Nonmonotonic Temporal Reasoning

Erik Sandewall
Email: erisa@ida.liu.se

Yoav Shoham
Email: shoham@cs.Stanford.edu

This is a draft manuscript, intended for publication in the Handbook of Artificial Intelligence and Logic Programming*. Please do not redistribute.*

IDA RKLLAB Memo 1994
IDA-RKL-94-TR11
ISSN-0281-4250

Department of Computer
and Information Science,
Linköping University,
S-581 83 Linköping, Sweden

1 Introduction

Temporal reasoning is not restricted to reasoning about time itself; it also includes reasoning about phenomena that take place in time, i.e. *reasoning about actions and change*. Such reasoning is required for a number of purposes in A.I. programs, in particular for prediction, planning and plan modification, plan recognition, and diagnosis. These types of reasoning tasks depend crucially on the use of *action laws*, i.e. axioms which specify what effects an action will have or may have in a particular situation. Besides the action laws, each reasoning task typically involves (1) designation of certain actions which have been (will be, may be) performed, as well as their order of execution; (2) statements about the state of the world before the actions; (3) statements about the state of the world after the actions. In a planning task (2) and (3) are given and (1) is sought; in a prediction task (1) and (2) are given and (3) is sought, and so on.

Of particular interest for A.I. is *common-sense* reasoning about actions and change, where the purpose is to accomodate various phenomena that are characteristic of the everyday world and of our ways of thinking about it. Such reasoning is inherently nonmonotonic because of the need to concentrate on the normal or "default" cases of situations and actions, and to avoid excessive consideration of exceptional cases except when they are specifically relevant.

In technical terms, one considers that for a given set of known facts ("axioms") there is a set of *intended conclusions* which is a superset of those conclusions obtained by classical logic. Furthermore, although the set of classical conclusions is of course a monotonic function of the set of axioms, the same does not hold for the set of intended conclusions.

Both nonmonotonic reasoning and temporal reasoning are rich and multi-faceted subjects, with relevance to many areas of artificial intelligence. Other chapters in the Handbook deal with each of the two in detail. Our purpose here is to present varieties of nonmonotonic reasoning which are specialized for reasoning about actions and change.

The following are some related and relevant topic areas which could not be included in the present article: methods for scheduling i.e. reasoning about time constraints for a given set of actions; computational issues such as algorithms and software architectures; the problems of belief update and theory revision; and methods based on modal logics.

1.1 Methodologies

Two methodologies or "paradigms" are being used in this research. The classical paradigm or *paradigm of common-sense examples* is closely tied to the notion of developing a logic for common-sense reasoning, and has been concisely summarized by Davis [7] as follows

> *The basic approach used here, as in much of the research on automating commonsense reasoning, is to take a number of examples of commonsense inference in a commonsense domain, generally deductive inference; identify the general domain knowledge and the particular problem specification used in the inference; develop a formal language in which this knowledge can be expressed; and define the primitives of the language as carefully and precisely as possible...*

The "primitives of the language" include the primitive inference operations. This paradigm has been applied not only for temporal reasoning (where the "Yale shooting problem" is the most often referenced example of commonsense inference), but also for inheritance reasoning where one has used examples such as the Nixon dilemma ("Nixon diamond"), the royal elephant, and the beer-drinking marine chaplain.

The use of common-sense examples was universally accepted for a while, and there have been proposals e.g. by Lifschitz [29] to form a standard collection of common-sense examples for nonmonotonic logics. More recently one has realized that the described paradigm also has a number of problems:

- The methodology of common-sense examples assumes implicitly that there *is* such a thing as a logic of common-sense reasoning, which merely remains to be discovered. If it exists, then presumably it can be reconstructed from examples, but how do we know that it exists at all?

- There are often "clashes of intuitions" where different informers make incompatible statements about what are the admissible common-sense inferences for a given example. This means that the empirical data for this line of research are inherently unreliable. Also since the informer usually equals the researcher, there is always a danger that he(she) will be biased by the particular theory he has developed (Touretzky [65]). A major reason for the uncertainty is that it is hard to delimit the "general domain knowledge" that has been used for a particular inference.

- As one goes beyond the trivial examples to more complicated ones, the potentially relevant general domain knowledge broadens, so the uncertainty increases about what are the admissible common-sense inferences. A logic of common-sense reasoning based on examples has difficulties to develop, therefore. Progress is hampered by the increasing uncertainty in the empirical data, as one proceeds beyond the simplest examples.

- The use of nonmonotonic logics aggravates these problems. If classical logic is being used, then a proposed formalization can arguably be accepted even if it obtains a too small set of conclusions, since that

> can always be corrected by "adding more knowledge", but it must definitely be rejected if it obtains some unintended conclusions. However with nonmonotonic logics, *any* proposed formalization can possibly be salvaged by adding more axioms which obliterate previously obtained conclusions.

Several researchers have therefore recently switched to a *systematic paradigm* where each proposed logic is accompanied by an *assessment*, i.e. a statement of a class of reasoning tasks for which the logic is guaranteed to obtain exactly the intended conclusions. A precise definition of the set of intended conclusions for arbitrary members of the class is then required.

Each of these paradigms has its merits and demerits. The paradigm of common-sense examples has been very useful for identifying interesting types of reasoning problems, and for reminding us of the great variation of difficulties in common-sense reasoning. In particular it has been instrumental for identifying three problems that arise when one tries to formalize common-sense reasoning about actions and change: the frame problem, the ramification problem, and the qualification problem. The assortment of difficulties which have been identified in this fashion is commonly called the *frame problem*, in the singular or the plural. The systematic paradigm eliminates the uncertainties that plague the first methodology, although it does so by moving the uncertainty outside the scope of investigation. The definition of the set of intended conclusions, or models, is still subject to choice and must be based on intuitions. It is only after that choice has been made and formalized that the systematic methodology can be applied.

1.2 The structure of time

Various possible models of time are reviewed separately in the article *Time and Change for AI* by Anthony Galton in this Handbook. Research on logics for actions and change in A.I. has traditionally mostly used the *situation calculus*, which was introduced by J. McCarthy around 1960 [35]. Its underlying assumption is that the state of the world can be described in terms of *situations*, whose exact character is somewhat different between different authors, and in terms of *actions* which transform the world from one situation to another. Other common choices have been to characterize time in terms of *explicit integer timepoints*, thereby providing a discrete and linearly ordered time domain, or in terms of *intervals* whose endpoints are supposed to be points on a linear time axis. In all cases the temporal entities in question (situations, timepoints, integers) are represented explicitly in the logic, usually as a separate sort in a multi-sorted first-order logic.

1.2.1 Structure and authorship of this chapter

Sections 2 and 3 of this article are based on an original manuscript by Shoham, which was later revised and extended by Sandewall. Section 2 makes an overview of issues and aspects of the frame problem from a historical perspective. Section 3 describes a number of logics which have been proposed within the methodology of common-sense examples, and mostly using the situation calculus. To facilitate the reading we will use a uniform situation-calculus notation throughout sections 2 and 3, and normalize the examples of different authors accordingly.

The following sections have been written by Sandewall. Section 4 is a brief overview of recent research using the systematic methodology. In section 5 follows a detailed account using one particular systematic approach, providing assessments of logics for the basic version of the frame problem where only persistence is assumed. It uses a framework of explicit integer time.

2 The frame problem(s)

2.1 The situation calculus

The situation calculus has traditionally been the major notational tool for this branch of research. A *situation* is intended as the actual or hypothetical state of the world at a particular time. Given an *action* A and a situation s, $\texttt{Result}(A, s)$ denotes the new situation after action A is performed in situation s. Given a *property* p and a situation s, $\texttt{Holds}(p, s)$ indicates that the property p is true in situation s. With this notation it is possible to quantify over properties while staying in a first-order logic. The property domain is often restricted to a finite set of explicitly given properties, but some authors have also used composite properties such as $p \wedge q$ with the obvious intended meaning.

McCarthy's original proposal did not include the use of such reified properties; his examples use functions like $\texttt{location}(x, s)$ for the location of object x in situation s. However reified properties and the `Holds` predicate are now commonly used. The notation defined now will be used uniformly in sections 2–4.

The situation calculus is basically just a notation. Different authors introduce and use different axiomatizations for it, reflecting not only technical differences in their logics but also different notions of what a situation *is*. One intuition is that the situation domain forms a tree structure with an initial situation `S0` as its root and where the function `Result` is used to form successive descendants. Situations are therefore considered as nominal entities *at which* various facts are true or false, and constitute in fact a branching time structure. As time progresses one can never get back to the same situation as one was in before. In accordance with this approach, Reiter and Lin [55, 32] define a Peano-like axiomatization for the situation calculus with the axioms

$$\texttt{S0} \neq \texttt{Result}(A, s)$$
$$\texttt{Result}(A, s) \ = \ \texttt{Result}(A', s') \ \rightarrow A = A' \wedge s = s'$$

and an induction rule to the effect that if $\Phi(\texttt{S0})$ and $\Phi(s) \rightarrow \forall A[\Phi(\texttt{Result}(A, s))]$ then $\forall s[\Phi(s)]$. This choice of semantics for situations will be referred to as *Herbrand situations*.

Some authors reject this intuition. Baker and Ginsberg write: *"our intent is to think of a situation not as a sequence of actions, but instead in terms of the values taken by various fluents"* [2]. The notion of situation must then be tied to the notion of *states*, i.e. mappings from properties to corresponding truth-values. States (designated by r) are characterized by the *uniqueness property*:

$$\forall r \forall r' [\forall p [\texttt{Holds}(p, r) \equiv \texttt{Holds}(p, r')] \rightarrow r = r']$$

and the *existence property*

$$\forall r \forall p \exists r' [(\texttt{Holds}(p,r) \equiv \neg\texttt{Holds}(p,r')) \quad \wedge \\ \qquad \forall p' [p \neq p' \rightarrow (\texttt{Holds}(p',r) \equiv \texttt{Holds}(p',r'))]].$$

A *strict state semantics* would identify situations and states, but this alternative has not been used to our knowledge. However Baker [1] has proposed to use a *state-oriented semantics* where situations have the existence property but not the uniqueness property. This means that each situation "contains" a state, and for each state there is at least one situation containing it. (However Baker formulated the existence property differently, using circumscription and a special abnormality predicate). – Finally there exist also other notions of situations e.g. as partial states.

When a scenario has been formalized in the situation calculus, each model contains the whole situations space i.e. it assigns a truth-value to each combination of property and situation. Several alternative histories or futures are therefore represented within each model. Multiple models can be obtained if $\texttt{Holds}(p,\texttt{S0})$ is unspecified for some p, and also if the effects of some actions are incompletely known or nondeterministic.

2.2 The nonmonotonicity property

The frame problem was very early recognized as one of the topics requiring nonmonotonic reasoning. Let us consider a simple example, namely a blocks world allowing properties of the form $\texttt{On}(x,y)$ representing that block x is on top of block y, $\texttt{Clear}(x)$ representing that nothing is on top of block x, and $\texttt{Compressed}(x)$ representing that block x is compressed. We assume a definition of common-sense conclusions for this world where the position, clearedness, and compressedness of a block can be inferred to remain unchanged except for the immediate effects of the actions Stack and Squeeze.

If Γ is a set of propositions, we write $[\![\Gamma]\!]$ for the set of classical models for Γ and $\Sigma(\Gamma)$ for the set of intended models. For example, if Γ has the following axiom

$$\texttt{Holds}(\texttt{Clear}(x),s) \wedge \texttt{Holds}(\texttt{Clear}(y),s) \wedge x \neq y \\ \quad \rightarrow \texttt{Holds}(\texttt{On}(x,y),\texttt{Result}(\texttt{Stack}(x,y),s))$$

as its single member, then any member of $[\![\Gamma]\!]$ as well as of $\Sigma(\Gamma)$ must satisfy this axiom. Suppose s_1 is a particular situation term, a and b are blocks, $\texttt{Clear}(a)$ and $\texttt{Clear}(b)$ hold in s_1, and the term $s_2 = \texttt{Result}(\texttt{Stack}(a,b),s_1)$. An interpretation which satisfies Γ but where also $\texttt{Holds}(\texttt{On}(c,d),s_1)$ is different from $\texttt{Holds}(\texttt{On}(c,d),s_2)$ for some c and d different from a and b, is a member of $[\![\Gamma]\!]$ but not of $\Sigma(\Gamma)$, since it violates the assumption that properties do not change without a reason. Similarly an interpretation which satisfies the axiom but where $\texttt{Compressed}(b)$ is false in s_1 but true in s_2 is also a member of $[\![\Gamma]\!]$ but not of $\Sigma(\Gamma)$.

