conceivable facts which are *not* changed" [Hayes, 1971]. That is, one could have *frame axioms* stating that for each proposition, whenever *some non-affecting action occurs*, that proposition remains unchanged. But, as Hayes argues, "it is clearly going to be impractical in any elaborate system. That is the frame problem". In other words, the problem is to specify the actions of an agent "in some economical and principled way" [Hayes, 1987] such that all of the inferences sanctioned by a complete set of *frame axioms* are likewise permitted under the economical encoding.

It is clear that these inferences cannot be ignored, since beliefs about distant future states (such as that we are holding a banana after attempting to grab it) will depend crucially upon what we believe is true in the proximal future (such as, that the bananas do not disappear as we approach them). Without drawing such inferences, our knowledge about the world is inadequate for planning even the simplest tasks.

The difficulty of this problem is evidenced by the number of researchers attacking it, and the variety and ingenuity of approaches. These range from excplicitly encoding frame axioms in first-order logic and using theorem proving in order to plan [Green, 1969, Kowalski, 1979], to embedding frame axioms procedurally, or within some aspect of the representation [Fikes and Nilsson, 1971, Hewitt, 1972, Pednault, 1985], to monotonic approaches that use either domain-dependent "explanation closure axioms" [Haas, 1987, Schubert, 1989, Weber, 1989], that attempt to enumerate the complete set of actions which can alter the truth value of a proposition, or equivalently that use complete sets of independence assertions [Hayes, 1971, Georgeff, 1987], to (most popularly), the definition, development, and use of non-monotonic logics [Reiter, 1980, McCarthy, 1977, Lifschitz, 1987, Haugh, 1987, Shoham, 1986, Kautz, 1986] that use unsound inferences to conclude that things do not change unless there is explicit knowledge to the contrary.

Although there are subtle differences between the various approaches, there are two broad classes into which they can be grouped: the permissive and conservative approaches. Each of these will be looked at in turn.

3 Permissive Approaches

The permissive class includes the explicit frame axiom, the procedural, and the non-monotonic approaches. The intent is to represent the *independence* of actions and world state, ([Hayes, 1987]):

> If you pick up a cup from a table, lots of things don't happen: The cup doesn't break, the table doesn't move, the walls don't move, your hair doesn't change color, and Jerry Fodor's refrigerator door doesn't move; *or rather, if any of these things happen, it's not because of your picking up the cup.* [emphasis added]

Despite this intent, these approaches say quite a different thing: they say that the world is obliged to stay precisely as it is whenever an agent applies an action, save for the known, local effects of that action. That is, rather than expressing

(1) "action A does not affect P,"

the permissive approaches instead say

(2) "P does not change when A occurs."

The first allows P to change or not, irrespective of the occurence of A, while the second prohibits P's change during A. The cup cannot break, the walls cannot move, Pat's hair color must remain unchanged, and Jerry Fodor's refrigerator door does not move whenever I pick up my coffee cup, *by definition*, regardless of any efforts to the contrary by these other agents. Assertions of type 2 allow the truth value of P to be known following A if it was known previously, and is thus preferable to assertions of type 1 in terms of minimizing knowledge loss. But these inferences are sanctioned by the stronger axioms only under the following assumptions:

1. all change is precipitated by an acting agent,

2. all acts occurring in the world are known by the reasoning agent,

3. the relationship between an act and *all* of its consequences are known by the reasoning agent.

And all permissive approaches embed these assumptions in some form.

Although Green [1969] uses frame axioms that explicitly state assertion 2, for every such P and A, most other permissive approaches involve an assumption that relations *persist* over time, in the absence of knowledge to the contrary. STRIPS [Fikes and Nilsson, 1971] persists all assertions in a knowledge base not explicitly deleted by an action-denoting *operator*, Kautz [1986] and Shoham [1986] persist the truth of assertions in chronological order until there is a conflict with known future facts, and Lifschitz [1987], Haugh [1987], and Weber [1988] persist assertions in the absence of contradicting causative acts.

Regardless of the particular method, no permissive approach addresses the *intended* issue that "most of the world carries on in just the same way that it did before, or would have done" [Hayes, 1987], since most of the world is not taken into account. To formalize reasoning about change in a world brimming with multitudinous active agents is a difficult task; but to assume the world acts lock-step and in full view of the agent, to whom all causation is transparent, is to invite disaster. It is ironic, then, that inferences intended to reflect that "[M]ost events ... make only small local changes in the world" [Hayes, 1981], have precisely the opposite meaning.

The problems with assuming knowledge completeness stem both from the existence of other active agents, and from the inherent inability of the agent to develop a complete world model. Every kindergarten child knows that blocks that have been stacked do not remain in that state for long, due to the other children that might want to build structures of their own. Even the child playing alone develops an understanding of the instabilities of block towers. We can well imagine a future Robot-Chaplin inadvertently knocking over the stack of blocks it has just constructed with some attached projection as it bends over to begin another task.

Further, the agent will be subject to errors both in its perceptions and its motor control. There are inherent limitations in the measurements of physical parameters, such as size, shape, and location. An agent might thus traverse a path in the wrong direction, upsetting a predicted persistence in the process. Similarly for the agent's ability to finely control its own motor behavior.

3.1 Consequences of the Completeness Assumption

My criticism thus far is that assumptions of completeness wll lead to incorrect conclusions about any non-trivial real world domain. I will examine the following consequences of embedding this assumption.

Permissive approaches

1. are not scalable to worlds in which there are many active agents, and the agent has limited sensory bandwidth,
2. are difficult to assess since the relevant evidence is not explicit,
3. do not depreciate knowledge as it obsolesces, allowing for the propogation of error,
4. have been too constrained in the types of knowledge they encode.

3.1.1 Scalability

Implicit in the permissive approaches is the argument that solutions to simple problems will first be obtained [Hayes, 1971]:

> This approach to the frame problem makes few presuppositions. The two most important are that the world is *deterministic*, that is, changes do not occur spontaneously; and that there is only a single agent in the universe.

and then extended to more complex domains [Haugh, 1987]:

> "Much work remains to extend the simple formulation presented here to handle ... more complex facts, ... multiple simultaneous events."

until eventually, solutions to general problems will be obtained.

But this position is fatally flawed, since all permissive approaches rely *crucially* on either explicit (and overly strong) frame axioms, or on completeness assumptions as the means for drawing inferences about what relations are unchanged in subsequent world states. Abandoning the completeness assumptions is equivalent to abandoning the approaches built upon these assumptions. For instance, if the agent is *not* assumed to know all acts that occur, nor all consequences of each act, then the very essence of Lifschitz's solution [1987] that minimizes the causality and precondition predicates evaporates. The beauty and the substance of his solution is in formalizing this very assumption.

3.1.2 Loss of Evidence

As Shoham points out, [1986], the permissive approaches, (particularly the use of non-monotonic logic) follow the "*what-you-don't-know-won't-hurt-you*" principle[2].

