Theory Preference Based on Persistence

Scott D. Goodwin
Randy G. Goebel

Technical Report CS-86-34
September 1986

Theory Preference Based on Persistence

Scott D. Goodwin
Randy G. Goebel

Logic Programming and Artificial Intelligence Group
Department of Computer Science
University of Waterloo

ABSTRACT

A recent paper by Hanks and McDermott calls into question the value of logic in AI. The paper describes how representing even simple default reasoning problems can give rise to multiple consistent yet conflicting solutions. The problem they describe is not due to any deficiency of the reasoning system, but is merely the result of a weak set of axioms. Since strengthening the axioms to eliminate unwanted models is a non-trivial problem (equivalent to the frame problem), our approach is to supplement the axioms with a preference criterion which restricts the models (just as strengthening would), but which is easier to specify. The preference criterion we propose is intended to reflect what McCarthy has called "...the common sense law of inertia." Our formalisation of this concept is based on a heuristic measure of *persistence*. We describe a planning framework in which a theory formation system uses frame default schemas to generate descriptions of situations. We show how the notion of persistence can be used to distinguish multiple competing situation descriptions and thereby determine whether the goal is predicted by the preferred situation description.

September 9, 1986

Theory Preference Based on Persistence

Scott D. Goodwin
Randy G. Goebel

Logic Programming and Artificial Intelligence Group
Department of Computer Science
University of Waterloo

1. Introduction

The advent of nonmonotonic reasoning systems provides the opportunity to explore consistency-based solutions to the frame problem. In particular, the conceptual framework underlying Theorist [Poole85a] not only supports such solutions, but also offers a foundation for their semantic analysis. Based on a philosophy inspired by Popper [Popper58], Theorist views reasoning as scientific theory formation (rather than as deduction). Reasoning in the Theorist framework involves building theories (henceforth *theory* means scientific theory — not logical theory) that explain a set of observations. A *theory*, consisting of instances drawn from a set of *possible hypotheses*, is said to *explain* a set of *observations* if the theory, together with the *facts*, logically implies the observations; it must also be consistent with the facts.

Although Theorist provides a framework in which to investigate solutions to the frame problem, there are difficulties. One problem is that of having multiple theories that explain the observations. Given that it is difficult, if not impossible, to determine a unique theory by strengthening the underlying set of axioms, it seems natural instead to accept the theories corresponding to a weak set of axioms as possible explanations and then to rank them according to some suitable criterion and thereby determine a preferred explanation. (For example, the Copernican view of the solar system is not the result of the complete axiomatisation of the properties of the universe, but rather it is a preferred explanation supported by a necessarily weak set of axioms.) The issue of theory preference has been considered for several domains [Poole85a, Poole85b, Jones85, Poole86]. Apparently, preference criteria are domain-dependent. For example, in domains with inheritance hierarchies, the explanation that uses the most specific knowledge is preferred; while in learning domains, the most general explanation is desired; and in diagnosis, the most probable, the most serious, or the most specific explanation may be sought.

The problem of multiple theories is also encountered in planning domains. Here we present a theory preference measure intended to reflect what McCarthy calls the "common sense law of inertia." [1] Our formalisation of this common sense concept is

[1] See [Lifschitz86b, p 408]

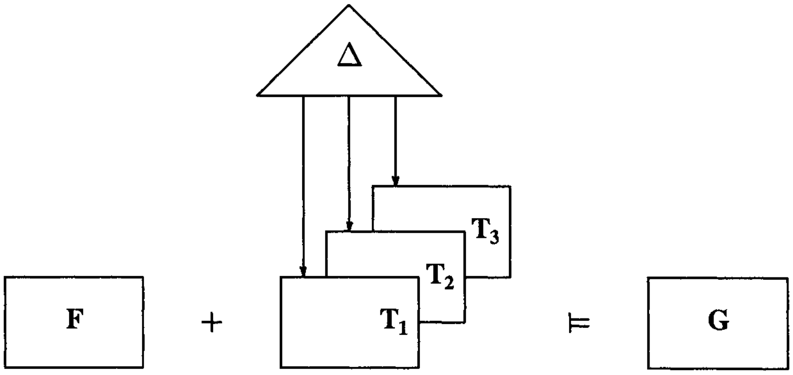
based on a theory formation framework in which some theories are *more persistent* than others. The proposed formalisation handles the difficulties described by Hanks & McDermott [Hanks85, Hanks86] and can be extended in various ways to enrich the strategic levels of automatic planning.

The next section provides a brief description of the planning representation scheme. The section is intended to enable the reader to understand the axiomatisations used in the examples. A more detailed description of the representation is described elsewhere [Goodwin86]. Section 3 defines the criteria that we expect our heuristic preference measure to satisfy. A semantic definition of persistence together with examples that show how alternative persistence measures can distinguish theories are provided. We show how the global maximisation of persistence is inadequate for theory preference and provide an alternative that incrementally maximises persistence. Next we indicate how accuracy may be traded off for computational savings. Finally, the heuristic nature of our preference measure is discussed. Section 4 mentions related work; of these, the work of Kautz is most closely related to ours. We give an example showing how information about relative weights of persistence can be used to distinguish solutions which are incomparable in Kautz's framework. Section 5 gives our conclusions, the most important of which is that theory formation together with theory preference form a framework which is simple and intuitive.

2. Planning in the Theorist Framework

Planning can be viewed as two processes: deductive question answering [Green81] and search. For example, in a simple forward planner, a tree of possible situations is searched to find one that satisfies the *goal description* (properties of the goal). Question answering is used to determine whether the goal description is satisfied in the current situation. When it is not satisfied, neighbour nodes of the current situation in the search tree can be generated by using question answering to determine which actions are possible, i.e., which actions have their preconditions satisfied in the current situation. This separation of planning into question answering and search allows the representation of planning domain knowledge to be considered independently of the planning search strategy.

A Theorist-based representation scheme, intended to form the basis of the question answering component of a planning system, has been designed [Goodwin86]. Planning problems are described (Fig. 1) using a combination of *facts* and *defaults* (possible hypotheses). In worlds with complete information, the initial situation is described solely by facts. The effects of an action partition relations describing a world into two groups. Relations that are known to be changed by the performance of the action form one group, and all other relations — those *presumed* to be unaffected by the performance of the action — form the other group. Laws of motion [Hayes71] describing the relations that are known to change are expressed as facts, while the laws of motion for the relations presumed invariant are expressed as defaults (as suggested by Reiter



Find $T_i \subseteq \Delta$ such that:
$F \cup T_i \models G$, and
$F \cup T_i$ is consistent
where Δ is a set of frame defaults
T_i is a set of instances of frame defaults
F is a set of facts describing the initial situation and laws of motion
G is a goal description (or its negation).