Conventional logic provides the definitions and the means of using $[\![\Gamma]\!]$. The research problem is how to obtain $\Sigma(\Gamma)$ or the corresponding conclusions (formulae true in all members of $\Sigma(\Gamma)$) in terms of operations on formulae in Γ.

One obvious way of obtaining $\Sigma(\Gamma)$ is by introducing additional axioms. If the intended set of models for Γ is a subset of the actual (classical) set of models, then there are some other constraints that one intends but which have not been made explicit in Γ. Arguably the correct course of action would be to add those constraints to the set of axioms. However there are several difficulties with this seemingly straightforward answer. Some of the more complex notions of common-sense conclusions can not be expressed as additional axioms except after considerable extensions of the logic's expressiveness. The assumptions in our example represent a simple case which *can* be completely characterized using additional axioms in situation calculus, but it has often been argued that this solution is impractical due to the large number of resulting so-called frame axioms. These considerations have resulted in a search for other ways of obtaining $\Sigma(\Gamma)$ where the language of the logic is retained but a more complex inference mechanism is used.

It is easily verified that the operator Σ is nonmonotonic, for the chosen definition of common-sense inferences, whereas of course the $[\![.]\!]$ operator is monotonic. In other words, if $\Gamma \subseteq \Gamma'$ then one can conclude $[\![\Gamma']\!] \subseteq [\![\Gamma]\!]$ but not $\Sigma(\Gamma') \subseteq \Sigma(\Gamma)$. For example, if Γ' is obtained from Γ by adding the following axiom

$$\begin{aligned}&\texttt{Holds(Clear(}x\texttt{),}s\texttt{)} \wedge \texttt{Holds(Clear(}y\texttt{),}s\texttt{)} \wedge x \neq y\\ &\quad \rightarrow \texttt{Holds(Compressed(}y\texttt{),Result(Stack(}x,y\texttt{),}s\texttt{))},\end{aligned}$$

saying that if x is stacked on top of y then y becomes compressed, then the intended model sets $\Sigma(\Gamma)$ and $\Sigma(\Gamma')$ are clearly disjoint except for odd cases such as a model containing just one block. In $\Sigma(\Gamma)$ a block can never become compressed when another block is stacked on it, whereas in $\Sigma(\Gamma')$ this always happens.

For these reasons, and since $\Sigma(\Gamma) \subseteq [\![\Gamma]\!]$ for any Γ, it was believed quite early that a systematic logic-based solution of the frame problem could be obtained by developing an appropriate nonmonotonic logic. In fact the frame problem was a large part of the motivation for the emergence of those logics in A.I. research.

The phenomenon of nonmonotonicity in reasoning about actions and change has now been formulated in terms of the model-sets $[\![\Gamma]\!]$ and $\Sigma(\Gamma)$, but it could also have been formulated in terms of the classical set of conclusions $Cn(\Gamma)$ and the nonmonotonic set of conclusions $C(\Gamma)$ as used in the article *General Patterns in Nonmonotonic Reasoning* by David Makinson in this Handbook. The formulation in terms of model-sets lends itself more easily to the specification of preferential entailment criteria which will follow later on in the present article.

2.3 Persistence, ramification, surprises, and qualification

The example in the previous subsection illustrated a particular case of the frame problem which is often called the *persistence problem.* (Sometimes the term "frame problem" is only used in the restricted sense of the persistence problem; sometimes it is given a much broader meaning). Underlying it there is an *assumption of persistence* or *inertia*: properties of the world do not change except as the direct and known effect of an action. The problem is how to realize this assumption in a convenient and efficient fashion in a logic.

Some other and similar problems have also been defined. It is important to discuss them together, since the implicit assumption in much of this research has been that a logic of common-sense reasoning must provide solutions for all of the problems at once. This is true in particular for the research in the paradigm of common-sense examples.

The *ramification problem* is the difficulty of formalizing all of the things that *do* change as the result of an action, given that the effects of an action may be very complex. For instance, if a man dies then his wife becomes a widow, friendship relations cease to apply, the ownership of his belongings change, etc. Some of these changes are also obtained as the results of other events. Rather than list all effects explicitly for each action, one would like to state them in more general and implicit terms. Some of the effects may be implicitly characterized by constraints on the possible states of the world. Other effects may be specified by cause-effect rules.

The simplest possible variant of the ramification problem arises if one allows a pair of two complementary properties such as `Alive` and `Dead`, related by the obvious axiom

$$\texttt{Holds(Alive,}s\texttt{)} \equiv \neg\texttt{Holds(Dead,}s\texttt{)},$$

and one expects that if an action law specifies a change for one of these properties, then the corresponding change in the other property will follow as a consequence. In conventional logic this is of course not a problem, but if an assumption of persistence has been built into the logic then one must take care that it is overridden properly in case of ramification. The interesting variants of ramification, from a common-sense perspective, are of course significantly more complex than this.

There are two major ways of dealing with the ramification problem. In one approach, one chooses to assume that each action results in a minimal set of changes. Changes of property value are accepted if they are motivated in view of constraints, cause-effect rules, or other available information; all other properties are assumed to retain their old value according to the case of the persistence assumption. However the notions of "motivated" and a "minimal set of changes" require some additional analysis. We will refer to this approach as *constraint-based* ramification.

Another approach is to distinguish between *independent* and *derived* properties, where persistence is only assumed for the former. The values of derived properties are defined or at least constrained by axioms which relate them to the independent properties or to other derived properties, but within those bounds they vary freely. For some applications this approach is not expressive enough, but it does have a clearer motivation than the minimization of changes.

The assumptions for common-sense reasoning may choose to allow or not to allow *surprises*. Consider for example the "Stolen car scenario" first described by Kautz [22]. The world is characterized in terms of a single property p representing that a given car is in a given garage. There are a few actions such as parking the car, leaving the car in the garage overnight, etc. In line with the persistence assumption it is considered that if the car is parked and left overnight, then it is still in the garage the next morning. However it is also known that the car was parked in the garage on day t and on day $t+2$ it is observed not to be in the garage.

One possible analysis of this scenario is that it is inconsistent, since the observation that the car is gone on day $t+2$ contradicts a conclusion from the other assumptions including the assumption of inertia. In this case the logic does not enjoy the property of consistency preservation, as defined in the Handbook article of D.Makinson. In another analysis one argues that the scenario must necessarily have *some* model, since by definition the informant can not have lied, and models can only be obtained by assuming that some *surprise* has occurred. Presumably the possible surprises in this case are that the car was stolen during the first night, or that it was stolen during the second night. In general, one takes recourse to surprises in order to avoid obtaining an empty model set, and one wishes to minimize the extent of those assumed surprises. Different varieties of *surprise assumptions* specify how surprises may be chosen and by what criteria they are to be minimized.

The *qualification problem* arises from the observation that many actions are defined for a "normal" case and are infeasible or without effect when the normal case is not satisfied. For example, normally the result of turning the ignition key of the car is to start the engine, but there are many exceptions, or qualifications, to this rule: the battery may be dead, the gas tank may be empty, the tailpipe may be blocked, and so on. Typically one wants the reasoning system to bypass these unlikely possibilities. It should jump to the conclusion that the action will succeed, and only consider the qualifications when necessary. There are several situations where qualifications come into play. Like surprises, qualifications may serve to avoid obtaining an empty set of models in postdictive reasoning. They can also be applied in predictive reasoning when the unusual case is known to occur. If we know that the gas tank of the car is empty then we conclude that the car can not be started; if we do not know whether it is empty or not we may (in some contexts) assume that is is in order. Finally, of course, a cautious planning system should also consider those things that may go wrong in a plan – it is not always wise to assume that things we do not know about will not fail.

Although we have so far used the situation calculus for the examples, the frame problem(s) reappear in all approaches to reasoning about action and change, and in particular for the other choices of the time domain which were mentioned in subsection 1.2. In [63], Shoham and McDermott attempt to define the problems in a formalism-independent manner.

Some authors have argued that the frame problem is not really a problem, that its solution is trivial, or that it has already been solved [52, 4]. The generality of the frame problem has also drawn the attention of cognitive scientists not directly interested in either temporal reasoning or nonmonotonic reasoning. Their much broader interpretations of the frame problem are collected in [50].

2.4 Nondeterministic actions

Most of the common-sense examples in the literature assume that actions are deterministic, i.e. that the state of the world when the action ends is uniquely determined by the state of the world when the action starts. However, nondeterministic actions arise in many applications, not only because of intended randomization such as the tossing of coins and spinning of roulette wheels, but also due to discretization. Even if the world is deterministic "in itself" or according to its description in Newtonian physics, the shift to a discrete and qualitative level of description involves a loss of information which may cause actions to appear nondeterministic. For example, if the vertical position of an object is characterized in terms of which floor it is on, then raising the object two meters may or may not shift it to the next floor up.

The use of the function name `Result` in the situation calculus does not indicate a restriction to deterministic actions. An action `A` which causes either of the two properties `P` and `Q` to shift to the opposite truth value could be completely characterized (in the context of a persistence assumption) by the following action law:

$$(\mathtt{Holds(P,}s\mathtt{)} \equiv \neg\mathtt{Holds(P,Result(A,}s\mathtt{))}) \oplus (\mathtt{Holds(Q,}s\mathtt{)} \equiv \neg\mathtt{Holds(Q,Result(A,}s\mathtt{))}),$$

where $\oplus$ represents exclusive or. Now suppose `Holds(P,S0)`, `S1 = Result(A,S0)`, `S2 = Result(A,S1)`, and in both `S1` and `S2` it is the value of `P` that has been commuted relative to the preceeding situation. Therefore `S0` and `S2` are associated with the same state of the world. What does one then want to conclude about the value of `Holds(P, Result(A,S2))`?

In a strict state semantics for situations, s is considered to refer precisely to the state of the world where the action is performed. In this case `S0` and `S2` are equal, so `Result(A,S2)` equals `S1`, and `P` is false there. The disjunction in the action law expresses that either of `P` or `Q` may change depending on what state one does the action in, but not for a given state. The strict state semantics therefore results in a deterministic reading of the action law.

In a Herbrand semantics for situations, on the other hand, this action law will obtain a set of different models, where each model contains its own outcome for each occurrence of the action `A`. In particular it allows models where different occurrences of the action have different effects, for example where `S1` is obtained by a change of `P` but `Result(A,S2)` by a change of `Q`. This is the nondeterministic reading of the action law.

If the same action law is used in a state-oriented semantics, where several situations s_i may contain the same state r, it is still possible to obtain nondeterminism like for Herbrand semantics. Even in the case of strict state semantics one can expand the situation space by introducing additional properties, for example a time-counting property with the axiom

$$\texttt{Holds(Time(}t\texttt{),}s\texttt{)} \rightarrow \texttt{Holds(Time(}t+1\texttt{),Result(A,}s\texttt{))},$$

with a corresponding modification of the situation-existence axiom. The situation space then becomes large enough for nondeterminism. The property `Time(`t`)` must then be treated as a nonpersistent property in the same way as those discussed in section 2.3.

Even though a particular variant of the situation calculus is *representationally* able to express nondeterministic actions as required by an application, there may still be problems with the persistence assumption. For example, define the action `Toss` of tossing a coin so that the value of `Holds(P,Toss(`s`))` is either true or false, regardless of the value of `Holds(P,`s`)`. The straightforward expression for the action law would be

$$\texttt{Holds(P,Result(Toss,}s\texttt{))} \vee \neg\texttt{Holds(P,Result(Toss,}s\texttt{))},$$

but this is just a tautology. Even worse, the principle of persistence in its simplest form would dictate that the coin never changes side as a result of tossing!