> To use the standard example from the literature, when we prefer models in which birds can fly to those in which they cannot, we do not have to specify the flying capabilities of most birds; their ability to fly will follow "automatically." The price we pay is when reasoning about "abnormal" birds such as penguins: if we do not specify their (nonexistent) flying capabilities then we will erroneously infer that they can fly (whereas in an ordinary monotonic logic, in the absence of explicit information about it, we would neither infer that they can fly or that they cannot). When the frequency of flying birds overwhelms the expected danger of making wrong predictions about penguins, we assume that birds fly by default.

[2] One might not think that the monotonic approaches [Green, 1969, Kowalski, 1979, Fikes and Nilsson, 1971, Pednault, 1985] are subject to the same objections. But their approaches seem to be the biggest casualties, since not only do they attempt to provide a complete set of frame axioms, they do not admit the possibility that some of their inferences might be defeasible, as do the non-monotonic approaches.

Complete knowledge is therefore not *required*, according to Shoham; but the consequences of lack of knowledge must not be too costly. Unfortunately, none of the frequency statistics, nor the utility functions inherent in the above argument are made explicit in any of the permissive approaches. The evidence upon which the defaults are based is no longer part of the formal theory. It is therefore difficult to determine when it is advisable to use these strategies, and when a particular inference should be abandoned due to lack of statistical support or an unacceptable risk of error.

One can certainly construct theories of action using the inference rules of one's favorite permissive approach, as well as statistics and cost functions which appear to make the appropriate cost/benefit tradeoff. Likewise, one can provide, for the same theory, a different set of statistics and a cost function for which the tradeoff is poorly made. But when the statistics and cost function are not part of the theory, then it is only our individual dispositions that determines whether or not we believe the tradeoff has been appropriately made. But, then, arguing that all *rational* agents with the same knowledge should likewise have drawn the same inferences is impossible to defend.

3.1.3 Knowledge Obsolescense

Suppose that agent A knows fact P at time T. Further, suppose that A knows of no acts occurring between time T and $T + \Delta$ that affect P. The permissive approaches will, in general, persist the belief in P through an inductive argument from T to $T + 1$, $T + 2$, ..., $T + \Delta$. For example, Weber [1989] describes the *battery paradox*, where the agent initially knows that it's car battery is charged, and persists this belief inductively in the absence of knowledge to the contrary. Thus, the agent believes its battery is *always* charged, contradicting our intuition that batteries eventually become discharged[3].

Accepting default conclusions as true, allows their use in subsequent inferences; in doing so, however, beliefs become eternally perpetuated. Conversely, if one prohibits the chaining of defaults (or no defaults, as in the conservative approaches), then one is only able to perform inference for at most a single moment into the future.

[3] Note that this is an instance of Kyburg's *lottery paradox* [1970], where the agent believes of each individual ticket in a large lottery that it will not win the lottery, and also that *some* ticket will win.

Instead, an agent's commitment to its beliefs should weaken as its knowledge obsolesces. This is *not* because we want the agent's weakening beliefs to proportionately correspond to a continuously changing physical parameter (such as a battery's charge). Rather, it is an acknowledgement of the inherent limitations of the agent's ability to know about the external world. Commitment should reflect epistemic state, since, due to the agent's limited perceptual and computational resources, it *cannot* reflect metaphysical state; too many relations can change between the times when an agent is able to attend to them.

Despite Hayes's claim that "it should be clear that the underlying problem is not peculiar to the 'logical language' approach to model-building" [Hayes, 1971], the problem of knowledge obsolescence does involve at least the *straightforward* logical language approach. Since, if we wish to establish some belief in a proposition that we cannot prove deductively, we are obliged to either push its value up to true, or down to false. An agent must believe either that the battery is still charged, or it is not charged. "I don't know" is anathema to a permissive-ist. And there is no middle ground with first-order logic; the continuum between truth and falsity is the event horizon into which knowledge irretrievably falls.

3.2 Constraints on Knowledge Encoding

Suppose that rather than reasoning about the persistence of a battery's charge, one wishes to reason about whether an object O known to be at location L is still at L after time Δ has elapsed, and no disconfirming actions are known to occur in this interval. Relevant to this reasoning task are the following factors.

1. The type of object that O is (contrast a $100 bill and a used pair of jockey shorts).

2. The type of location that L is (contrast a safe deposit box, and a subway station).

3. The agent's ability to know about what is occurring at or near L. That is, *would* the agent know about a contravening event *if* such an event occurred?

4. The physical relationship between O and other objects of which the agent might be unaware.

5. The length of Δ.

6. The rate at which other agents are likely to perceive, or know about O at L.

7. The goals of other agents (with respect to O).

Each of these aspects might have a profound influence on our belief in the given proposition. Altering each aspect in turn while holding the others fixed gives vastly differing intuitions about the persistence of O at L. For instance, suppose O is a $100 bill, the agent is nowhere near L, Δ is one year, and 10 other agents know about O being at L, all of whom would happily benefit from ownership of O. Whether L is a bank vault, or the sidewalk in front of Macy's will likely determine O's future.

If an agent actually knows nothing about the type of object or location involved, the length of elapsed time, etc., then under such severe lack of knowledge it is reasonable for the agent *not* to commit to the persistence of O at L, unlike the permissive approaches. To expect a solution to the persistence problem to yield consistently intuitive answers in the absence of this information is an exercise in futility[4]. We are biased toward particular outcomes (the gun remains loaded, the car starts, the dollar bill remains on the sidewalk) largely because of shared, but hidden knowledge about common situations. Unless this information is made explicit *in a useable fashion*, our agents will be handcuffed, consistently defeating our intuitions.

Unfortunately, the knowledge encoded in the permissive approaches has been limited to either ground facts ("A is on B at time T", "The battery is charged at time $T+1$"), or universally quantified truths ("All red blocks are large," "Two things can't be at the same place at the same time"). There appears no straightforward way to encode useful knowledge of the form "batteries eventually wear out", "A was on B *yesterday*", "other children tend to want to use large blue blocks about as frequently as I do", "adding anti-freeze to your car increases the likelihood that it will start in January in Rochester."

[4] This theme is taken up with considerably more energy by [Weber, 1989, Schubert, 1989, Loui, 1987a].

Assuming that such knowledge is encoded, we would still require some means for relating this knowledge to our commitment to the truth of *particular* assertions. For example, we will want to relate knowledge that "batteries wear out" and "anti-freeze helps cars start ... " to our commitment to the assertion that "my car will start when I engage the ignition at time T." I will attempt to demonstrate how this statistical knowledge can be encoded and related to the probabilities we attach to causal assertions following an examination of the conservative approach.

4 The Conservative Approach

The conservative approaches have arisen from two separate motivations: those concerned with abandoning domain-independent completeness assumptions in favor of monotonic "explanation-closure" axioms, and those concerned with the explicit representation of multi-agency and concurrent activity.

4.1 Explanation Closure

The first group [Schubert, 1989, Weber, 1989, Haas, 1987] takes a novel approach to the frame problem: rather than encoding that "P does not change when A occurs", they instead encode

> (3) "P changes between time T and $T + 1$, only when one of actions $A_1, \ldots, A_n$ occurs at T."