Figure 1. Planning in the Theorist Framework

[Reiter80, p 85]). These defaults correspond to frame axioms [Green81]. Collections of ground instances of these *frame defaults* form theories from which predictions can be made.

For some problems, it is desirable to treat some action effects and some initial conditions as defaults. For example, the normal effect of an action can be represented as a default. As well, the closed world assumption can be expressed with defaults. Here, however, we restrict the set of possible hypotheses to frame defaults. We do this so we can concentrate on the problem of defining a preference measure for planning without the need to consider possible interactions between various types of defaults.

Once a planning problem has been represented as described above, it is solved by finding a situation and a supporting theory that satisfies the goal description. In this theory formation framework, a situation is named by the sequence of actions from which it results and it is described by the facts together with a theory. Since there can be many theories describing a situation, the issue of theory preference arises. For example, an action is presumed applicable if its preconditions are predicted by the preferred theory. In this framework, a solution to a planning problem has two components: a sequence of actions (a plan) which achieves the goal, and a theory which predicts the goal description. Since it is not known in advance whether the goal description or its negation holds, they are treated as competing *predictions*. From the theories predicting the goal or its negation, one seeks a preferred theory from which the truth of the goal description is determined.

This is quite different from the treatment of symptoms in diagnosis, where symptoms are *observations* to be explained. There is, however, a close relationship between observations and predictions; both are consequences of a theory of the world. The difference between them is that observations (if accurate) are true in the world whereas predictions may or may not be true in the world. From this we see that the quality of a heuristic measure of preference depends on its ability to select a theory that makes predictions which are usually true in the world.

In addition to the problem of multiple theories arising in other domains, the planning domain also exhibits multiple *conflicting* theories. This is illustrated by the theories corresponding to the example in Figure 2.

F = {	The set of facts:
	Initial Situation:
loaded(0),	The gun is loaded and aimed at John
alive(0),	John is alive
rich(0),	He is rich
	Action: get_married
¬ rich(do(get_married,S)),	If John gets married he won't be rich
	Action: shoot
¬ alive(do(shoot,S)) ← loaded(S),	John dies when shot with a loaded gun
¬ loaded(do(shoot,S))}	After shooting, the gun is not loaded
Δ = {	The set of Frame Defaults:
[A,S] loaded(do(A,S)) ↔ loaded(S),	
[A,S] alive(do(A,S)) ↔ alive(S),	
[A,S] rich(do(A,S)) ↔ rich(S)}	

Figure 2. Example 1

In this example, two sets of statements are given. One set, **F**, describes the relations that are accepted as true in the world. Within this set, there are two kinds of axioms: those that describe the initial situation, and those that describe the changes caused by the performance of actions. A second set of statements, Δ, contains a frame default for each *primitive* [Fikes71] relation occurring in **F**. Collections of instances (the variables in the square brackets are to be instantiated) of these defaults form theories. An instance of a frame default asserts that the truth-value of the corresponding relation is preserved when performing the particular action in the particular situation. For example, $\mathbf{alive(do(get\_married,0))} \leftrightarrow \mathbf{alive(0)}$ means the truth-value of **alive** is unaffected when performing the action **get_married** in situation **0**. Note that instances of frame defaults assert the *equivalence* between properties in two adjoining situations. Equivalence is used instead of implication since, in addition to preserving positive information between situations, negative information is also preserved (i.e., $\mathbf{[p(do(a,s)) \leftarrow p(s)] \;\&\; [notp(do(a,s)) \leftarrow \neg p(s)] \equiv [p(do(a,s)) \leftrightarrow p(s)]}$).

In the above example, consider whether John will be alive after the actions **get_married** and **shoot**. There are 22 consistent theories that describe the invariance of relations over the path from the initial situation **0** to the situation **do(shoot,do(get_married,0))**, of which two are:

$$\begin{aligned}\mathbf{T_1} = \{&\mathbf{loaded(do(shoot,do(get\_married,0)))} \leftrightarrow \mathbf{loaded(do(get\_married,0))},\\ &\mathbf{alive(do(get\_married,0))} \leftrightarrow \mathbf{alive(0)},\\ &\mathbf{alive(do(shoot,do(get\_married,0)))} \leftrightarrow \mathbf{alive(do(get\_married,0))},\\ &\mathbf{rich(do(shoot,do(get\_married,0)))} \leftrightarrow \mathbf{rich(do(get\_married,0))}\} \text{ and}\end{aligned}$$

$$\begin{aligned}\mathbf{T_2} = \{&\mathbf{loaded(do(get\_married,0))} \leftrightarrow \mathbf{loaded(0)},\\ &\mathbf{alive(do(get\_married,0))} \leftrightarrow \mathbf{alive(0)},\\ &\mathbf{rich(do(shoot,do(get\_married,0)))} \leftrightarrow \mathbf{rich(do(get\_married,0))}\}.\end{aligned}$$

The statements in the theories are instances of frame defaults that record two different sets of assumptions about the action sequence **get_married** then **shoot**. Note that these two theories conflict since

$$\begin{aligned}&\mathbf{F} \cup \mathbf{T_1} \models \mathbf{alive(do(shoot,do(get\_married,0)))} \text{ while}\\ &\mathbf{F} \cup \mathbf{T_2} \models \neg\mathbf{alive(do(shoot,do(get\_married,0)))}.\end{aligned}$$

Of the remaining 20 theories,

three predict **alive(do(shoot,do(get_married,0)))**,
five predict $\neg$**alive(do(shoot,do(get_married,0)))**, and
12 make no prediction regarding **alive** or $\neg$**alive**.

In our intended model (i.e., the one that corresponds to our intuitions), however, $\neg$**alive(do(shoot,do(get_married,0)))** is true (Table 1).