Another example of the same problem is obtained for the action `A` above if the exclusive or ($\oplus$) in the action law is replaced by inclusive or ($\vee$).

There are several ways of dealing with this problem. One possibility is to extend the language and the machinery of the logic so that exceptions from persistence can be stated explicitly; we shall soon see examples of this. Another possibility is to technically assume that all actions are deterministic, but to introduce auxiliary properties that serve as "random number generators". If `R` is such a property one could write the action law for tossing the coin as

$$\texttt{Holds(P,Result(Toss,}s\texttt{))} \equiv \texttt{Holds(R,}s\texttt{)}.$$

Thus the `R` property in situation s may be understood as the combination of all that information – unknown and unknowable – that together defines the outcome of tossing the coin. Such auxiliary properties must then be considered as nonpersistent in the same way as `Time(`t`)`.

2.5 The historical development

The frame problem was identified by J. McCarthy in the pioneer days of A.I. research, and was described by McCarthy and Hayes in [40]. That article discussed persistence under the title of the frame problem, and the problem of qualification but without assigning a name to it. The term "qualification problem" appeared in [36]. The ramification problem was originally considered as a part of the broader frame problem, but it was assigned a separate name and identity in [10]. Surprises are usually discussed as an aspect of the restricted frame problem.

McCarthy first suggested to deal with these frame problems by generalizing formal proofs into "literature" sequences which allow explicit reference to the previous elements in the same sequence, in particular through nonstandard inference steps of the form

$$\frac{\mathtt{T}}{\mathtt{Consistent}(\phi)}$$

provided that ϕ is consistent with the conjunction of all previous elements of the sequence. Defaults for a statement ϕ would be expressed as axioms of the form `Normally(`ϕ`)` and used in the nonstandard inference schema

$$\frac{\mathtt{Normally}(\phi)\mathtt{,\ Consistent}(\phi)}{\mathtt{Probably}(\phi)}.$$

Sandewall [56] proposed to use instead an operator `Unless(`ϕ`)` with the intended meaning "it can not be proven that ϕ", and to write the assumption of persistence as a nonstandard inference rule which in contemporary notation would be the default rule

$$\frac{\mathtt{Holds}(p(x), s)\ :\ \neg\mathtt{Ends}(p(x),\mathtt{Result}(A, s))}{\mathtt{Holds}(p(x),\mathtt{Result}(A, s))}$$

saying that if the object x has the property p in the situation s, and if x can not be proven to end being p in `Result(`A`,`s`)` then x is p in `Result(`A`,`s`)` as well. (The operator `Unless` has been rewritten as $:\neg$). Action laws must specify both the effects of actions and the resulting `Ends` propositions. It was proposed to deal with ramification using axioms that imply additional `Ends` propositions, such as `Ends(Alive,`s`)` $\rightarrow$ `Ends(Dead,`s`)`. The major purpose of the article was to argue that the frame problem could best be addressed by modifying conventional logic into a nonmonotonic variant, instead of using the procedural approach and the `Thnot` operator advocated by Hewitt [18]. It also discussed several possible approaches to the definition of such a logic, identified some of the difficulties, in particular the phenomenon of multiple extensions, and described several approaches in outline but none in detail.

The following years saw a debate about whether nonmonotonic logics were appropriate or deserved to be called logics [24, 19], as well as the actual development of several nonmonotonic logics before 1980 [37, 53]. The attention was concentrated on the formal properties of these logics, whereas their relevance for e.g. the frame problem was not analyzed. However in an article first published in 1984 [38, 39] McCarthy discussed principles for the formalization of various common-sense phenomena using nonmonotonic logics, and in particular he proposed a solution for the frame problem in a blocks world using situation calculus and circumscription. His suggestion, addressing both the persistence and qualification problems, was to use a number of abnormality predicates which would be minimized by circumscription. Rewritten in terms of reified properties, McCarthy's proposal used the following *default frame axiom* for persistence

$$\neg\texttt{Ab}(p, A, s) \;\rightarrow\; [\texttt{Holds}(p, s) \;\equiv\; \texttt{Holds}(p,\texttt{Result}(A, s))].$$

Such an axiom says that the truth-value of a property persists from one situation to the next unless something is abnormal. The action law for the move action would consist of two parts, namely

$$\texttt{Ab}(\texttt{Loc}(x, l),\texttt{Move}(x, l'),s)$$

saying that the (old) location-property of an object is abnormal (and therefore exempt from persistence) when the object is moved, and secondly

$$\neg\texttt{Abq}(\texttt{Move}(x, l),s) \;\rightarrow\; \texttt{Holds}(\texttt{Loc}(x, l),\texttt{Result}(\texttt{Move}(x, l),s))$$

saying that if an object x is moved to a location l then its location is l in the resulting situation, unless something about x, l, or the originating situation is abnormal w.r.t. the move action. Various qualifications can then be added in a modular fashion by axioms such as

$$\neg\texttt{Holds}(\texttt{Clear}(l),s) \;\rightarrow\; \texttt{Abq}(\texttt{Move}(x, l),s)$$

$$\neg\texttt{Holds}(\texttt{Clear}(\texttt{Top}(x)),s) \;\rightarrow\; \texttt{Abq}(\texttt{Move}(x, l),s)$$

and so on. Two abnormality predicates are used namely `Ab` for persistence and `Abq` for qualification.

McCarthy's approach therefore required each action law to explicitly specify both the new values of the affected properties, and the abnormality of their old values. This can be considered as a moderate price that has to be paid in order to avoid having to write out all the frame axioms; it can also be used as the declaration of possible change that is required for nondeterministic actions such as `Toss`.

It was hoped that minimization of these abnormality predicates would obtain the intended common-sense conclusions. Unfortunately this is not the case in general. With their trend-setting *Yale shooting problem*, Hanks and McDermott [16] demonstrated that even if abnormality minimization is only used for persistence with the predicate `Ab` one does not always obtain the intended set of models. The problem arises in scenarios having the following pattern:

- There are at least two actions in succession and there is some intended model where the following holds:

- Some property p persists during the first action.

- The effects of the second action are conditional on its starting state. In the intended model it affects some property which it would not affect in the modified interpretation where the property p did not persist in the first action.

Hanks and McDermott presented this counterexample with more dramatical effect as the following concrete example which we will simplify slightly. There are two properties, `Loaded` and `Alive`. Ther are also two actions, `Wait` and `Shoot`, which are performed in succession. If the gun is shot while it is loaded, a certain person (typically named Fred) dies. In the initial situation `S0`, the gun is loaded and Fred is alive. The property `Loaded` persists during `Wait`, and `Shoot` has more effects if the gun is loaded than otherwise.

Combining the result part and the abnormality part of the action law for `Shoot` into one formula, we obtain the following domain axioms:

$$\texttt{Holds(Loaded,}s\texttt{)} \rightarrow \neg\texttt{Holds(Alive,Result(Shoot,}s\texttt{))} \wedge \texttt{Ab(Alive,Shoot,}s\texttt{),}$$

$$\texttt{Holds(Loaded,S0),}$$

$$\texttt{Holds(Alive,S0).}$$

We are interested in what would happen if the actions `Wait` and then `Shoot` are performed in succession, i.e. what properties hold in the situation `Result(Shoot,Result(Wait,`

Since there is no action law for `Wait`, and the persistence assumption says that there are no changes besides those that have been specified in the axioms, we expect that the `Wait` action would have no effect. The gun would remain loaded, and the shooting would kill Fred. We therefore consider $\neg$`Holds(Alive,Result(Shoot,Result(Wait,S0)))` to be an intended conclusion.

Consider then the result of circumscribing the set of four axioms namely the three domain axioms and the default frame axiom mentioned before. Unfortunately this provides another model, besides the intended one, namely one where the gun becomes unloaded mysteriously during the waiting action, and Fred survives. This second model contains an abnormality during the waiting that was not present in the first model, but there is no longer an abnormality during the shooting. Since the gun is unloaded, Fred does not change from `Alive` to not `Alive`. Therefore, both models are minimal. If one wishes the logic to obtain exactly the intended set of models (i.e. to obtain completeness and not only soundness relative to the set of intended models) then the formalization must be altered in some way to rule out the anomalous model.

Hanks and McDermott also tried the same formalization in default logic, replacing the circumscription of Ab with the normal default rule

$$\frac{:\neg\mathtt{Ab}(p,A,s)}{\neg\mathtt{Ab}(p,A,s)}.$$

Several extensions are obtained, one corresponding to the intended model and one corresponding to the anomalous model in the circumscriptive formulation.

At the same time it was noticed by Lifschitz and reported by McCarthy [39] that the minimization of abnormality for qualification leads to a similar anomaly. The problem is explained in terms of a simple blocks world. Suppose we state that normally the action of stacking x on y succeeds:

$$\neg\mathtt{Abq(Stack(}x,y\mathtt{)},\ s\mathtt{)} \rightarrow \mathtt{Holds(On(}x,y\mathtt{),Result(Stack(}x,y\mathtt{)},s\mathtt{))},$$

but the action is qualified if anything is on x (the robot is too weak to lift more than a single block):

$$\mathtt{Holds(On(}w,x\mathtt{)},s\mathtt{)} \rightarrow \mathtt{Abq(Stack(}x,y\mathtt{)},s\mathtt{)}.$$

There may of course also be other qualifications. Failed actions are considered to have no effect. Now suppose we try to first move block A onto block B, and then to move B onto block A. Intuitively, the first action should succeed, and then since B would no longer be clear, the second action would fail. But we cannot reach this conclusion by circumscribing Abq, because there will be an alternate minimal model in which the first action fails so that the second can succeed.

2.6 Repercussions of the shooting problem

Hanks and McDermott made their own formalizations of the Yale shooting problem, using circumscription, default logic, and autoepistemic logic, and observed that the same non-intended model is obtained in all of their formalizations. They concluded that the choice of nonmonotonic logic is not the major issue, and became skeptical about whether *any* nonmonotonic logic could deal adequately with an interesting range of common-sense scenarios and reasoning tasks.

Arguably, they obtained these negative conclusions because essentially the same formalization had been expressed in all three nonmonotonic logics. Another approach which was available in the literature at the time the YSP was presented, namely Sandewall's 1972 formalization of the persistence assumption as a rule in default logic and Reiter's 1980 semantics for default logic, together do obtain the intended set of conclusions for the Yale shooting problem. However the argument of Hanks and McDermott was that even if *some* variants of nonmonotonic logic turn out to be adequate for one or the other scenario example, it would still be unclear how to obtain an appropriate variant each time a new reasoning task is presented. For those researchers who wanted to pursue the logicist approach, the natural way to meet this challenge was to try and find a logic that could provide exactly the intended set of conclusions over a convincingly large class of examples of common-sense reasoning.

3 The classical paradigm

Following the publication of the Hanks-McDermott article, a number of researchers proposed other variants of nonmonotonic logics which at least would obtain the intended conclusions for the Yale shooting problem specifically. Additional counterexamples to the new proposals were quickly found, triggering new revisions of the logics. The resulting line of research used the methodology of representative examples, where each proposed logic is supported by the examples for which it obtains the intended conclusions, and refuted by those examples where it does not. In this section we will discuss some of those solutions.

3.1 Principles for assignment of blame

In order to use the classical paradigm at all one must formulate what is the problem if a proposed logic fails for a particular example. First of all one must distinguish between *insufficient expressiveness* where even the syntax and the intended meaning of the proposed logic language does not allow one to express the example at hand, and on the other hand those cases where the example can be expressed but *wrong conclusions* are obtained. Already the subsection about nondeterminism showed how both of these aspects may be important.

Insufficient expressiveness may for example be due to the choice of time structure and the choice between propositional or first-order logic. For the wrong set of conclusions, which was the case in the Yale shooting problem, there are a number of possibilities:

- The problem is with the choice of nonmonotonic logic: circumscription, default logic, autoepistemic logic, etc., and can be remedied by switching to another one of those logics.