These axioms are "explanation-closure" axioms, since they completely account for the causes of change for particular fluents. If the agent knows that none of the A_i occurred at T, then from axiom 3, one can *deductively* infer the persistence of P. For instance, I might have an axiom saying that if I have money in my pocket at time T, and I do not have money in my pocket at $T + 1$, then I must have intiated a purchase act at time T. If I know that I did not purchase anything between T and $T + \Delta$, then I can deductively infer that money is still in my pocket at $T + \Delta$.

The appeal of these approaches is their simplicity, that they don't require an axiom for each fluent/action pair (as with frame axioms), and that the inferences they draw are purely deductive. Thus, they do not suffer from the permissive problems of drawing overly strong inferences. But their obvious limitations are severe, as noted by the researchers themselves [Schubert, 1989]:

> [With] external agencies of change ... both effect axioms and explanation closure axioms may be invalidated. For example ... if the money in my pocket may be arbitrarily lost or stolen, I cannot assert an axiom that its depletion requires an expenditure. ... It would not help to include loss and theft among the possible explanations for depletion of funds, since the occurrence of these events cannot be ruled out ... "

Unlike the permissive approaches, explanation closure does not depend upon completeness *assumptions*. The deductive soundness of the theory remains intact, regardless of the number of closure axioms that the agent is fortunate enough to possess. But in large, multi-agented worlds, such axioms will be rare; those that exist will be constantly under the threat that new causes for change will be discovered. In this sense, providing complete explanation-closure axioms bears a striking similarity with providing complete sets of antecedent conditions for laws of motion (the qualification problem[5]). With few closure axioms, however, this approach is emasculated, the agent being able to draw few conclusions about what relations persist in future states.

4.2 Multiple Agency and Concurrency

As has been noted, the existence of outside agency causes considerable difficulty for solving the frame problem. As Rich Pelavin aptly points out [Pelavin, 1988], in a situation-calculus language in which future states are determined by ***function*** application on state-denoting terms, there is no way to distinguish between "action A occurs at time T" and "action A occurs at time T ***and no other actions occur.***" Thus, the language itself is placing unnecessary and severe constraints on the theory.

To overcome this problem, there has been a considerable amount of research in developing representations that allow for the explicit representation of concurrent events and multi-agency [Allen, 1984, Pollack, 1985,

[5] As pointed out by Steve Whitehead, personal communication.

McDermott, 1982, Pelavin and Allen, 1986, Georgeff, 1987]. However, even in these representations, the frame problem persists. Being able to represent that events A_1 through A_n all occur at time T is not sufficient to infer that some fluent P has not changed truth value. As Pelavin points out, sufficiency is obtained only if

1. the agent knows all events that affect each proposition, and
2. the agent knows that no such events also occurred at time T.

And, as noted earlier, agents are rarely fortunate enough to possess such knowledge.

Equivalently, Georgeff [1987] formalizes this principle, using a situation-calculus notation extended to account for concurrent actions. His "law of persistence" states that P holds in a succeeding state only if every action which occurs is *independent* of P. These more expressive representations only highlight the problem that few inferences are possible without complete information.

4.3 Knowledge Obsolescence

The arguments regarding the inadequacy of the permissive approaches to deal effectively with knowledge obsolescence apply equally to the conservative approaches. Whereas the permissive approaches never allow relations to wither on the vine, under the conservative doctrine, all relations die aborning. If there exists an explanation closure axiom (or set of independence axioms) for a fluent, and all change causing events are known *not* to occur, then the persistence is inferrable. But more likely, no such closure axioms exist, or some causitive event cannot be definitively ruled out, in which case no belief regarding the fluent is allowed. In the absence of knowledge completeness, the conservative approaches are too weak for use in dynamic, multi-agented domains.

5 Statistical Inference

Both of the solution classes that have been explored founder on the issue of knowledge completeness, although in opposite ways. In the absence of completeness, the permissive approaches are led to incorrect conclusions, while the conservative approaches are able to draw few, if any, conclusions.

An alternative approach is to allow agents to associate probabilities with propositions indicating their commitment toward the truth of these propositions. This allows assertions to be believed with less than full acceptance, enabling the agent's commitment to degrade as its knowledge obsolesces.

Using probabilistic reasoning is not new in artificial intelligence, but has rarely been explored for action reasoning (the exceptions being [Pearl, 1988, Dean and Kanazawa, 1988]). The novelty of the approach to be presented is that probabilities are *not* taken as subjective degrees of belief, but are instead based upon the set of statistical assertions that the agent possesses. It can thus be viewed as an application of Kyburg's theory of statistical epistemology as described in [Kyburg, 1974, Kyburg, 1983a, Kyburg, 1983b]. A brief description of using this theory for action reasoning is given here, with a more detailed, formal version provided in both [Weber, 1989] and [Tenenberg and Weber, 1990].

5.1 Kyburg's probability theory

Kyburg's theory relies upon the existence of a set of statistics that the agent possesses in its *rational corpus*[6] regarding set proportions. For instance, the agent will have sentences of the form "the proportion of birds that fly is in the interval [.93, .95]." Note that both "birds" and "flyers" are taken as sets of objects, and that proportion is taken simply as *set* proportion, or measure. Suppose additionally the agent also knows "Tweety is a bird" (Tweety is an element of the set of birds), and Tweety is a random element of the set of birds, *as far as the agent knows* (of which more will be said below). Then the agent can assign a probability interval [.93, .95] to the assertion "Tweety flies." That is, Tweety is an element of a set, *the reference class*, from which the proportion of flyers is known, and this proportion can be taken as the probability that Tweety is a flier.

As an example of action reasoning using this framework, I will consider the problem of inferring that my mug is in the lounge (`mug-in-lounge`) at time $t + \Delta$, given knowledge that it is in the lounge at some earlier time

[6] The set of statements that the agent accepts as *practical certainties*, which will be treated as truths, including such things as axiomatic set theory and number theory, and statistics that have a reliable basis.

t. Applying Kyburg's theory to action reasoning requires some means for relating set proportions to the propositions of interest. Intuitively, this will be done by having frequency statements of the form "the proportion of *times* (moments, worlds, instants) when $P_1 \wedge \ldots \wedge P_n$ occurs (for some set of fluents P_i) in which the mug is still in the lounge after time Δ is in the interval [p,q]."

I will use a first-order language with set theory and arithmetic, and will take time points as a linearly ordered set isomorphic with the integers and fluents to be sets of time points at the object level (rather than as set denoting, as in situation-calculus). That is, the fluent **mug-in-lounge** is the set of worlds in which my mug is in the lounge. Expressing that my mug is in the lounge at time t can be done with a simple set membership assertion "$t \in$ **mug-in-lounge**". In such a case, we say that t is a **mug-in-lounge**-world.