At this point, one might wonder if the existence of multiple consistent yet conflicting theories indicates that our axiomatisation of the problem is incorrect. In their work, Hanks and McDermott [Hanks85] show that such multiple consistent conflicting solutions are inevitable and they conclude that nonmonotonic reasoning systems are inherently incapable of adequately representing even simple default reasoning problems. However, the problem is not due to any deficiency of the reasoning system; it is merely the result of a weak set of axioms. One might ask whether it is possible to restrict the set of consistent theories by strengthening the set of facts so that the new set of consistent theories all predict $\neg$**alive(do(shoot,do(get_married,0)))**. The answer, of course,

Situations:
1 ≡ do(get_married,0)
2 ≡ do(shoot,do(get_married,0))

Intended Model Features			
	0	1	2
loaded	T	T	F
alive	T	T	F
rich	T	F	F

T_1 Model Features			
	0	1	2
loaded	T	F	F
alive	T	T	T
rich	T	F	F

T_2 Model Features			
	0	1	2
loaded	T	T	F
alive	T	T	F
rich	T	F	F

Table 1. Essential Features of Models for Example 1.

is yes — frame axioms could be added to the set of facts — but we are trying to avoid the inherent problems of frame axioms. Therefore, we initially allow theories which disagree with our intended model and then use theory preference to discriminate between them.

In view of the possibility of having multiple and potentially conflicting theories, how can a selection be made between them? To answer this question, the concept of persistence must be defined — persistence will provide a way to compare theories.

3. Persistence

The notion of persistence is intended to reflect what McCarthy calls the "common sense law of inertia"; that is, when an action is performed, most things remain unchanged. When presented with competing theories, each making different predictions, we desire a heuristic that prefers the theory that corresponds to our intuition about persistence. The heuristic should simultaneously satisfy three criteria:

1) *accuracy* — it should select a theory which makes predictions that correspond to our expectations;

2) *sufficiency* — if the goal description (or its negation) is expected, then it should be predicted by the selected theory;

3) *resource conservatism* — it should select a theory with maximal obtainable *accuracy* for minimal computational effort.

In order to formalise this intuition, a semantic account of persistence is necessary. We will use standard Tarskian semantics; that is, the world is described in terms of individuals and relations on individuals. The domain of discourse contains three types of individuals: situations, actions, and ordinary objects. This ontology corresponds to that of Green's formulation II [Green81]. For the purpose of discussion, the language in which logical theories are expressed will be full first-order clausal logic; however, any language capable of expressing a contradiction is sufficient (cf. [Goebel85]).

Definition 1.

Let **R** be a relation in the domain with a vector of arguments $\bar{\mathbf{x}}$ as well as a situation argument **s**. We say that $\mathbf{R(\bar{x})}$ is a *propositional fluent* [McCarthy69] because the truth-value of the corresponding relation, $\mathbf{R(\bar{x},s)}$, varies with the situation. When we say $\mathbf{R(\bar{x})}$ is true in situation **s**, we mean $\mathbf{R(\bar{x},s)}$ is true.

Definition 2.

Let **do** be a function that maps actions and situations to situations. Thus, **do(a,s)** names the situation that results from doing action **a** in situation **s**.

Definition 3.

Let $\mathbf{s_n = do(a_{n-1},do(a_{n-2},\ldots,do(a_1,s_1))\ldots)}$ be a situation in the domain. The situations $\mathbf{s_1}$ to $\mathbf{s_n}$ determine a *path* which we write as $\mathbf{<s_1,s_n>}$. Furthermore, the *length of a path* $\mathbf{<s_1,s_n>}$ is defined to be **n−1**, one less than the number of situations on the path from $\mathbf{s_1}$ to $\mathbf{s_n}$. A *unit path* is a path of length one.

Definition 4.

Given a consistent theory **T** and a path $\mathbf{p = <s_1,s_n>}$:

a) A *primitive* propositional fluent $\mathbf{R(\bar{x})}$ is said to *persist* in **T** over the unit path $\mathbf{<s,do(a,s)>}$ if $\forall \mathbf{S} \in \mathbf{<s,do(a,s)>, T} \models \mathbf{R(\bar{x},S)}$ or $\forall \mathbf{S} \in \mathbf{<s,do(a,s)>, T} \models \neg\mathbf{R(\bar{x},S)}$;

b) The set of propositional fluents $\mathbf{P_T^i}$ which persist in **T** over unit path $\mathbf{<s_i,s_{i+1}>}$ is called a *persistence set*;

c) A domain-dependent ranking of persistence sets (denoted by $>^{\mathbf{p}}$) can be defined to reflect the relative likelihoods of each persistence set. Thus if two persistence sets differ on a highly persistent propositional fluent $\mathbf{R(\bar{x})}$, the persistence set which includes it would be higher ranked.

We will restrict our attention to problems where all propositional fluents are equally likely to persist. Under this assumption, persistence sets are ranked according to their cardinality. Hence $\mathbf{P_{T_1}^i} >^{\mathbf{p}} \mathbf{P_{T_2}^i}$ if $|\mathbf{P_{T_1}^i}| > |\mathbf{P_{T_2}^i}|$. Furthermore, we can define the *persistence* of a propositional fluent $\mathbf{R(\bar{x})}$ in **T** over path **p** to be equal to the number of unit paths contained in **p** over which $\mathbf{R(\bar{x})}$ persists in **T**. We can also define the *persistence* of **T** over the path **p** to be equal to the sum of the persistence of the propositional fluents in **T** over path **p**. Also note that persistence is defined in relation to primitive [Fikes71] propositional fluents. This ensures that the truth-value of a defined propositional fluent is preserved only when its associated primitives persist.

This semantic definition of persistence can be used to distinguish theories. For instance, the persistence for each propositional fluent of example 1 is given in Table 2. This table is computed by counting the unit paths over which each propositional fluent

	T_1	T_2
loaded	1	1
alive	2	1
rich	1	1
	4	3

Table 2. Persistence over the path <0,do(shoot,do(get_married,0))>

persists in the indicated theory (cf. Table 1), e.g., in theory $\mathbf{T_1}$ the truth-value of **loaded** is invariant only over the unit path **<do(get_married,0),do(shoot,do(get_married,0))>**, hence the entry '1' at the intersection of column $\mathbf{T_1}$ and row **loaded** in Table 2. The final row of the table shows the persistence of each theory (i.e., the sum of each column). The table indicates that the persistence of the propositional fluent **alive** is greater in $\mathbf{T_1}$ than in $\mathbf{T_2}$. It also shows that the persistence of $\mathbf{T_1}$ is greater than the persistence of $\mathbf{T_2}$. The persistence of theories over each unit path can also be used to distinguish theories (Fig. 3). This is computed as in Table 2, except persistence is split by unit path, e.g., for $\mathbf{T_1}$, **alive** is the only propositional fluent that persists over unit path **1** while **loaded**, **alive**, and **rich** each persist over unit path **2**. By comparing the above two methods of distinguishing theories, we hope to show that the second method, that is, the method which considers persistence incrementally rather than globally, is the preferable method. The observation that theories can be distinguished based on their persistence motivates the following definitions.