- The problem is instead in the choice of *selection criterion* that maps the larger set of classical models to the narrower set of intended models.

In the second case one perceives that each selection criterion can be expressed in several of the non-monotonic logics, and that switching to another logic while retaining the same selection criterion does not help. In addition the problem may be due to the formalization of the scenario:

- Some intended axioms may have been overlooked and omitted, whereby some unintended models were obtained.

- The problem is with the choice of time domain, and can be remedied by changing that choice.

- Particularly for descriptions of continuous phenomena, the use of a discrete-level formalization means a certain loss of information, which may also result in obtaining some unintended models.

Most research in A.I. has used the explicit-time approaches, particularly because one has considered them to be more adequate in terms of expressiveness. The blame for wrong conclusions is usually assigned to the first and/or the second of these five items.

3.2 Original chronological minimization

The first idea in the post-YSP debate was that reasoning should follow the direction of time, that is, the default assumptions should be applied in temporal order. Following this principle in the shooting scenario, we would first conclude that the waiting action does not lead to any change. Then, since the gun would remain loaded, we would be able to conclude that Fred dies. Three separate authors formalized this principle. Kautz [22] introduced a logic of *persistence*, Lifschitz [28] used *pointwise circumscription*, and Shoham [62] introduced *chronological minimization* and then *chronological ignorance*. Their solutions differ in the technical presentation, and for some examples they give different conclusions, but the basic principle in them all is to choose a partial order of the models (as used e.g. in circumscription) so as to prefer models in which the abnormalities happen as late in time as possible. In other words, they minimize the abnormalities in earlier situations at a higher priority than those in later situations.

We will discuss Kautz's variant but use Shoham's term "chronological minimization" as a generic term for all of these solutions. The method was formulated in terms of integer time, but we disregard that detail here and try a corresponding formulation in situation calculus and using our unified notation. Kautz wrote the scenario description essentially as follows

$$\texttt{Holds(Loaded,}s\texttt{)} \rightarrow \neg\texttt{Holds(Alive,Result(Shoot,}s\texttt{))},$$

$$\texttt{Holds(Loaded,S0)},$$

$$\texttt{Holds(Alive,S0)}.$$

so abnormality was not explicitly implied in the action law. The default frame axiom was replaced by an axiom of the form

$$(\texttt{Holds(}p\texttt{,}s\texttt{)} \equiv \neg\texttt{Holds(}p\texttt{,Result(}A,s\texttt{)))} \equiv \texttt{Clip(}p\texttt{,Result(}A,s\texttt{))}$$

so that Clip is true exactly when Holds has just changed its value. The set of selected models is then chosen as $Min(\ll, [\![\Gamma]\!])$, where Γ is the set of all the axioms, and $\ll$ is a preference relation which prefers one model over another if (1) they are equal in all components except Holds and Clip, and (2) there is some point of comparison s_c such that the two models agree with respect to Clip for situations (really: timepoints) earlier than s_c, and $\lambda p.\texttt{Clip(}p\texttt{,}s_c\texttt{)}$ is a smaller set in the preferred model than in the other one.

While this approach does in fact give the intended answer to the Yale shooting problem, counterexamples were also quickly found. Its applicability is limited to temporal prediction problems, or in other words, problems in which given the initial conditions, we are asked to predict what will be the case at a later time. It also requires that the initial state is completely known. Suppose we know that Fred is initially alive:

```
Holds(Alive,S0),
```

but we do not know whether the gun is loaded or not. What happens if we shoot Fred? There are two intended models: either the gun is loaded and Fred dies, or the gun is not loaded and Fred does not die. Chronological minimization prefers the latter one of these models, and concludes that Fred will not die. Thus an unintended conclusion is obtained.

To continue with the same example, suppose we are also told that Fred is dead in the situation resulting from the shooting and then waiting. Chronological minimization will then prefer a model where the gun is initially not loaded, Fred survives the shot, and then he dies as the result of waiting, since this alternative satisfies all the axioms and postpones all changes as late as possible.

We should note that in this last example, the original (pre-YSP) circumscription approach also fails to obtain the intended result. Our point is merely that the switch to simple chronological minimization does not help.

3.3 Modified chronological minimization

The problems with basic chronological minimization can be solved, as far as the persistence problem is concerned, by certain modifications to the approach. Sandewall [58] proposes three modifications to the original chronological minimization that was described above.

One of the counterexamples involved an incompletely specified initial state, suggesting that the method only works when the initial state is completely specified. This problem can be solved by changing the second condition in the preference order $\ll$ on models to read: (2) there is some point of comparison s_c such that the two models agree with respect to `Clip` <u>and `Holds`</u> for situations/timepoints earlier than s_c, and $\lambda p.\mathtt{Clip}(p, s_c)$ is a smaller set in the preferred model than in the other one. This is also in agreement with Shoham's variant of chronological minimization. Note how this handles the example in which we are told that Fred is originally alive, but we are not told whether the gun is loaded or not. The original chronological minimization preferred that the gun be unloaded so that the aliveness of Fred would persist; with the new definition, these two possibilities are not comparable.

Other counterexamples involve temporal explanation, where some of the axioms specify facts in the world at later points in time, and expects to make conclusions about the state of the world at earlier points in time. In order to handle such reasoning problems correctly, one must delimit the set of axioms to which chronological minimization is applied. Suppose Γ_ℓ is the set of action laws, and Γ_o is the set of observation axioms, i.e. those axioms specifying which properties hold at various times. Ordinarily, chronological minimization is applied to all of the axioms, so the set of selected models is defined as

$$Min(\ll, [\![\Gamma_\ell \cup \Gamma_o]\!])$$

i.e. the set of those classical models which are minimal with respect to the preference order $\ll$ defined above. In a modification called *filter preferential entailment*, the minimization is only applied to the causal axioms, and then the observation axioms are conjoined afterwards. The set of selected models is therefore defined as

$$Min(\ll, [\![\Gamma_\ell]\!]) \cap [\![\Gamma_o]\!]$$

For some Γ this will obtain a smaller set of selected models and therefore stronger conclusions. It is easily verified that in general,

$$Min(\ll, [\![\Gamma_\ell]\!]) \cap [\![\Gamma_o]\!] \subseteq Min(\ll, [\![\Gamma_\ell \cup \Gamma_o]\!]),$$

and also that this method does not have the property of consistency preservation.

Consider the example in which we are told that Fred dies after being shot, but we are not told whether the gun was originally loaded or not; recall that ordinary chronological minimization would not let us conclude that the gun was originally loaded even though this is the only available explanation for Fred's death. With the new approach, $Min(\ll, [\![\Gamma_\ell]\!])$ entails that the shooting kills Fred if the gun was loaded and not otherwise; when we conjoin the fact that Fred does in fact die, it will follow that the gun must have been loaded.

A third type of counterexamples to simple chronological minimization are obtained for nondeterministic actions, for example the action `Toss` of tossing a coin which was discussed in subsection 2.4. In order to deal properly with nondeterministic actions in the context of chronological minimization, Sandewall proposed to use an *occlusion operator* whereby one can explicitly specify whether a property is affected by an action. The action law for `Toss` would be written simply as

$$\texttt{Occluded(P,Result(Toss,}s\texttt{))},$$

and in general every action law should be written so as to imply the occlusion of all properties that *may* be influenced by the action. Syntactically this just means the reintroduction of the explicit implication of `Ab` (or `Ends`), but now occlusion is thought to apply even for a potential abnormality with respect to persistence, and not only for an actual one.

Approaches similar to chronological minimization have been explored by other authors as well. Pearl [46], for example, represents defaults as probabilistic statements with probabilities very close to 1. He then shows that one can get the right answer to the Yale shooting problem by making certain assumptions regarding conditional independence. Pearl argues that his logic of probabilistic independence gives chronological minimization "its operational and probabilistic legitimacy."

3.4 Causal minimization

Another approach, developed independently by Haugh [17] and Lifschitz [27] is that of *causal minimization*. This method modifies the original domain description in a manner that captures the notion of causality directly. The intuition here is that there is a crucial difference between the abnormality of a gun becoming unloaded while waiting, and the abnormality of Fred dying when shot with a loaded gun: there is a *cause* for the second while the first is totally arbitrary. For concreteness, we will discuss Lifschitz's version. He begins by stating *explicitly* that shooting causes the person to die:

```
Causes(Shoot,Alive,False),
```

and that the `Shoot` action has the single precondition of the gun being loaded:

```
Precond(Loaded,Shoot).
```

Also of course the initial situation is specified:

```
Holds(Loaded,S0),
Holds(Alive,S0).
```

Finally there are two domain-independent axioms which "interpret" the relation `Causes`. One axiom allows to conclude whether `Holds(`p`,Result(`A,s`))` on the basis of `Causes(`A`,`p`,`v`)` (where v provides the truth-value), `Precond(`p'`,`A`)`, and `Holds(`p',s`)`. The other general axiom states that if `Holds(`p`,Result(`A,s`))` differs from `Holds(`p,s`)` then it must be the case that `Causes(`A,p,v`)` with the appropriate v, and also the preconditions of A must be satisfied in s.

With this axiomatization both `Causes` and `Precond` are circumscribed with `Holds` varied. In the simplest case with only persistence, the extent of those predicates is then reduced to what is stated explicitly in the axioms. To see how it works, consider what will happen with the anomalous model of the Yale shooting problem, where the gun becomes unloaded during the waiting. In order to satisfy all the axioms, such a model must assign the value `T` to `Causes(Wait,Loaded,False)`. However `Causes(Shoot,Alive,False)` is still present since it is one of the axioms, although in this model the precondition of `Shoot` is not satisfied. This model is therefore less preferred than the intended one. It is easy to see that in all minimal models, the gun remains loaded and Fred dies.

Causal minimization is more restricted than situation calculus in terms of expressiveness, because of the use of the `Causes` predicate for specifying the effects of actions. There is no way, for instance, to express context-dependent effects; that is, we cannot say that an action causes one thing in one circumstance, and something else in another circumstance. Similarly there is no way to express action laws for nondeterministic actions.

Minimization of `Causes` and `Precond` will sometimes result in predicates that hold for additional cases besides those explicitly stated in the axioms.

One might hope that the minimization of `Causes` could provide a mechanism for constraint-based ramification, namely if domain constraints force additional cases of `Causes` besides those stated explicitly. Unfortunately this effect is not always obtained. A counterexample, due to Ginsberg, can be obtained from Hanks and McDermott's original version of the Yale shooting problem in which the gun was initially unloaded and there is a `Load` action that loads it:

`Causes(Load,Loaded,True).`

Suppose we know that in order to be breathing, you have to be alive (this is the domain constraint):

`Holds(Breathing,`s`)` $\rightarrow$ `Holds(Alive,`s`).`

If the minimization of `Causes` were to handle the ramification, then all minimal models should state that not only does shooting make Fred not alive; it also makes him not breathing:

`Causes(Shoot,Breathing,False).`

Unfortunately, there is another causally minimal model in which `Load` kills Fred, and where

`Causes(Load,Alive,False),`

`Causes(Load,Breathing,False).`

In this model, there is no situation in which the gun is loaded while Fred is alive; therefore, `Causes(Shoot,Breathing,False)` does not have to be added. By setwise minimization of the predicates, both models are selected.

In order to deal with ramification in causal minimization, one must therefore either record *all* of the effects of the actions in the action laws, or use the distinction between independent and derived properties.

One might also have hoped that the minimization of `Precond` could support the postdictive use of qualification: if an action, according to observations, fails to have the effects that it should have according to the action laws, then an additional precondition may be assumed in order to account for the non-applicability of the action. However it is not clear whether the concurrent minimization of `Causes` and `Precond` actually gives the intended results in such cases.

In summary, causal minimization has a number of difficulties. It is however very instructive by showing what kinds of difficulties a nonmonotonic logic can have with the various kinds of frame problems.