Making predictions will involve reasoning about sets of the form

$$\{t : t + \Delta \in \texttt{mug-in-lounge}\},$$

that is, the set of worlds such that my mug is in the lounge after time Δ. Sets of this form will be abbreviated by $\texttt{mug-in-lounge}_\Delta$ (where Δ is an integer constant). Under this notation,

$$\{t : t \in \texttt{mug-in-lounge}\} = \texttt{mug-in-lounge}_0 = \texttt{mug-in-lounge}$$

Assertions about set proportions will be denoted using a binary function symbol "%". For example, "$\%(A, B)$" stands for "the proportion of B's that are A". To avoid confusion about the ordering of the arguments, a vertical bar will be used instead of a comma to indicate the similarity to conditional probabilities: $\%(A|B)$.

Suppose that an agent has only the following statistical assertions:

$$\%(\texttt{mug-in-lounge}_1|\texttt{mug-in-lounge}) \in [.9, .95], \tag{1}$$
$$t \in \texttt{mug-in-lounge}. \tag{2}$$

These state that the proportion of worlds in which my mug is in the lounge that are followed by worlds in which the mug is still in the lounge is in the interval [.9,.95], and in world t, the mug is in the lounge. One can thus associate a probability of [.9,.95] with the assertion $t \in \texttt{mug-in-lounge}_1$, or equivalently $t + 1 \in$ **mug-in-lounge**.

The above statistic functions like a frame axiom: if my mug is in the lounge at some time, then it will remain in the lounge at the next moment with some high probability. The difference between the probability interval of this statistic, and the interval $[1, 1]$, or certainty that the mug remains in the lounge, reflects the fact that the frequency with which actions occur in the world that result in the mug leaving the lounge occurs between 5% and 10% of the time. That is, although there might be many agents that can remove the mug, and many ways of doing so, the coarse statistic (1) above concisely summarizes their combined influence. The agent thus need not know about the activities of all agents, nor all laws of cause and effect. The agent need only know the proportion of times in which such influences defeat its intentions. These statistics are obtainable either through direct observation, or through reporting by other agents.

This is similar to a default approach, where the agent believes P, since it has no knowledge to the contrary; the agent does not need to have a complete theory of the world. In this default case, however, the agent's uncertainty is not reflected in a probability between 0 and 1, but only in the ability to completely abandon a belief when a proof to the contrary is obtained.

5.1.1 Competing Reference Classes

The complexity that arises in this framework is if the agent has conflicting statistical knowledge. For example, suppose in addition to assertions 1 and 2 above, the agent knows

$$\%(\texttt{mug-in-lounge}_1 | \texttt{mug-in-lounge} \cap \texttt{holding-mug} \cap \texttt{leave-lounge}) \in [.0, .02].$$

There are thus competing *reference classes*

$$(\texttt{mug-in-lounge} \text{ versus } \texttt{mug-in-lounge} \cap \texttt{holding-mug} \cap \texttt{leave-lounge})$$

with competing reference intervals ($[.9, .95]$ versus $[.0, .02]$) for the probability of the assertion $t + 1 \in \texttt{mug-in-lounge}$.

The problem of choosing between reference classes is addressed by Kyburg [Kyburg, 1983b]. I will not provide a complete explanation of his theory[7] but will mention two of the criteria he provides for choosing between competing classes. Given two candidate reference sets X and Y, if

[7] A full formal accounting can be found in [Kyburg, 1974, Kyburg, 1983b].

the reference interval of one nests inside the other, choose the strongest (nested) interval. If the intervals do not nest, and X is a subset of Y, choose X as the reference set. This last criterion directs us to choose the most *specific* reference class for which we have statistics, a familiar principle in defeasible reasoning systems [Touretzky, 1984, Bacchus, 1988, Loui, 1987b]. Thus, according to this principle, we would choose the reference class

$$\texttt{mug-in-lounge} \cap \texttt{holding-mug} \cap \texttt{leave-lounge}$$

over the class mug-in-lounge, and hence associate the interval [.0, .02] with the probability of $t+1 \in$ mug-in-lounge.

Another way of viewing this problem is to recognize that t is not a random element of the set mug-in-lounge, relative to what the agent knows, since there exists knowledge of a smaller set to which t belongs for which there is a different proportion of $\texttt{mug-in-lounge}_1$-worlds. This notion of randomness is epistemological, since it depends upon the agent's state of knowledge regarding the sets to which particular worlds belong.

It is not guaranteed that one can attach a probability to every assertion. There might be some such assertions for which there there is no way to reconcile the different statistics, for example, if the agent knows

$$\%(\texttt{mug-in-lounge}_1 | \texttt{mug-in-lounge}) = [.9, .95]$$

and

$$\%(\texttt{mug-in-lounge}_1 | \texttt{holding-mug}) = [.1, .2],$$

but not

$$\%(\texttt{mug-in-lounge}_1 | \texttt{mug-in-lounge} \cap \texttt{holding-mug}).$$

This is analogous to having competing default rules[8] of the form

> "If my mug is in the lounge at time t then defeasibly it will be there at $t+1$"

(because of inertia) and

[8] In terms of Nute's [1986] or Loui's [1987b] default logic, but similarly for other nonmonotonic approaches.

> "If I am holding my mug at time t, then defeasibly it is not in the lounge at time $t + 1$"

(since I am rarely in the lounge), without having a specific default to cover the situation in which I am holding the mug while being in the lounge.

The advantage of using the statistical approach is that beliefs are related to measurements of the external world, compiled as frequency assertions. The agent is thus able to refine its knowledge by tracking particular relations in the world, for example, by seeking to determine a value for

$$\%(\texttt{mug-in-lounge}_1 | \texttt{mug-in-lounge} \cap \texttt{holding-mug}).$$

Kyburg's contribution is summarized nicely by Loui [Loui, 1987b]:

> Here is the achievement. Distinguish probability assertions,
>
> $$Prob(\text{"}x \in Z\text{"}) = [p, q]$$
>
> from statements about specific frequencies in classes,
>
> $$\text{"}\%(Y, Z) = [p, q].\text{"}$$
>
> Then provide rules for selecting among such frequency statements in order to determine probability.

The contribution of the present paper is to cast action reasoning in Kyburg's theory, and to demonstrate that the frame problem reduces to the problem of choosing the appropriate reference class. For each assertion whose truth value we would like to know, (e.g., whether t is a `mug-in-lounge`-world), we obtain a probability by finding a dominating reference class from among our statistical knowledge.

5.2 Yale Shooting Problem

One of the standard problems used to test solutions to the frame problem is the Yale Shooting Problem. Despite the fact that this problem has no correct answer, in the sense that an instance of the *16-puzzle* might have a solution, it nonetheless serves as a common foil against which to compare competing theories and intuitions. It will be helpful to describe how the statistical approach can be applied to this problem.

At some starting time, t_0, Fred is alive (`alive`), and a gun is loaded[9] (`loaded`). After some time Δ has elapsed, the gun is fired at Fred. Fred dies when fired upon with a loaded gun. The question for the theory to answer, is whether Fred is alive or not after time $t + \Delta$.