Definition 5.

A theory $\mathbf{T_1}$ is said to be *more persistent* than theory $\mathbf{T_2}$ over the path $<\mathbf{s_1},\mathbf{s_n}>$ if

a) $\exists \mathbf{j} \leq \mathbf{n},\ \mathbf{P}^{\mathbf{j}}_{\mathbf{T_1}} >^{\mathbf{p}} \mathbf{P}^{\mathbf{j}}_{\mathbf{T_2}}$

b) $\forall \mathbf{i} < \mathbf{j},\ \mathbf{P}^{\mathbf{i}}_{\mathbf{T_1}} =^{\mathbf{p}} \mathbf{P}^{\mathbf{i}}_{\mathbf{T_2}}$

where $\mathbf{P}^{\mathbf{i}}_{\mathbf{T}}$ is the persistence set for $\mathbf{T}$ over $<\mathbf{s_i},\mathbf{s_{i+1}}>$.

Definition 6.

A theory is said to be *maximally persistent* over a path if there does not exist a theory which is more persistent over that path.

We can use these definitions to examine Reiter's [Reiter80, p 85] formalisation of the STRIPS assumption. In Reiter's terms, a *default theory* $\Delta = (\mathbf{D},\mathbf{W})$ consists of defaults $\mathbf{D}$ and facts $\mathbf{W}$. The extensions of a default theory with

Unit Paths:

1 ≡ <0,do(get_married,0)>

2 ≡ <do(get_married,0),do(shoot,do(get_married,0))>

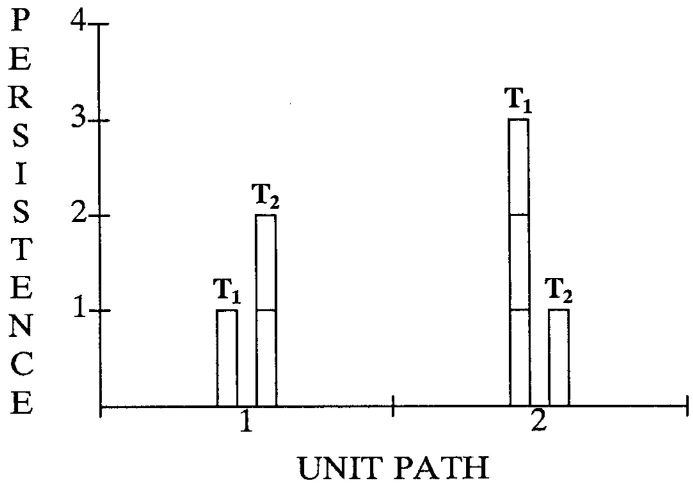


Figure 3. Persistence by unit path for Theories from Example 1

$$\mathbf{D} = \frac{\mathbf{R(S): M\ R(do(A,S))}}{\mathbf{R(do(A,S))}}$$

are the maximal consistent sets of consequences of **D** ∪ **W**. Such an extension corresponds to the consequences of a maximally persistent theory **T**. Returning to the example of Figure 2, of the 22 consistent theories, $\mathbf{T}_1$ is the maximally persistent theory (in this example, there is a unique maximally persistent theory—this is not always the case); but $\mathbf{T}_1$ predicts **alive(do(shoot,do(get_married,0)))**. This prediction is optimistic but unlikely. Therefore, at least in this case, maximal persistence is not the heuristic measure of preference we desire. To further support this claim, consider another example (Fig. 4).

For example 2, suppose that we are interested in whether **alive(do(jump,do(guzzle,0)))** is true. Again we can consider the maximally persistent theory to find out what it predicts. It turns out that, in this example, there are two maximally persistent theories.

F = {	The set of facts:
	Initial Situation:
defective(0),	The parachute is defective
alive(0),	John is alive
¬ happy(0),	He is not happy
	Action: guzzle
happy(do(guzzle,S)),	Guzzling beer makes John happy
	Action: jump
¬ alive(do(jump,S)) ← defective(S)}	Sky diving with the defective parachute is fatal
Δ = {	The set of Frame Defaults:
[A,S] defective(do(A,S)) ↔ defective(S),	
[A,S] alive(do(A,S)) ↔ alive(S),	
[A,S] happy(do(A,S)) ↔ happy(S)}	

Figure 4. Example 2

They are:

$$\begin{aligned} T_1 = \{&\mathbf{defective(do(guzzle,0))} \leftrightarrow \mathbf{defective(0)},\\ &\mathbf{defective(do(jump,do(guzzle,0)))} \leftrightarrow \mathbf{defective(do(guzzle,0))},\\ &\mathbf{alive(do(guzzle,0))} \leftrightarrow \mathbf{alive(0)},\\ &\mathbf{happy(do(jump,do(guzzle,0)))} \leftrightarrow \mathbf{happy(do(guzzle,0))}\} \text{ and}\end{aligned}$$

$$\begin{aligned} T_2 = \{&\mathbf{defective(do(jump,do(guzzle,0)))} \leftrightarrow \mathbf{defective(do(guzzle,0))},\\ &\mathbf{alive(do(guzzle,0))} \leftrightarrow \mathbf{alive(0)},\\ &\mathbf{alive(do(jump,do(guzzle,0)))} \leftrightarrow \mathbf{alive(do(guzzle,0))},\\ &\mathbf{happy(do(jump,do(guzzle,0)))} \leftrightarrow \mathbf{happy(do(guzzle,0))}\}.\end{aligned}$$

Each theory is consistent but makes a different prediction:

T_1 predicts **¬alive(do(jump,do(guzzle,0)))** while
T_2 predicts **alive(do(jump,do(guzzle,0)))**.