3.5 Nonnormal defaults and autoepistemic logic

Based on a comparison with truth maintenance systems, Morris [44] observed that the Yale shooting problem is correctly solved if one expresses the default frame axiom as a nonnormal default rule in Reiter's default logic [53]. The normal default rule used by Hanks and McDermott is replaced by

$$\frac{\mathtt{Holds}(p,s)\,:\,\neg\mathtt{Ab}(p,A,s)}{\mathtt{Holds}(p,\mathtt{Result}(A,s))}.$$

(A similar rule for $\neg\mathtt{Holds}(p,s)$ is also needed). Default logic is described in the article with that name by David Etherington in this Handbook. For the present purpose the following semiformal view is sufficient. A default rule of the form

$$\frac{\alpha\,:\,\beta}{\gamma}$$

means that if α is known and β is consistent with our knowledge, then one may conclude γ; the rule is nonnormal if β is different from γ. One builds up an *extension* of a default theory by starting with the axioms, and then applying the various default rules in some arbitrary order until all applicable rules have been applied. If any of the later default rules undermines an earlier one, then this ordering does not yield a valid extension.

Suppose now that the new default rule is combined with the same three domain axioms as were used by Hanks and McDermott, namely

$$\mathtt{Holds(Loaded,}s\mathtt{)} \rightarrow \neg\mathtt{Holds(Alive,Result(Shoot,}s\mathtt{))} \wedge \mathtt{Ab(Alive,Shoot,}s\mathtt{)},$$

$$\mathtt{Holds(Loaded,S0)},$$

$$\mathtt{Holds(Alive,S0)}.$$

The frame default rule is a schema representing all its ground instances. In the current example there are two ground instances that are of interest:

$$\frac{\mathtt{Holds}(\mathtt{Loaded},\mathtt{S0}) : \neg\mathtt{Ab}(\mathtt{Loaded},\mathtt{Wait},\mathtt{S0})}{\mathtt{Holds}(\mathtt{Loaded},\mathtt{S1})}$$

and

$$\frac{\mathtt{Holds}(\mathtt{Alive},\mathtt{S1}) : \neg\mathtt{Ab}(\mathtt{Alive},\mathtt{Shoot},\mathtt{S1})}{\mathtt{Holds}(\mathtt{Alive},\mathtt{S2})}$$

where `S1` is an abbreviation for `Result(Wait,S0)` and `S2` abbreviates `Result(Shoot,S1)`. If we initially apply the first default rule, its antecedent will be consistent, so we will conclude `Holds(Loaded,S1)`. Then, from the causal axiom it follows that Fred dies, and thus `Ab(Alive,Shoot,S1)`. At this point, the second default rule is disabled, so our extension is complete. This is the intuitive extension. If we initially apply the second default rule, its antecedent $\neg$`Ab(Alive,Shoot,S1)` will be consistent so we will conclude that Fred survives. The first default rule, however, will still be viable at this point ($\neg$`Ab(Loaded,Wait,S0)` is also consistent), so we conclude that the gun remains loaded, and thus from the causal axiom it follows that `Ab(Alive,Shoot,S1)`. Since this fact contradicts the antecedent of the previously applied default rule, this fails to be a valid extension. Therefore, the only extension is the one in which Fred dies.

Gelfond [11] later formulated the same solution in autoepistemic logic [42]. Recall that autoepistemic logic and default logic are intimately related [23]. Syntactically one then only requires a single change to McCarthy's default frame axiom, namely changing $\neg\texttt{Ab}(p, A, s)$ to $\neg\mathbf{L}\texttt{Ab}(p, A, s)$, where $\mathbf{L}$ is the belief operator in autoepistemic logic.

The approach described here has the advantage of being extremely simple. Unfortunately it does not do very well on the additional examples that we introduced in Sections 3.2 and 3.4. In the first temporal explanation example where we are not told whether the gun is loaded or not, `Ab(Alive,Shoot,S1)` is not provable, so $\neg$`Ab(Alive,Shoot,S1)` is consistent. Therefore `Alive` persists from `S1` to `S2`. The nonnormal default approach will mistakenly conclude that Fred will survive being shot. In the second temporal explanation example where we are then told that Fred is not alive in `S2`, the system will fail to have *any* extensions, since all the conclusions from the previous case remain and one of them has been contradicted.

The nonnormal default approach also does not handle constraint-based ramification. If we merely assume

$$\texttt{Holds(Breathing,}s\texttt{)} \rightarrow \texttt{Holds(Alive,}s\texttt{)}$$

without explicitly adding `Ab(Breathing,Shoot,`s`)` on the right-hand side of the causal axiom, then the system will again have no extensions since the default for the persistence of `Breathing` will not be disabled. The nonnormal default approach therefore seems to share the restrictions of both simple chronological minimization and causal minimization.

3.6 Varying `Result`

Recall that in our discussion of the default frame axiom, the straightforward method of minimizing abnormality was to circumscribe `Ab` with `Holds` varied. Baker [1] suggests that the shooting problem can be solved by varying the `Result` function instead of the `Holds` predicate during this circumscription. This is an important difference because in circumscription's partial order on models, model M_1 is preferred to model M_2 if the predicate being circumscribed has a smaller extent in M_1 *and* if the only other predicates and functions that differ between M_1 and M_2 are the ones that are varied.

The standard approach amounts to thinking of `Result(Wait,S0)` as a Herbrand situation, with circumscription being used to determine which properties hold in this situation. Baker uses instead a state-oriented semantics for situations, as discussed in subsection 2.1, and requires that for each consistent combination of properties, there is some situation in which these properties hold. Circumscription is then used to determine which of the situations might be the result of waiting in `S0`. To see why this works, consider the model in which waiting unloads the gun. This model has the abnormality `Ab(Loaded,Wait,S0)`. This abnormality can be eliminated by varying `Result` so that the result of waiting in `S0` is equal to some situation in which the gun is still loaded. Since this operation need not introduce any new abnormalities, the gun will remain loaded (and Fred will die) in all minimal models.

Baker and Ginsberg [2] later proposed a revised version of this approach, where they allow explicit representation of both situations (still with a state-oriented semantics) and of states themselves, and minimize abnormality over states rather than situations.

Baker's approach deals correctly with the persistence problem for deterministic actions, regardless of whether observations are given for earlier situations (prediction) or also for later situations (postdiction), as verified by Kartha [20]. It has problems with nondeterministic actions such as `Toss`. Since it operates by minimizing changes without any temporal preference, it can also deal with constraint-based ramification and with surprises, and for some specific simple examples it is known to give the intended results. The use of one single minimization mechanism for all of these purposes is both an advantage and a potential weakness – an advantage because of its simplicity, and a potential weakness because it is not clear whether the intended results are obtained when several of such phenomena occur in the same reasoning example.

3.7 Surprises

The problem of surprises was introduced in subsection 2.3 above. We consider now the case where surprises are allowed in order to avoid obtaining an empty set of intended models. The usual assumption with respect to surprises is that the set of surprises is to be minimized. This assumption is debatable. Suppose one model requires two surprises in order to obtain consistency for the given axioms, and another model manages with a single surprise which is not among the first two. If surprises are minimized setwise then neither of these models is preferred over the other one. Arguably the choice and combination of surprises ought to be minimized with respect to its total degree of unlikeliness, and in the absence of more detailed information it would be better to minimize the cardinality of the set of surprises rather than the set itself. However the case of setwise minimization is to our knowledge the only one that has been investigated.

Another immediate question is whether surprises ought to be dealt with by the *same* logical machinery as is used for the persistence problem, or whether separate preference mechanisms for persistence and surprises must be identified and then combined. If one should choose to minimize surprises by cardinality or weight then it is at once clear that separate solutions are needed. If one is satisfied with setwise minimization of surprises then both paths are open.

The approach due to Baker where `Result` is allowed to vary, has the effect of allowing surprises and minimizing them setwise. Filtered preferential entailment does not allow surprises, since it necessarily selects an empty set of models rather than assuming a surprise. Causal minimization may react to surprises by extending the `Causes` predicate, but this leads to problems if the same action is performed more than once – it will be inferred that the surprise is to be repeated as well. Original global (= non-chronological) minimization will in principle allow surprises but it has too many other weaknesses to be of any interest. Approaches based on default logic with nonnormal defaults have a chronological character, so to the extent that they allow surprises at all they tend to prefer later surprises over earlier ones, which presumably is not intended. The same holds of course for chronological minimization without filtering.

The other possibility is to minimize surprises separately. Lifschitz and Rabinov [31] have proposed such a solution, built on top of causal minimization. In the extended method, nothing can change without a cause unless a *miracle* occurred; the new predicate `Miracle` is circumscribed at a lower priority than `Causes`. Therefore in the above example, Lifschitz and Rabinov's system would not mistakenly conclude that waiting *caused* the gun to unload; instead, it would simply conclude that a miracle must have happened.

Morgenstern and Stein [43] take another approach to this problem. They use a logic with explicit integer time and with an explicit treatment of actions or events, allowing also events that are not directly caused by agents. This brings several advantages; in particular they can represent chains of causation from one event to another. Unexpected change is explained by assuming additional events, rather than by assuming arbitrary changes of property values (miracles). This is a way of imposing some restrictions on what kinds of unexpected changes can be allowed, but it assumes that the repertoire of possible surprises has been defined in advance in terms of events. For the stolen car scenario of Kautz, for example, one would need a `Steal` action which can be used as explanation for the surprise, even if no steal action is explicitly known in the scenario. The minimization of the set of actions is achieved by circumscribing "unmotivated" actions, where roughly speaking an unmotivated action is one that is not triggered by a previous action.

A general issue with respect to surprises is how the repertoire of possible surprises is specified. This may strike one at first as a strange question since the very point with surprises is that they represent exceptions to the standard situations. The necessity of imposing constraints has been illustrated by the three lightbulbs scenario of Struss and Dressler [64]: suppose one battery is connected to three lightbulbs in parallel; normally the battery is active, but it is observed that only one lightbulb is shining. One explanation is that the other two bulbs are broken; another explanation which also contains two "surprises" is that the battery is flat but one lightbulb is shining anyway!

In the approach of Lifschitz and Rabinov it is straightforward to restrict the repertoire of surprises by imposing constraints on the predicate `Miracle`. In the approach of Morgenstern and Stein the repertoire of surprises is defined by the available set of events, but since the same minimization of events is used for dealing with both causality and surprises it may be difficult to obtain exactly the right repertoire for each purpose individually. In either case, certain undesirable surprises may be excluded using global state constraints. However since the systematic methodology has not yet been applied to this class of problems, it is hard to establish exactly what each proposed approach is able and not able to do.

3.8 Explanation closure

The most straightforward way of obtaining the set $\Sigma(\Gamma)$ of intended models for the scenario description Γ is to introduce an additional set of axioms Φ and to approximate (or even define) $\Sigma(\Gamma)$ as $[\![\Gamma \cup \Phi]\!]$. It has traditionally been observed that under simplifying assumptions (persistence only, simple action laws) Φ can be obtained mechanically from the set of action laws in Γ together with the vocabularies of action symbols and property symbols. For each action A, given the axioms that specify its effects, one can identify those properties that are not affected by the action, and for every such property p one can construct *frame axioms* of the form

$$\psi(s) \;\rightarrow\; (\texttt{Holds}(p,s) \;\equiv\; \texttt{Holds}(p,\texttt{Result}(A,s))),$$

saying that if the action A is performed in a situation satisfying the condition ψ then the property p is not affected. This syntactic transformation on the axioms has been considered problematic only because the resulting number of frame axioms will be quite large if the number of actions and properties is large, presumably causing difficulties for any theorem-proving system or other implementation of the logic.

One may also ask whether or under what circumstances the transformation resulting in Φ is at all feasible. This question does not seem to have been addressed, probably because of an assumption that practical action laws have a simple structure so that also the construction of the frame axioms is trivial.