If the gun has not been unloaded between t and $t + \Delta$, then the answer is that Fred dies. Unfortunately, we have no explicit knowledge as to what events occur after t, other than the firing of the gun. It is important that we separate out the issues of conversational implicature here, since, from the story-telling point of view, it is reasonable to assume that I would have mentioned if someone had unloaded the gun.

This problem clearly motivates the use of the conservative approach of Schubert [1989] and Weber [1989], where we have axioms stating that guns get unloaded only through an unload act. When the reasoning agent is the same agent that is holding the gun and would know about all unloadings, we can deductively infer that Fred dies.

Kautz [1986] and Shoham [1986] provide solutions whereby we prefer interpretations in which *all* propositions persist as long as possible forward in time. Fred thus dies, since this interpretation allows both `alive` and `loaded` to persist right up until the gun is fired.

Lifschitz [1987] and Haugh [1987] present solutions preferring interpretations that minimize causitive acts. Thus, Fred dies again, since we know of no act occurring that causes the gun to be unloaded.

However, all of these solutions strain our intuitions when we consider increasingly large Δ's ([Hanks and McDermott, 1986]):

> If several years had elapsed between the WAIT and SHOT, for example, it would be reasonable to assume that the gun was no longer loaded.

In fact, for sufficiently large Δ, it would be *unreasonable* to conclude that the gun remained loaded. And this intuition seems to be based upon an implicit assumption that unloaded acts occur with some relative frequency, so that the longer we wait, the likelier it is that one such act will occur.

This can be explicitly encoded in the statistical framework. Note, however, that the probabilities obtained will rely upon the particular statistics

[9] For simplicity, I will dispense with the loading act originally described as part of the problem [Hanks and McDermott, 1986], and assume the gun has been successfully loaded.

in the knowledge base. Notationally, I will take t_0 as the initial time point, and t_i as the i^{th} subsequent time point. We first have that Fred is alive, the gun is loaded, and a wait occurs, all at time t_0:

$$t_0 \in \mathtt{alive} \cap \mathtt{loaded} \cap \mathtt{wait}.$$

For simplicity, I will initially assume that the gun is fired at time t_1 (rather than at some later time):

$$t_1 \in \mathtt{shoot}.$$

We would like to obtain a probability for the assertion "$t_2 \in \mathtt{alive}$." That is, is Fred alive after a WAIT act and a SHOOT?

There is a low likelihood that Fred is alive if he has just been shot with a loaded gun:

$$\%(\mathtt{alive}_1 | \mathtt{shoot} \cap \mathtt{alive} \cap \mathtt{loaded}) \in [0.1, 0.3] \qquad (4)$$

It is not difficult to encode that guns tend to stay loaded, and that Fred tends to remain alive from one moment to the next:

$$\%(\mathtt{loaded}_1 | \mathtt{loaded}) \in [0.9, 0.95] \qquad (5)$$
$$\%(\mathtt{alive}_1 | \mathtt{alive}) \in [0.99, 1.0] \qquad (6)$$

This approach appears to require explicit statistics for the likelihood of persistence, for each separate proposition, rather than a general inference mechanism (such as circumscription) that persists the entire set of propositions. Although this looks suspiciously like providing an entire set of frame axioms, there are significant differences. First, the statistics can be based upon observations of the reasoning agent. Second, the statistics can be sensitive to context – that is, there can be different reference sets associated with the same probability assertion. For instance, the statistic

$$\%(\mathtt{loaded}_1 | \mathtt{loaded} \cap \mathtt{unload}) \in [0.0, 0.05]$$

specifies a more finely grained context than the one above for predicting whether the gun is loaded. And third, statistics can be provided for broad reference classes, using the subset relation. For instance, suppose that our language has sets denoting the fact that other people are alive (`aliveBill`, `aliveKaren`, `aliveJake`). Then an abstract alive set (`absAlive`) can be defined which is the union of the specific alive propositions, and a general statistic can be obtained in terms of this abstract proposition:

$$\texttt{absAlive} = \texttt{alive} \cup \texttt{aliveBill} \cup \texttt{aliveKaren} \cup \texttt{aliveJake}$$
$$\%(\texttt{absAlive}_1 | \texttt{absAlive}) \in [0.95, 0.98] \qquad (7)$$

This general statistic reflects the likelihood that some random individual (among Fred, Karen, Jake, and Bill) will be alive from one moment to the next. It can be used, in the absence of other dominating statistics, to associate a probability with any particular individual persisting to be alive at a particular time.

Further, these statistics might be relative to larger intervals of time, rather than simply from moment to moment. Although the temporal model in the presented language is weak, one can still specify truth at some time during an interval using set union. First, the symbol $\mathrm{f}_{i,j}$ is defined as follows, where f stands for some set of time points (such as `alive`, or `loaded`), and i, j are integers, with $i < j$:

$$\mathrm{f}_{i,j} = \mathrm{f}_i \cup \mathrm{f}_{i+1} \cup \ldots \cup \mathrm{f}_{j-1} \cup \mathrm{f}_j.$$

Thus, $\texttt{alive}_{1,5}$ is the set of time points in which Fred is alive in at least one of the next five time points. Thus, we can now encode the statistic for the proportion of time points in which some agent is alive, given that it was alive *sometime* (we may not be sure precisely when) in the last 10,000 time points:

$$\%(\texttt{absAlive}_{10001} | \texttt{absAlive}_{0,10000}) \in [0.8, 0.9]$$

Note that were an agent to compile this statistic, it would *not* require that Fred, Bill, etc., all be monitored at each time instant. Instead, samples of their aliveness taken at only a few points over an extended time interval are all folded together. Thus, this very general statistic, taken over a very large class of propositions, argues strongly against the requirement of specific persistence statistics for every proposition.

Let us suppose, for simplicity, that we have only the statistics 5 and 6. We can thus derive the following probabilities:

$$Prob(\text{“}t_1 \in \texttt{loaded}\text{''}) \in [0.9, 0.95]$$
$$Prob(\text{“}t_1 \in \texttt{alive}\text{''}) \in [0.99, 0.1]$$

Thus, there is a high likelihood that at time t_1, Fred is still alive, and the gun is still loaded. As stated earlier, the inference task is to associate a probability with Fred's being alive at time t_2. Unfortunately, we have no appropriate statistics upon which to base this probability, since using statistic 4 requires that we be *certain* that both Fred is alive and the gun is loaded, and we have only high probability of these assertions. We are thus not able, under the given statistics, to associate any probability with the assertion "$t_2 \in$ `alive`."

One possible approach to arriving at a probability for this assertion was first was taken up by Dean and Kanazawa [1988], where, associated with each proposition (such as `alive`), is a single probability function, usually exponentially decreasing with time, indicating the likelihood of that proposition remaining true. Actions are treated as special propositions for which particular action instances might have a probability function associated with this action occuring at a *particular* time. A new probability function can then be derived for a proposition, by convolving its associated probability function with the probability functions of any affecting actions. Thus, the probability function for `alive` would be combined with the probability functions for the occurence during a particular interval of both an `unload` (which may be unlikely under our scenrio), and a `shoot` (which we take as certainly occuring at time t_1). In this fashion, Dean and Kanazawa obtain a new probability function for `alive` *conditioned upon the occurence of the shoot,* and are thus able to arrive at a probability for "$t_2 \in$ `alive`."