Because maximal persistence does not distinguish these conflicting solutions, this example further demonstrates that maximal persistence is not the measure that we are looking for.

Though maximal persistence fails to provide the desired preference heuristic, it leads to an important observation. Of the two theories above, only $\mathbf{T_1}$ corresponds to our intended model (cf. Table 3). The persistence (by unit path) of these theories is shown in Figure 5 ($\mathbf{T_3}$, $\mathbf{T_4}$, and $\mathbf{T_5}$ will be used later.) Notice that over the first unit path **<0,do(guzzle,0)>**, $\mathbf{T_1}$ has a persistence of 2 while $\mathbf{T_2}$ has a persistence of only one.

Situations:
$1 \equiv$ **do(guzzle,0)**
$2 \equiv$ **do(jump,do(guzzle,0))**

Intended Model Features			
	0	1	2
defective	T	T	T
alive	T	T	F
happy	F	T	T

T_1 Model Features			
	0	1	2
defective	T	T	T
alive	T	T	F
happy	F	T	T

T_2 Model Features			
	0	1	2
defective	T	F	F
alive	T	T	T
happy	F	T	T

T_3 Model Features			
	0	1	2
defective	T	T	-
alive	T	-	F
happy	F	T	-

T_4 Model Features			
	0	1	2
defective	T	F	-
alive	T	T	T
happy	F	T	-

T_5 Model Features			
	0	1	2
defective	T	T	-
alive	T	T	F
happy	F	T	-

Table 3. Essential Features of Models for Example 2.

Over the second unit path, the persistences are 2 and 3 respectively. $\mathbf{T_1}$, therefore, has more persistence in the earlier unit path. Similarly, in example 1 (fig. 2), the "better" theory $\mathbf{T_2}$ also had more persistence in the earlier unit path.

Since actions are performed in sequence, it does not make sense to globally maximise persistence over a path as implied by the maximal persistence criterion. Instead, persistence should be maximised step by step (in chronological [2] order — cf. [Hanks85, Hanks86, Shoham86a, Shoham86b]). This notion of step by step maximisation is formalised in the following definitions.

Definition 7.

A theory $\mathbf{T_1}$ is said to be *chronologically more persistent* than theory $\mathbf{T_2}$ over the path $<s_1,s_n>$ if there exists a sub-path $<s_1,s_i>$ $\mathbf{i \leq n}$ in which $\mathbf{T_1}$ is more persistent than $\mathbf{T_2}$ and there does not exist a smaller sub-path $<s_1,s_j>$ $\mathbf{j < i}$ for which $\mathbf{T_2}$ is more persistent than $\mathbf{T_1}$.

[2] this term is due to Yoav Shoham

Unit Paths:

1 ≡ <0,do(guzzle,0)>

2 ≡ <do(guzzle,0),do(jump,do(guzzle,0))>

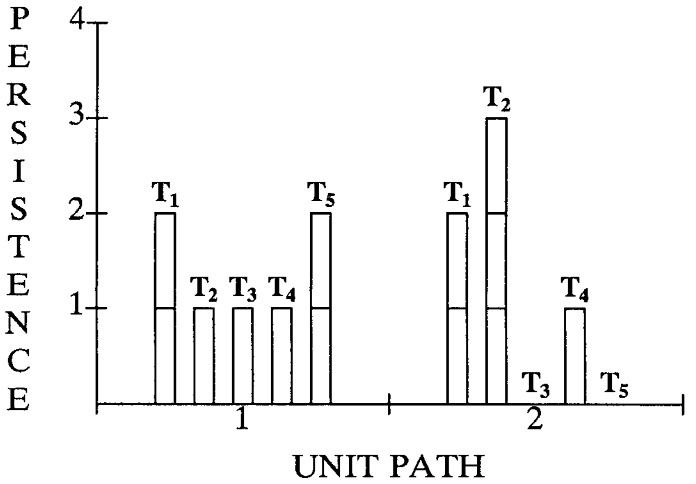


Figure 5. Persistence by unit path for Theories from Example 2

Definition 8.

A theory is said to be *chronologically maximally persistent* (CMP) over a path if there does not exist a theory which is chronologically more persistent over that path.

In example 2 (fig. 4), $\mathbf{T_1}$ is the unique chronologically maximally persistent theory over **<0,do(jump,do(guzzle,0))>**. It should be mentioned that, in general, there is no guarantee of a unique CMP theory. When there are multiple CMP theories, our intuition about persistence offers no further assistance — the problem is simply underspecified. (In the absence of any additional domain knowledge, there is no basis by which to determine a "best" theory — a conditional plan is called for.) Returning to the example, how well does the notion of chronological maximal persistence capture our intuitions about the world? As previously mentioned, the theory $\mathbf{T_1}$ corresponds to the intended model. It accurately reflects our expectations about the stability of the world. Therefore, at least in this case, chronological maximal persistence satisfies the accuracy criterion and thus could serve as a heuristic preference measure.

CMP theories, however, are usually not the most economical choice. In a CMP theory, many of the propositional fluents that persist are simply irrelevant to the problem at hand. A better theory would exclude all irrelevant instances of frame defaults. The need to simplify theories by eliminating irrelevant details motivates the following definitions (cf. [Poole86]).

Definition 9.

Let $\mathbf{T_1}$ and $\mathbf{T_2}$ be distinct theories predicting **G**. $\mathbf{T_1}$ is *simpler* than $\mathbf{T_2}$ iff $\mathbf{F \cup T_2 \models T_1}$ but $\mathbf{F \cup T_1 \not\models T_2}$. Syntactically, $\mathbf{Theorems(F \cup T_1) \subset Theorems(F \cup T_2)}$.

Definition 10.

A theory is *simplest* if there is no simpler theory.

Simplest theories contain the minimal amount of information necessary to make a given prediction. Because simplest theories can be generated in a goal directed manner (cf. [Poole85a]), it is expected that they are computationally less costly than CMP theories.

In the example above, there is a single simplest theory predicting **¬alive(do(jump,do(guzzle,0)))**, namely:

$$\mathbf{T_3} = \{\mathbf{defective(do(guzzle,0)) \leftrightarrow defective(0)}\}.$$

There is also a simplest theory predicting **alive(do(jump,do(guzzle,0)))**. It is:

$$\mathbf{T_4} = \{\mathbf{alive(do(guzzle,0)) \leftrightarrow alive(0),}$$
$$\mathbf{alive(do(jump,do(guzzle,0))) \leftrightarrow alive(do(guzzle,0))}\}.$$

From this it should be clear that simplicity alone is not an appropriate heuristic for theory preference in planning, since it does not satisfy the accuracy criterion (e.g., $\mathbf{T_4}$ is a simplest theory but its prediction does not agree with the intended model).