Haas pointed out in 1988 [15] that an adequate and much more compact set Φ of auxiliary axioms may be constructed in a different way by considering the whole set of action laws at once, and not for one action at a time. Basically for each property p one constructs an *explanation closure* axiom of the form

$$\begin{aligned}&(\texttt{Holds}(p,s) \;\not\equiv\; \texttt{Holds}(p,\texttt{Result}(A,s)))\\ &\qquad \rightarrow\;\; (\psi_1 \wedge A = A_1) \vee (\psi_2 \wedge A = A_2) \vee ...(\psi_k \wedge A = A_k)\end{aligned}$$

enumerating *all* the actions A_i and corresponding preconditions ψ_i where p may have changed its value between s and $Result(A,s)$. Interpretations that violate the assumption of persistence will fail to satisfy this axiom. This approach of *explanation closure* has been further investigated by Schubert [61] and Reiter [54].

If the original action laws in Γ have a sufficiently simple structure then it is straightforward to construct the explanation closure axioms automatically. The transformation is a syntactic one, and as such it is sensitive to certain distinctions that are semantically irrelevant. Consider again the action of tossing a coin, so that its position after the action is either heads or tails regardless of its position before the action. If its action law is written as

$$\texttt{Holds(P,Result(Toss,}s\texttt{))} \vee \neg\texttt{Holds(P,Result(Toss,}s\texttt{))},$$

then although semantically a tautology, it does contain syntactic information which may be read as saying that the property `P` is exempt from persistence. The syntax-level derivation of the explanation closure axioms may be sensitive to this distinction, and generate a rule where `Toss` occurs as one of the A_i. This is the syntactic counterpart of the occlusion operator or the explicit implication of abnormality.

The method of explanation closure is fairly straightforward if it is only used to address the persistence problem. For the qualification and ramification problems, Schubert [61] proposes that the explanation closure axioms can and should be written manually, and that it is then quite feasible to incorporate qualifications and ramifications of actions in the axiomatization. Very recently Lin and Reiter [32] have showed a systematic, syntactic way of transforming action laws and state constraints to a corresponding axiomatization which can be used monotonically and which seems to realize qualification and ramification.

Explanation closure is often described as a monotonic approach to the frame problem. Notice however that it is only monotonic after the set Φ of closure axioms has been constructed and added in. On the other hand if the original set of axioms Γ is extended by adding more effects for previously present actions, so that Φ must be revised accordingly, then the set of models $[\![\Gamma \cup \Phi]\!]$ does not change monotonically.

Explanation closure does not of course allow for surprises, but the method of Lifschitz and Rabinov can be combined with explanation closure as well as with causal minimization.

3.9 Meta level time

The previous approaches allow each interpretation in the logic to contain its own time structure and to define whether a certain proposition is true or false at a certain time. Another possibility is to let the time structure be the primary one, and to associate a set of atemporal models or a set of atemporal propositions with each point in the time structure. The step from one timepoint to its successor(s) is then defined in terms of the *update* of the set of models or the set of propositions. If s and $\texttt{Result}(A, s)$ are two successive nodes in the time structure, the update method specifies how the formulae or states associated with $\texttt{Result}(A, s)$ are to be obtained from those associated with s. Such an update method allows one to successively construct descriptions of future situations for different choices of action. Since time is external to interpretations as well as to propositions in this case, it will be referred to as *meta-level time*, in contradistinction to the previously described *object-level time* methods.

Ginsberg and Smith [13, 14] proposed to address a broad range of frame problems using a meta-level time approach where situations are associated with sets of formulae (not sets of states), and where update is based on minimal change in this set of formulae. They argued that their approach solves all the main varieties of the frame problem, i.e. the problems of persistence, qualification, and ramification. A similar approach was pursued by Rao and Foo [51].

Based on common-sense examples, Winslett [67] and later Brewka and Hertzberg [5] argued that update of the set of formulae is not a satisfactory approach. Winslett proposed to associate situations with sets of states, and to define update in such a way that if $M(s)$ is the set of states associated with a situation s, then $M(\texttt{Result}(A, s))$ is defined as

$$\bigcup_{r \in M(s)} \mathit{Incorporate}(r, \Psi \cup \mathit{Post}(A, r))$$

where Ψ is a set of domain constraints, $\mathit{Post}(A, r)$ are the postconditions of the action A in the state r, and $\mathit{Incorporate}(r, \Gamma)$ is the set of those states which satisfy Γ and which minimize the difference from r. Modifications of Winslett's approach have been proposed by Peppas and Wobcke [49], by Brewka and Hertzberg [5], and by Zhang and Foo [68]. In all cases the update is defined statewise so that

$$M(\texttt{Result}(A, s)) = \bigcup_{r \in M(s)} \Phi(r, A)$$

with some specific update function Φ. The purposes of their modifications is to deal more appropriately with non-deterministic actions and with ramification induced by state constraints. These works are also related to the large literature on data base update and belief revision, which are outside the scope of the present article.

Methods where the update function is defined statewise can be translated in a straight-forward fashion into methods of object-level time, and if the update function Φ is defined by minimization according to a preference relation on states, then a corresponding preference relation will apply for object-level time.

The use of meta-level time approaches instead of the more conventional (from the point of view of logic) object-level time, is not motivated by greater expressiveness or deductive correctness, therefore. Their particular attraction is that their formal definition is close to certain actual implementations, namely those using a single database containing all the properties that are true in some situation, in the tradition of the STRIPS system of Fikes and Nilsson [9]. The resulting perspective is biased in favor of "forward" temporal reasoning such as prediction and planning. For "backward" temporal reasoning such as postdiction the meta-level time approach appears to be inconvenient. Suppose the given information about an earlier situation is incomplete, but it is complemented with information about the state of the world after actions have been performed. The definition of $M(\mathtt{Result}(A, s))$ as the union of all statewise updates does not preserve the information about which later states were derived from which earlier states, and therefore makes it impossible (or at least inconvenient) to obtain conclusions about what must have been the state of the world at the earlier point.

4 The systematic paradigm

It should have been clear from the overview in section 3 that the methodology of common-sense examples has resulted in a somewhat chaotic development: logics, examples, and counter-examples have been confronted, and it has not always been clear which property of a logic was to be given the credit for its successes, or the blame for its failures. It has not even always been clear what was a success or a failure in terms of the proposed reasoning examples. The most important weakness, however, is that various logics have been proposed to be "solutions to the frame problem" just on the basis of intuitions and a few examples, only to be very quickly refuted by the arrival of more examples.

The systematic paradigm attempts to structure the problem area a bit better. Rather than allowing the definition of $\Sigma(\Gamma)$ to rely on intuitions that are independently applied for each particular choice of Γ, it requires a formal definition of $\Sigma(\Gamma)$ as the set of intended models. The task for any proposed nonmonotonic logic is then to correctly identify this $\Sigma(\Gamma)$.

In fact one needs not just one but several such functions Σ_i. Depending on whether surprises are allowed or not, whether ramifications are allowed or not, and so forth, and also depending on exactly how surprises are to be chosen and minimized or how side-effects due to ramification are constructed, one obtains different variants Σ_i of the function Σ. These variants may be more or less good approximations of one person's *intuitive* notion of intended conclusions; this consideration is intentionally and necessarily left outside the topic of formal study and precise results.

4.1 Definitions of intended models

Research in this area is presently just in its beginning, but some directions are already clear. In one aspect, contributions differ with respect to how the set of intended models is specified.

Lin and Shoham [33] define the set of intended models $\Sigma(\Gamma)$ as the classical models of an extended set of formulae, $[\![\Gamma \cup \Phi]\!]$. In the case of simple persistence (without qualification or ramification) Φ can be chosen as the frame axioms or the explanation closure axioms. However their approach also allows the case where Γ contains domain constraints in which case Φ must be made correspondingly weaker; the general requirement is that $\Gamma \cup \Phi$ fully determines the effect of any action in any situation. Their approach is therefore intrinsically limited to deterministic actions. – The definition of $\Sigma(\Gamma)$ in terms of classical models for an extended set of axioms has also been used by Lifschitz [30] and by Reiter et al [54, 32].

A different approach is to introduce a separate language for expressing action laws and action sequences. Such languages are regularly used in practical A.I. systems which reason about actions and change. In the present context they are used for defining the intended histories and thereby the intended models for a given scenario. This requires first of all that a formal semantics has been defined for the language. This is often not the case for the practical ones, although Pednault [47, 48] defined the semantics of his ADL language in an approach of meta-level time where a set of states was assigned to each situation.

Gelfond and Lifschitz[12] were the first to use this approach by introducing an *auxiliary language* $\mathcal{A}$ with its own syntax and semantics, which is especially designed for defining the intended conclusions from a given scenario description. The basic version of $\mathcal{A}$ is fairly restricted and corresponds to the following types of formulae in situation calculus, and in the terminology of the present article:

$$\pm\texttt{Holds}(\pi, \sigma)$$

$$\bigwedge_i \pm\texttt{Holds}(\phi_i, s) \quad \rightarrow \quad \pm\texttt{Holds}(\pi,\texttt{Result}(A, s))$$

where π and all ϕ_i are property names, $\pm$ represents presence or absence of negation, and σ is either S0 or is formed from S0 by any number of layers of the form $\texttt{Result}(A, \sigma)$.

Sandewall [60] defines Σ using an *underlying semantics* based on a notion of possible trajectories that characterize the various ways of performing each action. This may be understood as a way of having an auxiliary language with only its semantics but without any syntax.

A final approach is to specify the intended conclusions through a number of *postulates* for the relation between Γ and $\Sigma(\Gamma)$, or between $[\![\Gamma]\!]$ and $\Sigma(\Gamma)$. Del Val and Shoham [8] use the postulates of Katsumo and Mendelzon [21] as the basis, and Peppas and Wobcke [49] use an extended set of postulates for epistemic entrenchment as the basis for the correctness proof for their system which uses the approach of meta-level time.

In comparing these approaches to the definition of $\Sigma(\Gamma)$ there are two important aspects. It is important that the definition of the function Σ shall be intuitively convincing since it is intended to capture our notions of common sense. In particular, it must both be plausible in the abstract, and it must allow us to check it out with some concrete common-sense examples.

Another important aspect is what class of scenarios can be expressed in the definition of Σ. For example, the $\mathcal{A}$ language only allows deterministic actions. Two different strategies are being tried here. Sandewall's underlying semantics is defined to allow a relatively broad class of action laws, including nondeterministic actions and actions with extended duration, and it is used for assessments of the form "for which subset of these scenarios and reasoning tasks does a given logic obtain the intended conclusions?". The $\mathcal{A}$ language, on the other hand, was originally defined in a restricted manner and has then been extended to also allow concurrent actions [3]. Each time the auxiliary language is extended, it becomes necessary to redo the assessments of those logics that had been assessed with the more restricted language.

Different definitions of intended models may be intertranslatable. Ideally, one should have well-defined translations between them, so that each assessment of a proposed logic can be done with whatever definition of Σ is the most convenient one for the purpose at hand.

4.2 Which scenarios and reasoning tasks are addressed

Different contributions also differ with respect to which classes of scenarios and reasoning tasks are being analyzed. The common baseline consists of scenarios with sequential, deterministic actions each taking a single timestep, and where persistence is assumed but there are no surprises, qualifications, or ramifications. In some cases there is also a restriction that all observations refer to time zero or situation `S0`. Relative to this baseline, Sandewall has considered scenarios allowing nondeterministic actions with extended duration in time. Extensions allowing ramification have been studied by Lin and Shoham [33], Lin and Reiter [32], and Peppas and Wobcke [49]. The extension to concurrent actions has been studied by Lin and Shoham [34] and by Baral and Gelfond[3].

4.3 Assessment of applicability

Once one or more Σ_i have been defined, one can analyze some of the nonmonotonic logics that have been described above, or some new ones. Each logic must then be formulated as a function S from a set of axioms to a set of models, and one can ask the following question:

- For a given logic S and a given Σ_i, does the logic obtain exactly the intended conclusions, i.e. is $S(\Gamma) = \Sigma_i(\Gamma)$ for all Γ?

If Σ_i has been defined for a broader class of tasks than the given logic can handle (so that the answer is no), then the question must instead be:

- For a given logic S and a given Σ_i, identify a set of reasoning tasks $\mathcal{Z}$ such that $S(\Gamma) = \Sigma_i(\Gamma)$ for all $\Gamma \in \mathcal{Z}$.