Although Dean and Kanazawa's approach is in the same spirit as the current one described, and has some elegant features, it suffers from the problems of unmotivated probability functions, allowing conditioning only on action occurence, and conflating probabilities with statistics. Given the continuous probability functions that they propose, it will likely be difficult and costly for agents to obtain the statistics necessary to assure reliability. Further, if there is no *a priori* knowledge of the occurence of an `unload`, then an unjustified randomness assumption is made that t_1 is a random time point with respect to the occurence of unloadings, even though the agent might know otherwise–such as that it is holding the gun, and there are no other agents. This assumption is required because probability curves for a proposition are not sensitive to context, except what is provided by the occurence of actions. That is, these probabilities cannot be conditioned upon a broad range of other knowledge that the agent might possess. If such context-sensitive probabilities were allowed, then Dean and Kanazawa would be obliged to provide a mechanism for adjudicating between the competing probabilities.

These drawbacks arise largely because probabilities and statistics are conflated; it is not clear how these probabilities might be derived from a set of observed statistics. The advantage of the approach I have presented is that, under Kyburg's theory, probabilities over instances (of set membership, such as "$t_2 \in \texttt{alive}$") are distinguished from statistics, which are proportions over sets of instances (such as $\%(\texttt{alive}_1|\texttt{alive}) \in [0.99, 1.0]$). Probability is thus defined relative to a body of knowledge, including statistical knowledge, with the central problem being that of obtaining the dominating reference class among competing statistics.

Statistics 5 and 6 above allow us to derive a probability close to 1 that at time t_1, the gun remains loaded, and Fred is alive. But at what point are these probabilities sufficiently close to truth that they should be added to the rational corpus, making them available for further inference? And should this threshhold be held fixed, so that the reasoning agent is unable to change it under reasoning tasks having different criticalities and risks? If so, then under a sufficiently low acceptance threshhold, statistic 4 can be used to conclude that Fred dies at t_2, since 4 has a more specific reference class than 6. This then becomes a statistical translation of Loui's defeasible logic approach to the Yale Shooting Problem [Loui, 1987a]. But one of the significant advantages of the statistical approach is the flexibility of having access to the underlying numerical values. It is difficult to imagine domains in which the tradeoffs are so predictable that we would want a fixed threshhold. At the very least, it seems premature to exclude the possibility of change, to make changing the level of acceptance dependent upon human intervention.

This does not rule out that under different statistics, we can derive a high probability that Fred dies. One such statistic would say, "The frequency with which Fred dies[10] is very high, given that a few moments prior, the gun is loaded, and then a shoot takes place." This is encoded as

$$\%(\texttt{alive}_2|\texttt{shoot}_1 \cap \texttt{alive} \cap \texttt{loaded}) \in [0.12, 0.2]. \qquad (8)$$

t_0 is an element of the reference set of this statistic, and hence we can derive a probability for Fred being alive at time t_2.

[10] Or, as we stated earlier, we might phrase this in terms of the frequency with which people in general die, since Fred can only die once, presumably.

And as before, we can consider statistics over larger temporal intervals, so that if our WAIT period is longer than a few moments, we can derive different probabilities. For instance, we might have that "The frequency with which Fred dies is moderate, given that just prior the gun is fired, and sometime in the last 100 moments, the gun is loaded, and Fred is alive,"

$$\%(\mathtt{alive}_{101}|\mathtt{shoot}_{100} \cap \mathtt{alive}_{0,100} \cap \mathtt{loaded}_{0,100}) \in [0.4, 0.6].$$

Or, "The frequency with which Fred dies is low, given that just prior the gun is fired, and sometime in the last 10,000 moments, the gun is loaded, and Fred is alive,"

$$\%(\mathtt{alive}_{10001}|\mathtt{shoot}_{10000} \cap \mathtt{alive}_{0,10000} \cap \mathtt{loaded}_{0,10000}) \in [0.8, 0.95].$$

We can see, then, that with an impoverished set of statistics, there is little that one can conclude about the persistence of Fred's life. But that with a richer set, one can draw various conclusions about the likelihood of Fred dying, which takes intr account the length of the WAIT.

5.3 Implications of the Statistical Approach

I know that my mug is in the lounge at time t. Is it there after I open the refriferator? Lifschitz, a prototypical permissive-ist says that refrigerator openings do not cause mugs to change locations, and to assume nothing else has occurred which might move the mug. Schubert, a conservative-ist says that mugs move only when a pickup occurs, and by representational constraints, the refrigerator door is the *only* action that occurred, anywhere, at time t[11]. The statistical approach says that in the absence of mug pickups, mugs change location infrequently. Rather than assuming that no unknown mug pickups occur, I instead weaken my commitment to "$t + 1 \in \mathtt{mug\text{-}in\text{-}lounge}$" by the frequency with which mug pickups have occurred in the past.

There are several implications of the statistical approach that I will briefly explore.

[11] Schubert does allow for more than one action to occur, but all such actions are known by the agent.

5.3.1 Incremental Prediction

In the traditional approaches examined earlier, the paradigm is one in which a number of propositions are known by the agent, and knowledge is propogated forward incrementally a moment at a time. If the truth value of a fluent is not known with regard to a particular time, then a truth value for this fluent regarding subsequent times is likewise not known. For example, if I do not know whether my mug is in the lounge at time t, then I will likewise know nothing about my mug being in the loung at time $t+1$, (and by induction. *all* future times[12]). Knowledge is therefore carefully preserved from one moment to the next, since once it is lost, this knowledge is irretrievable.

The statistical approach provides much greater flexibility. We accord full belief to the assertion that "my mug was in the lounge yesterday at noon," since that is when I last saw it, but not to the assertion "the mug is in the lounge *now*." The relevant statistics are of the form "the frequency with which mugs change location from one day to the next is in the interval $[x, y]$." We do not require statistics that chart the moment to moment persistence of relations in the world; for relations that change infrequently, coarse statistics ("80% of the stacked blocks will remain as I have configured them one hour from now, "95% of the buildings in this neighborhood will be standing in 10 years") are more economical. Further, the coarser the statistic, the greater the cardinality of the associated reference set, and the greater number of worlds in which this statistic can serve as a basis for a probability statement. Thus, coarse statistics are more widely applicable. In addition, these coarse statistics appear much easier for an autonomous agent to acquire, since the agent is able to monitor only a limited part of its environment from one moment to the next. In fact, the agent can use these statistics to determine those highly dynamic relations which require careful, frequent monitoring.

Knowledge is allowed to obsolesce, but not necessarily in a smoothly continuous fashion, as is the case with [Dean and Kanazawa, 1988]. Rather, the decay function will have large plateuas, and deep drops, perhaps occasional bumps, depending upon the rate at which the agent measures the particular relations.