The sought preference heuristic is arrived at by combining the notions of simplicity and chronological persistence. From the partial ordering of theories defined by the chronological persistence, select a theory which agrees with the CMP theory and which balances the degree of simplicity and chronological persistence. The balance between simplicity and chronological persistence reflects the desire to balance resource conservatism and accuracy; we expect that decreasing simplicity increases computational cost and that increasing chronological persistence increases accuracy. The concept of balance is formalised by the following definitions.

Definition 11.

Suppose we are interested in whether **G** is true or false, and suppose there is a unique CMP theory $\mathbf{T_{CMP}}$ such that $\mathbf{F \cup T_{CMP} \models G}$. A theory $\mathbf{T_1}$ is *more balanced* than $\mathbf{T_{CMP}}$ if

a) $\mathbf{T_1}$ is simpler than $\mathbf{T_{CMP}}$ (and therefore $\mathbf{F \cup T_1 \models G}$ by definition 9);

b) there does not exist another theory $\mathbf{T_2}$ which is chronologically more persistent than $\mathbf{T_1}$ and for which $\mathbf{F \cup T_2 \models \neg G}$.

Definition 12.

A theory $\mathbf{T_B}$ is *balanced* if it is more balanced than $\mathbf{T_{CMP}}$ and there does not exist a theory which is simpler than $\mathbf{T_B}$ and more balanced than $\mathbf{T_{CMP}}$. If there is no theory which is more balanced than $\mathbf{T_{CMP}}$ then $\mathbf{T_{CMP}}$ is balanced.

Note that $\mathbf{T_B}$ can be viewed as an approximation to $\mathbf{T_{CMP}}$ since it has the property that if **P** is a prediction of $\mathbf{T_B}$ then it is also a prediction of $\mathbf{T_{CMP}}$ (the converse is not true). Consider the following theory:

$$\mathbf{T_5} = \{\mathbf{defective(do(guzzle,0))} \leftrightarrow \mathbf{defective(0)}, \\ \mathbf{alive(do(guzzle,0))} \leftrightarrow \mathbf{alive(0)}\}.$$

$\mathbf{T_5}$ is more balanced than $\mathbf{T_1}$ and there is no theory which is simpler than $\mathbf{T_5}$ and simultaneously more balanced than $\mathbf{T_1}$. Since in the above example, $\mathbf{T_1} \equiv \mathbf{T_{CMP}}$, by Definition 12, $\mathbf{T_5} \equiv \mathbf{T_B}$.

It is important to realise that the existence of $\mathbf{T_B}$ depends on the existence of $\mathbf{T_{CMP}}$. Furthermore, definition 11 depends on there being a unique CMP theory. If there are multiple CMP theories, it is still possible (though more complicated) to simplify them. Another important observation is that even when $\mathbf{T_B}$ exists, it may not be unique. In this case, either extra domain knowledge can be used to make a selection or a theory which is a disjunction of the multiple $\mathbf{T_B}$s can be used. The two definitions above could form the basis of an algorithm to find the theory $\mathbf{T_B}$; but the definitions assume we know $\mathbf{T_{CMP}}$. If we did, we could simply use $\mathbf{T_{CMP}}$. An algorithm for finding $\mathbf{T_B}$ without knowing $\mathbf{T_{CMP}}$ is presented elsewhere [Goodwin86].

The heuristic preference measure "balanced" satisfies the accuracy criterion since its predictions agree with $\mathbf{T_{CMP}}$ which corresponds to the intended model. It also satisfies the sufficiency criterion since it predicts the goal (or its negation depending on which is expected). In addition, it satisfies the conservatism criterion, since it contains only the information required to decide between competing predictions. Therefore, through the combination of the notion of chronological persistence and simplicity we have the heuristic measure we sought for theory preference in planning problems.

A word about the heuristic nature of this measure is in order here. The example of Kautz [Kautz86, p 404] shows that this measure is only a heuristic. In that example, a car is observed to be missing some time after it was parked. The "balanced" theory would predict the car was still in the parking lot until it was observed to be missing, but there is no reason to prefer this prediction — it could have been stolen anytime between when it was parked and when it was observed missing. Though our preference measure is only a heuristic, it is useful nevertheless, since it seems to correspond to our intuition about the invariance of relations.

4. Related Work

Hanks and McDermott [Hanks85, Hanks86] show that using default (or other nonmonotonic) reasoning to deal with the frame problem inevitably results in the need to chose between multiple models. Because of this, they come to the discouraging conclusion that logic is inadequate as an AI representation language. They turn to a direct procedural characterisation to describe default reasoning processes and give an algorithm that generates their intended model for a set of axioms. In our terms, this model is a model of $\mathbf{F} \cup \mathbf{T}_{\mathbf{CMP}}$.

Another related idea is that of Lifschitz [Lifschitz85, Lifschitz86a, Lifschitz86b]. He makes the observation that the usual forms of circumscription [McCarthy80] are inadequate for dealing with axiomatisations of planning problems (cf. [McCarthy86]). To overcome this inadequacy, he introduces *pointwise circumscription*. Our notion of chronological maximisation of persistence is analogous to a form of prioritised pointwise circumscription which prefers "minimisation at earlier moments of time".

The work of Shoham [Shoham86a, Shoham86b] is closely related to the work presented here. His work on the *initiation problem* led him to the idea of *chronological maximisation of ignorance*. While the initiation problem is different from the frame problem, solutions to each problem reflect the need to maximise (or minimise) step by step (i.e., chronologically).