The set of reasoning tasks $\mathcal{Z}$ may then involve restrictions such as "if all actions are deterministic" or "if the initial state of the world is completely specified". For some purposes one may also be interested in methods which fail to obtain all intended conclusions, as long as no unintended conclusions can be reached. The above mentioned questions can then be modified as follows:

- For a given logic S and a given Σ_i, does the logic obtain at most the intended conclusions, i.e. is $S(\Gamma) \supseteq \Sigma_i(\Gamma)$ for all Γ?

- For a given logic S and a given Σ_i, identify a set of reasoning tasks $\mathcal{Z}$ such that $S(\Gamma) \supseteq \Sigma_i(\Gamma)$ for all $\Gamma \in \mathcal{Z}$.

In order to obtain a complete picture of the range of applicability of a given model selection function S one is also interested in upper bounds i.e. in results of the following form

- For a given logic S and a given Σ_i, identify a set of reasoning tasks $\mathcal{Z}'$ such that if $\Gamma \notin \mathcal{Z}'$ then $S(\Gamma) \neq \Sigma_i(\Gamma)$,

although when we return to this topic below we shall see that the formulation must be modified slightly in order to make good sense.

4.4 Assessments of representation

An additional assessment issue is concerned with the change of representation. This topic has so far not been analyzed in a systematic fashion, to our knowledge, but it is clearly important and it sometimes appears in the informal discussion about the pros and cons of different approaches. For example, it can be argued that the formal system does not have to handle nondeterministic actions, since such action laws can always be rewritten using auxiliary properties as described in subsection 2.4. It is then necessary to verify, for each proposed nonmonotonic logic or method, whether it still behaves as intended after the proposed rewrite. For example, the rewrite defined above used an axiom of the form

$$\texttt{Holds(P,Result(Toss,}s\texttt{))} \equiv \texttt{Holds(R,}s\texttt{)},$$

which works fine with chronological minimization under the first of the corrections described in subsection 3.3. However if it is instead written on the form

$$\texttt{Holds(P,Result(Toss,}s\texttt{))} \equiv \texttt{Holds(R,Result(Toss,}s\texttt{))}$$

(which intuitively would not seem to make much difference), then the same variant of chronological minimization does not give the intended conclusions.

Another issue occurs with respect to the use of metric time. Although many of the contributions in this area use the situation calculus, others have used the integers as timepoints and as the second argument of `Holds`. In this case one needs separate statements as to which actions are performed at which time; alternative histories are represented in different models and never within the same model. This approach has the advantage of being able to express actions with extended duration. Alternatively, the duration of actions may be represented in the situation calculus with Herbrand semantics by allowing each situation `Result(A,s)` to contain the information about the starting time and the ending time of the action that led to it from s. How does the expressiveness of that approach compare with the expressiveness of the approach with explicit integer time? Also, does a particular logic which correctly obtains the intended models for one representation also obtain the intended models for the other representation? Those are some other examples of questions for representation assessment.

The paper by Lin and Shoham [33] in 1991 was the first one to argue seriously for the systematic approach. All research along those lines is very recent, therefore, as is also reflected by this brief overview. We expect that much more will happen in this area in the years to come.

5 Assessments of entailment methods for the persistence assumption

This section will use Sandewall's approach and summarize the results which have been obtained there. A more complete exposition can be found in a recent monograph[59], henceforth referred to as "the book".

5.1 Base logic

Each of the methods that were described in section 3 is characterized by its own model selection function S. In order to characterize the range of applicability of each S it is natural to first define a common and fairly broad *base logic* i.e. a logic in the conventional sense of a syntax and a Tarskian semantics. Ideally the base logic should be more general than the capabilities of any of the selection functions, so that the applicability of each of them can be characterized in terms of restrictions on the base logic.

We will therefore use a base logic which uses explicit metric time, with the non-negative integers as the time domain, and which allows non-deterministic actions and actions with extended duration in time. The effects of an action are therefore not restricted to the state of the world when the action begins and ends; they may also include the intermediate states while the action is being performed. The following notation will be used: $[t,t']p$ means that the atemporal proposition (property) p holds at each timepoint in the closed interval from time t to time t'. $[t,t]p$ will be abbreviated as $[t]p$, and therefore corresponds to $\mathtt{Holds}(p,t)$ except that the second argument is now an integer timepoint and not a situation. Since timepoints, unlike situations, do not "contain" actions, it is necessary to make separate statements about which actions are performed at which time. We write $[s,t]A$ to express that the action A is performed during the interval starting at time s and ending at time t, where $s < t$. The pre- and postconditions of the action apply at times s and t, respectively. Notice that s stands for "starting time" and not for "situation".

Furthermore we assume that all properties have the form $f \hat{=} v$ expressing that a *feature* f is being assigned a *value* v from the value domain for f. For example the momentary setting of a traffic light may may be a feature with at least three possible values `red`, `green`, and `amber`, and possibly also combinations or blinking variants of these. The statement "the color of box 5 is green at time 620" would be expressed as

$$[620]\,color(\mathtt{\#5}) \hat{=} \mathtt{Green}$$

corresponding to the situation-calculus expression

$$\mathtt{Holds}(color(\mathtt{\#5}) \hat{=} \mathtt{Green}, s_{620}).$$

The value of the term $color(\mathtt{\#5})$ is a feature, and `Green` is a member of the value domain for that feature. Even inherently two-valued properties such as "alive" or "loaded" will be considered as features with the binary value domain $\{\mathtt{T}, \mathtt{F}\}$, for uniformity.

The resulting logic is essentially a two-sorted first-order logic, with "timepoints" and "physical objects" as the two sorts. In addition there is the type of "features" and one or more types for the value domains of features. These latter types are second class in the sense that we will not use quantification over them. All domains are held fixed across interpretations.

Most scenarios for reasoning about actions and change need an intermediate between propositional logic and predicate logic: action laws and state constraints are best formulated in predicate logic terms and apply regardless of the choice of object domain, but each specific scenario defines a fixed and finite object domain. This is often expressed in terms of axioms for "unique names" and a "closed world" whose effect is to fix the same object domain across all models. In each specific scenario the deduction problem is therefore essentially propositional rather than first-order. Here, since our treatment is entirely model-theoretic, the same purpose is achieved by explicitly fixing object domains across models.

The following are those formal details in the chosen logic which may be required for the sequel. For the full definitions and for additional motivations about the choice of conventions, please refer to the book.

The temporal prefix may use an integer timepoint such as 620 or a timepoint constant symbol such as $\boldsymbol{t}_2$, or an interval between two such entities. Terms such as `#5` refer to specific physical objects, and are analogous to integer timepoints in that they are the same across all interpretations. The syntax also allows object constants, whose values are specified in each interpretation, e.g.

$$[620, \boldsymbol{t}_2]\, color(\boldsymbol{o}_5) \mathrel{\hat{=}} \mathtt{G}.$$

Actions are expressed using action symbols which may have arguments, e.g. as in

$$[530, 560]\, Paint(\boldsymbol{o}_5, \mathtt{G}).$$

In such a case the same action does *not* apply over subintervals or superintervals of $[530, 560]$. Ordering relations between timepoints are expressed in the natural way, for example as in $\boldsymbol{t}_2 > 620$.

An *action designator* is a term formed from an action symbol and its arguments, where all arguments are either object names (e.g. `#5`) or feature values (e.g. `G`), but not object constants such as $\boldsymbol{o}_5$. It is assumed that different object names refer to different objects, and that each object has an object name. For object constant symbols there is no similar uniqueness assumption.

The surface semantics is straightforward: for a given object domain $\mathcal{O}$, an interpretation I is a pair $\langle M, R\rangle$, where M is a valuation that assigns values to temporal and object constants independently of time, and R (analogous to `Holds`) assigns a value to each combination of feature and timepoint. Actions are not represented in the surface semantics.

In order to capture and make use of the fixed object domain, object names are used both in the syntax and in the semantics. A valuation maps object constant symbols to object names. A feature is a term formed by a feature symbol and its arguments, where all arguments are object names. Thus the expression $color(\texttt{\#5})$ is iself a feature. The expression $color(\boldsymbol{o}_5)$ is not a feature, but if the value of $\boldsymbol{o}_5$ is #5 then the value of the feature expression $color(\boldsymbol{o}_5)$ is the feature $color(\texttt{\#5})$.

The set of all features will be denoted $\mathcal{F}$. A *state* r is a mapping from features to feature values in the appropriate domain for each feature. The set of all states will be denoted $\mathcal{R}$. I will write $R(t)$ for $\lambda f.R(f,t)$, so $R(t)$ is a state for each non-negative t.

5.2 Underlying semantics: the game

The underlying semantics is intended to capture the intuition of a world where actions are performed over intervals of time, features have values at points in time, and inertia prevails so that features retain their value over time unless there is a positive reason why they must or may change it, for example as the direct result of an action. The underlying semantics is defined in two steps, first using a *game* between an ego and a world, and then using a formal characterization of the actions that the ego may invoke during the game.

For the purpose of the game, the ego is understood as the "knowledge level" [45] of an autonomous agent or robot, and the world is understood as the combination of the physical robot vehicle and the robot's environment. The ego and the world can therefore be assumed to communicate in terms of discrete descriptions of the environment, and invocations of actions which the ego requests and the world performs.

Formally, the ego and the world interact over *finite developments*, which are tuples $\langle \mathcal{B}, M, R, \mathcal{A}, \mathcal{C} \rangle$. Here $\mathcal{B}$ is a finite set of timepoints (integers) whose largest member n is referred to as the *now*. M is a valuation like in interpretations. R is a history of the world from time 0 to time n, and formally a mapping from $[0, n]$ to $\mathcal{R}$. $\mathcal{A}$ is a set of actions which have been completed at time n, i.e. a set of triples $\langle s, E, t \rangle$ where $s < t \leq n$ and E is an action designator. Finally $\mathcal{C}$ is a set of actions which have been started but not completed at time n, represented as pairs $\langle s, E \rangle$ where $s \leq n$. (s stands for "start" and not for "situation" here. Both s and t are timepoints).

The interaction between the ego and the world is realized as a game where the two players alternate, and the moves consist of sucessively modifying and extending a finite development which serves as the "board" of the game. If $\mathcal{J}$ is the domain of all finite developments, then an ego $\mathbf{K}$ is formally a mapping $\mathcal{J} \to \mathcal{J}$ specifying what move the ego does in each possible case, and a world $\mathbf{W}$ is formally a subset of $\mathcal{J} \times \mathcal{J}$ where $\forall J \exists J' [\mathbf{W}(J, J')]$, which specifies nondeterministically what move the world does in each case that may arise.

The moves of the ego must leave the first and third component of the development unchanged, which means in particular that the ego does not change the setting of "now". The exact definition of the ego and the world allows a fairly broad class of behaviors for the ego and the world, and then a harmless restriction to "typical" behaviors is added. It is shown that for the purpose of defining the set of intended models, it is no restriction to assume that "typical" behaviors are used.

Typically, then, the move of the ego is to add an element $\langle n, E\rangle$ to $\mathcal{C}$, meaning to invoke an action starting at the present now-time n, or to do nothing. The moves of the world are typically to increase *now*, to extend R correspondingly, and to terminate actions. If the ego has invoked the action $\langle n, E\rangle$ by adding it to $\mathcal{C}$, then the world can terminate it at time n' by resetting $\mathcal{C}$ to the empty set, and adding $\langle n, E, n'\rangle$ as an additional member of $\mathcal{A}$.

What has been described now is the definitions for sequential execution of actions. Concurrent actions require a slightly more elaborate definition.

5.3 Underlying semantics based on trajectories

Although the world is in the most general case a relation over $\mathcal{J} \times \mathcal{J}$, one needs to impose some more structure on it in order to obtain the results. We define worlds in terms of a *trajectory semantics*. A *trajectory* for a set $F \subseteq \mathcal{F}$ of features is a finite sequence $\langle r'_1, ...r'_k\rangle$ where $k \geq 1$ and each r'_i is a partial state assigning values to the features in F.