This style of representation, then, gives rise to a lazier method for determining beliefs: rather than propogating each proposition in the agent's

[12] Assuming that we do not have any knowledge gathering actions, as in [Feldman and Sproull, 1977].

knowledge base at each moment, one reasons about propositions only when needed. Thus, if I know my mug is in the lounge at noon yesterday, I need not concern myself with the location of my mug until such time as I need the mug. I am thus saved the computational expense of propogating irrelevant beliefs.

5.3.2 Independence

Encoding that "most of the world carries on in just the same way that ... it would have done" [Hayes, 1987] had I *not* performed action A, is equivalent to saying that some relations are *independent* of A. Traditional approaches based on completeness assumptions are, by definition, inadequate for encoding independence. But the present approach provides a precise, formal account of independence in terms of standard *statistical* independence. That the persistence of `mug-in-lounge` is independent of what I had for breakfast, or whether it is raining outside, is implicitly encoded whenever the following hold:

$$\begin{aligned}&\%(\texttt{mug-in-lounge}_1|\texttt{mug-in-lounge} \cap \texttt{omelet}) = \\ &\quad \%(\texttt{mug-in-lounge}_1|\texttt{mug-in-lounge}) \\ &\%(\texttt{mug-in-lounge}_1|\texttt{mug-in-lounge} \cap \texttt{raining}) = \\ &\quad \%(\texttt{mug-in-lounge}_1|\texttt{mug-in-lounge}).\end{aligned}$$

5.3.3 The Qualification Problem

The qualification problem is that of expressing sufficient preconditions guaranteeing that an effect will hold as a result of some action. For instance, if I wish to remove my mug from the lounge, then a *Remove* act is guaranteed to have the intended effect (of having the mug out of the lounge) only if the door to the lounge is not locked from the outside, the mug is not glued to the coffee table, the mug is not sitting on a button which will cause it to explode when lifted, etc. As with the frame problem, one does not want actually to explicitly mention all of the possible defeating conditions; the more remote and ludicrous the defeater (the mug is filled with matter from a neutron star), the less inclined one is to consider and axiomitize it.

Although it is beyond the scope of this paper to engage in a detailed discussion of both the statistical and previous approaches to the qualification problem, a brief presentation should provide an indication of how the qualification yields to the statistical approach. A full discussion can be found in [Tenenberg and Weber, 1990].

The central intuition of the current approach is that a statistic involving a particular reference set

$$(\text{e.g. } \%(\texttt{mug-in-lounge}_1|\texttt{mug-in-lounge}) \in [.9, .95])$$

summarizes the frequency with which all defeaters (exeptions) occur, *without the need to either consider or even know what these defeaters are*[13].

Suppose one could specify a reference set S of the form $S = P_1 \cap P_2 \ldots P_n$ for some sufficiently large n such that there exists a statistic

$$\%(\texttt{mug-in-lounge}_1|S) \in [1.0, 1.0]).$$

If world i is an S-world, it is guaranteed that $i+1$ is a `mug-in-lounge`-world. Either no such finite S exists, or its description (in terms of the number of intersected sets P_i) is impractically large. In either case, a reasoning agent will be obliged to reason with smaller reference class descriptions, containing only a few of the P_i. Ideally, one would like to disregard those P_i which change the reference interval only slightly, either because they occur so infrequently (e.g., mugs glued to the coffee table), or they exert relatively little influence (e.g., the lounge door is locked from the outside, but there exists another exit to the room). One is thus trading off the accuracy of more finely grained statistics against the cost of maintaining, and reasoning with these statistics. One would prefer the economy of short descriptions while sacrificing only a small amount of certainty. Rather than having the axiom writer make this a fixed, a priori tradeoff, as with previous approaches to the qualification problem [McCarthy, 1977, Lifschitz, 1987, Ginsberg, 1987], access to the underlying statistics allows the reasoning agent to make this tradeoff on-line, adjusting it when necessary to match the criticality of the reasoning task.

[13] In fairness, whether defeaters need to be considered depends upon whether the agent possesses any statistics regarding defeaters (mugs glued to coffee tables, locked lounge doors), and how the agent is computing reference classes. Under current implementations of Kyburg's theory [Loui, 1987b, Weber, 1989], defeaters might be considered if such statistics exist for them.

5.4 Computing Reference Classes

Abandoning the completeness assumptions does not come without costs. These costs will likely be related to the computational requirements of determining the appropriate reference class for the probability assertions of interest. There is, however, considerable cause for optimism.

The first of these causes pertains to the comments about coarse statistics in the previous sections. An agent that is responsible for managing a statistical knowledge base will be inherently limited in the quantity of statistics that it can collect, store, and access. Recall, however, that our theory does not require the agent to have a precise probability function over all of the sentences in its language. Thus, the computational requirements might not be as great as is often suspected when using probabilities. What remains a fruitful area of future investigation, however, is to explore strategies for explicitly reasoning about the tradeoffs in collecting and keeping knowledge at different levels of generality.

Secondly, there have been several serious attempts at automating Kyburg's probability theory. Loui [1987b] was able to achieve good results for a restricted language. Kyburg and Murtezaoglu have recently implemented a version that uses several heuristics to prune the set of candidate reference classes. And Weber [1989], offers an implementation for computing reference classes that provides an *anytime algorithm* (one that provides an approximation to the answer when interrupted at any time during its execution) by considering coarse-grained reference classes before more specific reference classes.

6 Conclusion

The frame problem involves formal attempts to capture inferences that an agent might make concerning those states of the world that it believes continue to remain unchanged after some period of time has expired. Previous solutions have fallen into two different classes, which I have termed the permissive and conservative approaches. Unfortunately, both classes rely crucially upon assumptions of knowledge completeness. In complex, multi-agented worlds, such assumptions are untenable.

Alternatively, I have proposed an approach that associates statistically founded probabilities with temporally scoped assertions. To suggest the use of statistical knowledge for action reasoning is not to trivialize the concomitant technical problems, in particular, computing reference classes, and trading off the management of statistics at different levels of granularity. However, there are several significant benefits to this approach not found previously:

1. predictions about the future do not depend upon knowing the truth value of particular fluents at the previous moment. In other words, moment to moment persistence is not the only type of reasoning available to the agent,

2. coarse statistical information is useful,

3. knowledge can decay as time passes,

4. predictions about the future do not depend on specific known events, just that certain types of events occur with certain frequency.

5. commitment to beliefs are related to the agent's statistical beliefs, and hence suggests a means for the agent to acquire its *own* evidence for accepting assertions about the world.

6. "A does not affect P" can be represented by *statistical* independence, which encodes the fact that "most of the world carries on in just the same way that it did before, or would have done" [Hayes, 1987].

7. this framework provides a uniform approach to the frame and qualification problem

8. by not relinquishing the underlying numerical values, decisions can be made on-line regarding what is an acceptable level of risk.

These benefits indicate the statistical approach's broad applicability, and argues strongly for it's receipt of serious consideration as both a theoretical and a practical model for reasoning about action and prediction.