Recent work by Kowalski and Sergot [Kowalski86a, Kowalski86b] also addresses the frame problem. More specifically, Kowalski [Kowalski86b] proposes a first order *persistence* axiom that specifies how one can deduce whether a relation holds at a given time in a particular temporal database. A database is formalised as the ground terms of a logical theory specified in *event calculus*; the first order ground atomic formula **holds(r,t)** is true when **r** is an instance of a relation, and **t** is a time interval over which the relation is true. Database update constraints are specified in terms of **terminate** and **initiate** conditions on relations, events and actions. The epistemological aspect of the frame problem is claimed to be solved by axiomatising the database in terms of a single relation **holds** (cf. [Kowalski79]), and relying on the persistence axiom's use of negation-as-failure to assume that nothing affects the truth of a relation unless explicitly declared. The heuristic aspect of the frame problem is viewed as efficiently using the persistence axiom to answer the general question of whether an instance of a relation **r** holds at time **t**. This problem is solved by describing an efficient method for implementing the use of the persistence axiom.

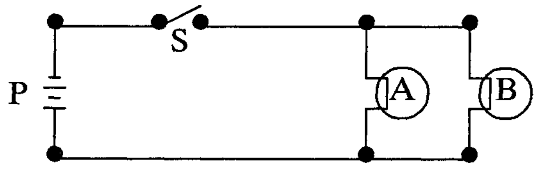
This notion of persistence, rendered as a first order axiom that specifies what "holds" in terms of how explicit initiate and terminate conditions affect the truth of relations, is only weakly related to the notion of persistence described here. Kowalski's formulation has no explicit concept of the update constraints being contingent or assumable if consistent, consequently there is no possibility of multiple conflicting answers to the question "does **r** hold at time **t**?" The persistence axiom relies on negation-as-failure while isolating those constraints that affect the truth of a relation so that different answers to the same question can be had due to intervening updates. However, there is no possibility of uncertain or multiple possible responses to questions of a relation's future persistence.

It seems that Kowalski assumes that frame axioms as defaults are unnecessary, as the concept of non-monotonicity can be handled more generally by using negation-as-failure within the persistence axiom. However, the burden to explicitly assert the affect of actions on relations remains; a default-like statement of the form "normally relation R persists" is not possible. Kowalski's alternative definition might be rephrased as "relation R persists, until I tell you otherwise."

Finally, Kautz [Kautz86] proposes a solution to the frame problem using a generalisation of circumscription. He defines the following partial ordering of models:

$$\begin{array}{l} \mathbf{M1} \leq \mathbf{M2} \text{ if and only if} \\ \forall \mathbf{t,f}\ .\ \mathbf{(t,f)} \in \mathbf{M1[Clip]} \supset (\mathbf{(t,f)} \in \mathbf{M2[Clip]}) \vee \\ \qquad \exists \mathbf{t2,f2}\ .\ \mathbf{t2}<\mathbf{t}\ \&\ (\mathbf{(t2,f2)} \in \mathbf{M2[Clip]})\ \&\ (\mathbf{(t2,f2)} \notin \mathbf{M1[Clip]}) \\ \text{and} \\ \qquad \mathbf{M1} < \mathbf{M2} \text{ if } \mathbf{M1} \leq \mathbf{M2} \text{ and not } \mathbf{M2} \leq \mathbf{M1} \end{array}$$

The predicate **Clip(t,f)** is true when the persistence of a fact **f** ceases at time **t**. From this definition, a model **M1** is strictly better than **M2** (i.e., $\mathbf{M1} < \mathbf{M2}$) when they are identical (in terms of Clip) up to some time **t** at which some fact **f** changes in **M2** but not in **M1** (i.e., $\mathbf{(t,f)} \in \mathbf{M2[Clip]}$ but $\mathbf{(t,f)} \notin \mathbf{M1[Clip]}$). This model ordering roughly corresponds to the chronological maximisation of persistence. To illustrate the difference consider the example in Figure 6 and two of the possible models corresponding to it (Table 4). Here we have two lights (A and B) connected in parallel through a switch (S) to a power source (P). For the period of interest, the switch and the two lights have an equivalent failure rate. The power source is completely reliable (and hence is irrelevant to the problem). Under Kautz's model ordering, $\mathbf{M_1}$ and $\mathbf{M_2}$ are incomparable (both are minimal) while a theory $\mathbf{T_2}$ corresponding to $\mathbf{M_2}$ is chronologically more persistent than $\mathbf{T_1}$ (corresponding to $\mathbf{M_1}$) over a path corresponding to the action between time **0** and time **1**. Thus, the assumption that all persistences are equally likely enables us to distinguish the two models. The preference heuristic reflects our expectation that the switch failing alone is a better explanation (theory) than both lights failing.



$F = \{$	
$\neg on(s,0),$	$\neg hold(0,on(s))$
$\neg on(a,0),$	$\neg hold(0,on(a))$
$\neg on(b,0),$	$\neg hold(0,on(b))$
$ok(s,0),$	$hold(0,ok(s))$
$ok(a,0),$	$hold(0,ok(a))$
$ok(b,0),$	$hold(0,ok(b))$
$on(s,1),$	$hold(1,on(s))$
$\neg on(a,1),$	$\neg hold(1,on(a))$
$\neg on(b,1),$	$\neg hold(1,on(b))$
$on(a,T) \leftarrow$	$hold(T,on(a)) \leftarrow$
$\quad on(s,T) \wedge ok(s,T) \wedge ok(a,T),$	$\quad hold(T,on(s)) \wedge hold(T,ok(s)) \wedge hold(T,ok(a))$
$on(b,T) \leftarrow$	$hold(T,on(b)) \leftarrow$
$\quad on(s,T) \wedge ok(s,T) \wedge ok(b,T)\}$	$\quad hold(T,on(s)) \wedge hold(T,ok(s)) \wedge hold(T,ok(b))$
$\Delta = \{$	
$[X,T]\ ok(X,T+1) \leftrightarrow ok(X,T),$	$hold(T+1,F) \oplus clip(T+1,F) \leftarrow hold(T,F)$
$[X,T]\ on(X,T+1) \leftrightarrow on(X,T)\}$	

Figure 6. Comparison with Kautz´s Model Ordering

5. Conclusion

The frame problem is a fundamental aspect of planning. In developing a representational scheme in a theory formation framework to deal with the frame problem, the problem of multiple theories arises. A method for selecting among competing (and possibly conflicting) theories based on a measure of persistence has been described. This method is restricted to the case where the possible hypotheses consist solely of frame defaults and where all propositional fluents are equally likely to persist. When other types of defaults are included, interactions may occur between them. The nature of these interactions and their management are under investigation. As well, ways of extending the preference heuristic to account for persistences whose likelihood varies over time and/or between relations is also being studied [Goodwin86].