7 Acknowledgements

I am indebted to several people who have helped make my musings and explorations surprising and joyous. Thanks to Brian Haugh, Lenhart Schubert, and Bulent Murtezaoglu, for their insightful comments and criticisms. To Henry Kyburg, for his sagacious knowledge and wonderful teachings, for his nurturing of my curiosity, and continuing patience with the thickness of my cranium which seems to leave so little room for wet, grey matter, I will be always in his debt. And to Jay Weber, who I wish were back in Rochester, whose ideas on the frame problem shaped so many of my own, I shudder to think of what this work might be like had he not patiently persisted in pointing out all of my multitudinous misconceptions.

References

[Allen, 1984] James F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154, July 1984.

[Bacchus, 1988] Fahiem Bacchus. *Representing and Reasoning with Probabilistic Knowledge*. PhD thesis, University of Alberta, Fall 1988.

[Dean and Kanazawa, 1988] Thomas Dean and Keiji Kanazawa. Probabilistic temporal reasoning. In *Proceedings of AAAI-88*, August 1988.

[Feldman and Sproull, 1977] Jerry Feldman and Robert Sproull. Decision theory and artificial intelligence ii: The hungry monkey. *Cognitive Science*, 1, 1977.

[Fikes and Nilsson, 1971] Richard E. Fikes and Nils J. Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2:198–208, 1971.

[Georgeff, 1987] Michael Georgeff. Actions, processes, and causality. In *Proceedings of the 1986 Workshop on Reasoning about Actions and Plans*, 1987.

[Ginsberg, 1987] Matthew Ginsberg. Possible worlds planning. In *Proceedings of the 1986 Workshop on Reasoning about Actions and Plans*, 1987.

[Green, 1969] Cordell Green. Application of theorem proving to problem solving. In *Proceedings of IJCAI-69*, pages 219–239, 1969.

[Haas, 1987] Andrew Haas. The case for domain-specific frame axioms. In Frank M. Brown, editor, *The Frame Problem in Artificial Intelligence: Proceedings of the 1987 Workshop*. Morgan Kaufman, April 1987.

[Hanks and McDermott, 1986] Steve Hanks and Drew McDermott. Default reasoning, nonmonotonic logics, and the frame problem. In *Proceedings of AAAI-86*, August 1986.

[Haugh, 1987] Brian Haugh. Simple causal minimizations for temporal persistence and projection. In *Proceedings of IJCAI-87*, pages 218–223, June 1987.

[Hayes, 1971] Patrick Hayes. A logic of actions. In *Principles for Designing Intelligent Robots*, pages 495–519. Metamathematics Unit, University of Edinburgh, 1971.

[Hayes, 1981] Patrick J. Hayes. The frame problem and related problems in artificial intelligence. In Bonnie Lynn Webber and Nils J. Nilsson, editors, *Readings in Artificial Intelligence*, pages 223–230. Tioga, 1981.

[Hayes, 1987] Patrick Hayes. What the frame problem is and isn't. In Pylyshyn, editor, *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, pages 123–138. Ablex Publishing Company, Norwood, New Jersey, 1987.

[Hewitt, 1972] Carl Hewitt. *Description and Theoretical Analysis (using Schemata) of PLANNER: A Language for Proving Theorems and Manipulating Models in a Robot.* PhD thesis, MIT, 1972.

[Kautz, 1986] Henry A. Kautz. The logic of persistence. In *Proceedings of AAAI-86*, August 1986.

[Kowalski, 1979] Robert Kowalski. *Logic for Problem Solving.* Elsevier North Holland, 1979.

[Kyburg, 1970] Henry E. Kyburg, Jr. Conjunctivitis. In M. Swain, editor, *Induction, Acceptance and Rational Belief.* Reidel, 1970.

[Kyburg, 1974] Henry E. Kyburg, Jr. *The Logical Foundations of Statistical Inference*. Reidel, 1974.

[Kyburg, 1983a] Henry E. Kyburg, Jr. Rational belief. Technical Report Cognitive Science Technical Report 3, University of Rochester, 1983.

[Kyburg, 1983b] Henry E. Kyburg, Jr. The reference class. *Philosophy of Science*, 50:374–397, 1983.

[Lifschitz, 1987] Vladimir Lifschitz. Formal theories of action: Preliminary report. In Frank M. Brown, editor, *The Frame Problem in Artificial Intelligence: Proceedings of the 1987 Workshop*. Morgan Kaufman, April 1987.

[Loui, 1987a] Ronald P. Loui. Response to hanks and mcdermott: Temporal evolution of beliefs and beliefs about temporal evolution. *Cognitive Science*, 11, 1987.

[Loui, 1987b] Ronald P. Loui. *Theory and Computation of Uncertain Inference and Decision*. PhD thesis, University of Rochester Computer Science Department, September 1987.

[McCarthy and Hayes, 1981] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In Bonnie Lynn Webber and Nils J. Nilsson, editors, *Readings in Artificial Intelligence*, pages 431–450. Tioga, 1981.

[McCarthy, 1977] John McCarthy. Epistemological problems of artificial intelligence. In *Proceedings of IJCAI-77*, pages 1034–1044, Cambridge, MA, 1977.

[McDermott, 1982] Drew McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6:101–155, 1982.

[Nute, 1986] Donald Nute. Ldr: A logic for defeasible reasoning. Technical Report ACMC Research Report 01–0013, Advanced Computational Methods Center, University of Georgia, Athens, Georgia, 1986.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, Inc., 1988.

[Pednault, 1985] Edwin Pednault. Preliminary report on a theory of plan synthesis. Technical report, SRI International, August 1985.

[Pelavin and Allen, 1986] Richard Pelavin and James F. Allen. A logic for planning in temporally rich domains. *IEEE transactions*, August 1986.

[Pelavin, 1988] Richard Pelavin. *A Formal Approach to Planning with Concurrent Actions and External Events*. PhD thesis, University of Rochester, 1988.

[Pollack, 1985] Martha Pollack. *Generating Expert Answers Through Goal Inference*. PhD thesis, Computer Science Department, University of Pennsylvania, 1985.

[Reiter, 1980] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13, 1980.

[Schubert, 1989] Lenhart Schubert. Monotonic solution of the frame problem in the situation calculus. In Henry Kyburg and Ron Loui, editors, *Selected papers from the 1988 Society for Exact Philosophy Conference*, 1989.

[Shoham, 1986] Yoav Shoham. Chronological ignorance. In *Proceedings of AAAI-86*, August 1986.

[Tenenberg and Weber, 1990] Josh Tenenberg and Jay Weber. A statistical solution to the qualification problem and how it also solves the frame problem. Technical Report forthcoming, University of Rochester, Rochester, New York, 1990.

[Touretzky, 1984] David S. Touretzky. Implicit ordering of defaults in inheritance systems. In *Proceedings of AAAI-84*, pages 322–325, 1984.

[Weber, 1988] Jay Weber. A versatile approach to action reasoning. Technical report, University of Rochester, March 1988.

[Weber, 1989] Jay Weber. *Principles and Algorithms for Causal Reasoning with Uncertain Knowledge*. PhD thesis, Computer Science Department, University of Rochester, 1989.