Model M_1		
	0	1
ok(s)	T	T
ok(a)	T	F
ok(b)	T	F
on(s)	F	T
on(a)	F	F
on(b)	F	F

Model M_2		
	0	1
ok(s)	T	F
ok(a)	T	T
ok(b)	T	T
on(s)	F	T
on(a)	F	F
on(b)	F	F

Table 4. Two of the Models for Figure 6.

It is widely believed that some form of default reasoning is a necessary for intelligence, but as Hanks and McDermott have observed, casting even simple problems in a default reasoning framework inevitably results in multiple solutions, some of which are extraneous. What's worse, in general, there is no simple way to extended the underlying set of axioms to eliminate the extraneous solutions. We have shown how this problem can be overcome in a theory formation/theory preference framework. The simplicity and intuitive appeal of this framework begs further study.

6. Acknowledgements

We would like to thank David Poole, Andre Trudel, Ken Jackson, Eric Neufeld, and Paul Van Arragon for valuable comments on an earlier draft of this paper. Thanks also to the other members of Logic Programming and Artificial Intelligence Group of the University of Waterloo for providing an intellectual environment conducive to research and to Charley Kasmierski for providing moral support.

References

[Fikes71]

R.E. Fikes and N.J. Nilsson (1971), STRIPS: A New Approach to the Application of Theorem Proving in Problem Solving, *Artificial Intelligence* 2(3&4), Winter, North-Holland, Amsterdam, 189-208.

[Goebel85]

R.G. Goebel, K. Furukawa, and D.L. Poole (1985), Using definite clauses and integrity constraints as the basis for a theory formation approach to diagnostic reasoning, Research report CS-85-50, Department of Computer Science, University of Waterloo, Waterloo, Ontario, December.

[Goodwin86]

S.D. Goodwin (1986), Representing Frame Axioms as Defaults, Master's dissertation, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada [In preparation].

[Green81]

C.C. Green (1981), Application of Theorem Proving to Problem Solving, *Readings in Artificial Intelligence*, B.L. Webber and N.J. Nilsson (eds.), Morgan Kaufmann Publishers, Los Altos, California, 202-222.

[Hanks85]

S. Hanks and D. McDermott (1985), Temporal reasoning and Default Logics, TR YALEU/CSD/RR# 430, Computer Science Department, Yale University.

[Hanks86]

S. Hanks and D. McDermott (1986), Default Reasoning, Nonmonotonic Logic, and the Frame Problem, *The National Conference for Artificial Intelligence* **1**, August 11-15, Philidelphia, 328-333.

[Hayes71]

P.J. Hayes (1971), A Logic of Actions, *Machine Intelligence*, vol. 6, B. Meltzer and D. Michie (eds.), Edinburgh University Press, Edinburgh, 495-520.

[Jones85]

M. Jones and D. Poole (1985), An expert system for educational diagnosis based on default logic, *Proceedings of the Fifth International Workshop on Expert Systems and their Applications*, May 13-15, Palais des Papes, Avignon, France, 573-583.

[Kautz86]

H. Kautz (1986), The Logic of Persistence, *The National Conference for Artificial Intelligence* **1**, August 11-15, Philidelphia, 401-405.

[Kowalski79]
R.A. Kowalski (1979), *Logic for Problem Solving*, Artificial Intelligence Series 7, Elsevier North Holland, New York.

[Kowalski86a]
R.A. Kowalski and M.J. Sergot (1986), A Logic-based Calculus of Events, *New Generation Computing* **4**, Ohmsha Ltd. and Springer Verlag, 67-95.

[Kowalski86b]
R.A. Kowalski (1986), Database Updates in the Event Calculus, Doc. 86/12, Imperial College, July, 29 pages.

[Lifschitz85]
V. Lifschitz (1985), Circumscription in the Blocks World, Stanford University, Stanford, December, 17 pages [unpublished].

[Lifschitz86a]
V. Lifschitz (1986), Pointwise Circumscription, Stanford University, Stanford, March, 15 pages [unpublished].

[Lifschitz86b]
V. Lifschitz (1986), Pointwise Circumscription: Preliminary Report, *The National Conference for Artificial Intelligence* **1**, August 11-15, Philidelphia, 406-410.

[McCarthy69]
J. McCarthy and P.J. Hayes (1969), Some Philosophical Problems from the Standpoint of Artificial Intelligence, *Machine Intelligence*, vol. 4, B. Meltzer and D. Michie (eds.), Edinburgh University Press, Edinburgh, 463-502.

[McCarthy80]
J. McCarthy (1980), Circumscription—a form of non-monotonic reasoning, *Artificial Intelligence* **13**(1&2), April, North-Holland, Amsterdam, 27-39.

[McCarthy86]
J. McCarthy (1986), Applications of Circumscription to Formalising Common-Sense Knowledge, *Artificial Intelligence* **28**(1), February, North-Holland, Amsterdam, 89-116.

[Poole85a]
D.L. Poole, R.G. Goebel, and R. Aleliunas (1985), Theorist: a logical reasoning system for defaults and diagnosis, *Knowledge Representation*, N.J. Cercone and G. McCalla (eds.), Springer-Verlag, New York [invited chapter, submitted September 10].

[Poole85b]
D.L. Poole (1985), On The Comparison of Theories: Preferring the Most Specific Explanation, *Proceeding of the Ninth International Joint Conference on Artificial Intelligence*, August 16-18, UCLA, Los Angeles, California, 144-147.

[Poole86]
D.L. Poole (1986), Default Reasoning and Diagnosis as Theory Formation, CS-86-08, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, March, 19 pages.

[Popper58]
K.R. Popper (1958), *The Logic of Scientific Discovery*, Harper & Row, New York.

[Reiter80]
R. Reiter (1980), A logic for default reasoning, *Artificial Intelligence* **13**(1&2), April, North-Holland, Amsterdam, 81-132.

[Shoham86a]
Y. Shoham (1986), Chronological Ignorance: Time, Knowledge, Nonmonotonicity, and Causation, Computer Science Department, Yale University, April, 24 pages.

[Shoham86b]
Y. Shoham (1986), Chronological Ignorance: Time, Nonmonotonicity, Necessity, and Causal Theories, *The National Conference for Artificial Intelligence* **1**, August 11-15, Philidelphia, 389-393.