Each world is characterized as a pair $\langle \texttt{Infl}, \texttt{Trajs}\rangle$, where $\texttt{Infl}$ (influenced features) and $\texttt{Trajs}$ (trajectories) are functions such that if E is an action designator and $r \in \mathcal{R}$ is a state, then $\texttt{Infl}(E, r) \subseteq \mathcal{F}$, and $\texttt{Trajs}(E, r)$ is a non-empty set of trajectories for $\texttt{Infl}(E, r)$.

Informally speaking, $\texttt{Trajs}(E, r)$ is the set of possible ways that the action can be realized in the world. A world description $\langle \texttt{Infl}, \texttt{Trajs}\rangle$ specifies a world $\mathbf{W}$ with the following behavior in the game. If the ego has invoked the action E at time s by adding the pair $\langle s, E\rangle$ to the $\mathcal{C}$ component of the development, then the world chooses an arbitrary member $\langle r'_1, ...r'_k\rangle$ of $\texttt{Trajs}(E, R(s))$. The world's move is to modify the current partial development by (1) adding $s+k$ as a member of $\mathcal{B}$, thereby making it the new "now" time; (2) extending the history component R by defining $R(s+i) = R(s) \oplus r'_i$, for $1 \leq i \leq k$. Here the symbol $\oplus$ represents "override" so $[u \oplus v](f)$ equals $v(f)$ when it is defined, otherwise $u(f)$; (3) adding the completed action $\langle s, E, s+k\rangle$ to the $\mathcal{A}$ component of the development; (4) resetting the $\mathcal{C}$ component for ongoing actions to the empty set.

After this it is the ego's turn to make its move, which it does in the new situation where *now* is $s + k$. Like before, the described procedure works for sequential actions, and otherwise a more elaborate definition is required.

Complete (infinite) developments are obtained as the limiting case of the game when time goes to infinity, and are characterized by $R(t)$ being defined for all non-negative t.

5.4 Chronicles

A scenario will not be described as a single set of logic formulae, but as a tuple $\langle \mathcal{O}, \Pi, \text{SCD}, \text{OBS} \rangle$, where $\mathcal{O}$ is a finite set of objects $\{\texttt{\#1}, \texttt{\#2}, ...\texttt{\#}k\}$ that are used in interpretations e.g. when constructing features. The other elements in the tuple are sets of formulae. Π is a set of "laws" characterizing the effects of actions, and is in fact an exhaustive description of `Trajs` in logic. SCD is a schedule i.e. a set of formulae characterizing the actions, chosen among the following two simple kinds:

- Action statements of the form $[s,t]\,A(...)$,
- A disjunction of one or more formulae of the form $s < t$ or $s \leq t$,

where A is an action name, e.g. *Paint* in the example above, and the arguments may be object names, object constant symbols, or feature values. In this way it is possible to specify what actions there are, and what are the values or constraints on the timing of each action. Finally OBS is a set of observation statements, which can be chosen as any formulae not containing action statements. A scenario description of this kind will be called a *chronicle*.

I will write $\Pi(\text{SCD})$ for the result of replacing each action statement in SCD by the effects specified by the laws in Π. For example the effect of loading the gun in the Yale shooting problem could be expressed as the following formula in Π:

$$[s,t]\,Load \Rightarrow [t]\,l \hat{=} \texttt{T}$$

saying that if the gun-loading action takes place over the interval from s to t, then the feature for the gun being loaded is true (T) at time t. If SCD has the member $[4,6]\,Load$, $\Pi(\text{SCD})$ would then in its place have the member $[6]\,l \hat{=} \texttt{T}$. From a formal point of view, action statements are considered to be a separate language, and action laws in Π are rules for translating that language to the main language. The truth of SCD in an interpretation $\langle M, R \rangle$ is not defined, since actions are not included in interpretations, but the truth of $\Pi(\text{SCD})$ is defined.

5.5 Intended models

The correctness of a logic was defined above in terms of equality between the sets of intended and selected models, and we can now finalize the definition of the intended models. If a chronicle $\Upsilon = \langle \mathcal{O}, \Pi, \text{SCD}, \text{OBS} \rangle$ is given, then the set of *intended models* $\Sigma_{\mathbf{IA}}(\Upsilon)$ is defined using the set of infinite developments obtained as follows. Select an arbitrary world $\mathbf{W}$ which is exactly described by Π, and select also an arbitrary ego $\mathbf{K}$, an arbitrary initial state r_0 using $\mathcal{O}$ as its object domain, and an arbitrary initial valuation M_0. Generate all possible developments which can be obtained in games between **W** and **K**, starting in the finite development $\langle\{0\}, M_0, \{0 \mapsto r_0\}, \varnothing, \varnothing\rangle$. Restrict the resulting set of complete developments to those where all formulas in $\Pi(\text{SCD}) \cup \text{OBS}$ are satisfied, and where the set of actions $\mathcal{A}$ in the development equals the set of actions stated in SCD, always using the valuation component M to relate constant symbols to their values. Let W be the set of those developments, and define $\Sigma_{\mathbf{IA}}(\Upsilon)$ as the set of those interpretations $\langle M, R\rangle$ such that $\langle \mathcal{B}, M, R, \mathcal{A}, \mathcal{C}\rangle \in W$ for some $\mathcal{B}$, $\mathcal{A}$, and $\mathcal{C}$. $\Sigma_{\mathbf{IA}}(\Upsilon)$ is the set of intended models.

The underlying semantics in terms of the ego/world game and the trajectory description of actions is intended to capture the essential aspects of actions in a world with a qualitative level of description where inertia (persistence) applies. This assumption is built in through the use of the override operator in the trajectory semantics, whereby all features which are not included in $\texttt{Infl}(E, R(s))$ are held constant during the execution of the action E starting at time s. The intended models $\Sigma_{\mathbf{IA}}(\Upsilon)$ do not allow surprises or qualifications, and ramifications are also not accounted for unless they are made explicit in the trajectories for each action.

5.6 The ontological and epistemological taxonomy

In order to assess the range of applicability of various non-monotonic logics, it is necessary to have a common "coordinate system": a taxonomy of scenario descriptions whereby one can specify where a logic is applicable and where it is not. I propose to organize the taxonomy using a distinction between main characteristics and sub-characteristics, and on the other hand a distinction between ontological and epistemological characteristics. "Characteristics" are then various features of the scenario description; I use this term since the word "feature" itself is used for a category in the logic.

Ontological characteristics refer to properties of the world that is described in the scenario, for example whether actions are non-deterministic or not. Epistemological characteristics refer to the character of the description, for example whether there are any observations for time later than zero.

5.6.1 Main characteristics

I will characterize classes of worlds using a set of bold-face letters, where each letter indicates the presence of a particular characteristic. Basic inertia or the classical frame problem is denoted as **IA**, where **I** stands for inertia as such, and **A** represents "alternative results": the results of an action are conditional on the starting state, for example as when firing the gun in the Yale shooting problem. **C** represents that concurrent actions are allowed, **L** that actions may have delayed effects (resulting changes that occur after the period of the action itself), etc.

Each combination of such characteristics requires its own definition of the set of intended models, and the definition of $\Sigma_{\mathbf{IA}}(\Upsilon)$ for the case of simple inertia is therefore only the simplest case. The status of ramifications needs to be made more precise. The trajectory semantics defined above corresponds to the ontological family **IAD**, where **D** represents dependencies between features. The simpler case **IA** is obtained by imposing the following restriction on the trajectory semantics. For each action there must be a *range* of influence for the action, consisting only of those objects which occur as arguments in the action designator. The result of the action may only depend on features all of whose arguments are in the action's range, and only those features can have their values changed by the action.

The **D** characteristic is related to the informal concept of ramification that was defined in section 2.3 above. Simple cases of ramification, such as "if *alive* ceases to be true, then *dead* ceases to be false", or "if *alive* ceases to be true, then *walking* ceases to be true", are constraints between features of one and the same object, and may be analyzed as a simple extension of the restricted **IA** family. However dependencies such as "if a married man is killed, then his wife becomes a widow" involve features of another object than what is specified as an argument of the action, and are therefore characteristic of the broader **IAD** family.

Some of the proposed logics are also restricted by epistemological constraints, so the taxonomy must be able to characterize them as well. I will use $\mathcal{K}$ to denote the basic assumption of full knowledge about actions, as defined above. $\mathcal{Q}$ will represent the weaker case that is often used in planning, i.e. that for each action there is a known precondition, and in case the precondition is satisfied there is full knowledge about the possible effects of the action, otherwise not.

A combination of an epistemological and an ontological descriptor are formed as for example $\mathcal{K}$-**IA**. Simple inertia in the frame problem literature is concerned with the $\mathcal{K}$-**IA** family of reasoning problems. By comparison, standard planning algorithms such as TWEAK [6] are concerned with the $\mathcal{Q}$-**I** family: the effects of actions are only known if preconditions are satisfied, but in that case the effects are not conditional on the starting state. An expression such as $\mathcal{K}$-**IA** will also represent the set of all chronicles having those epistemological and ontological characteristics.

5.6.2 Sub-characteristics

In addition there is a need to introduce *sub-characteristics* which provide additional detail in the taxonomy and which will be written with a small letter. For example **Is** denotes the subfamily of **I** systems where all actions take a single time-step. **Id** denotes the subfamily where in every action and for every feature affected by the action, the feature makes a single change from its old to its new value. **Ad** denotes the subfamily of **A** where all actions are deterministic, and so on.

All the subcharacteristics of **IA** can be precisely defined in terms of the trajectory semantics. For example **Is** is characterized so that every member of $\texttt{Trajs}(E, r)$ is a trajectory of length 1, for every E and r. **Ad** is characterized so that all members of $\texttt{Trajs}(E, r)$ for given E and r have the same last element, and so on. The full catalogue of characteristics and subcharacteristics, with their exact definitions, are in the book.

The assessment of a proposed logic is therefore always done within the framework of a given (main) characteristic. Given that $\mathcal{K}$-**IA** has obtained a precise definition, including a precise definition of the intended set of models $\Sigma_{\mathbf{IA}}(\Upsilon)$ for a chronicle $\Upsilon \in \mathcal{K}$-**IA**, one wishes to obtain results of the form "if $\Upsilon \in \mathcal{Z}$ then the set of selected models for Υ according to the logic equals $\Sigma_{\mathbf{IA}}(\Upsilon)$". For example if it is known that the logic applies correctly for chronicles where all actions are deterministic and take a single time-step, then $\mathcal{Z}$ in the assessment is chosen as $\mathcal{K}$-**IsAd**.

Corresponding to the ontological sub-characteristics, there is also a need for epistemological sub-characteristics. $\mathcal{K}\mathsf{s}$ represents that there is full knowledge about the initial state (time 0), and $\mathcal{K}\mathsf{p}$ that there are no observations about any timepoint after the initial one. The combination of the last two properties is written as $\mathcal{K}\mathsf{sp}$.

If a chronicle (scenario description) is given, then the epistemological sub-characteristics can be evaluated by a straight-forward syntactic or semantic analysis of the chronicle. For a given chronicle Υ one can determine which sub-characteristics apply there and which do not. The epistemological sub-characteristics s and p depend only on the SCD and OBS components of the chronicle, but others such as c for "the chronicle is consistent" depend on the combination of all four components.

The main epistemological characteristics, on the other hand, can not be identified from the chronicle itself; they express the assumptions that are intended to *accompany* the chronicle and to be used in its interpretation. For example, "all the actions have been explicitly specified here, so there are no actions other than those included in SCD" is a main epistemological characteristic and not a sub-characteristic.

5.7 Assessments of some simple nonmonotonic logics for action and change

The following are the assessments of some previously proposed logics for action and change. Each of them is first reformulated using the base logic i.e. the syntax and the surface semantics that were defined above. The distinguishing feature of each approach is only how it obtains the set of *selected models* for a given chronicle Υ. The analysis will therefore consider *entailment methods* S which are mappings from chronicles in $\mathcal{K}$-**IA** to sets of interpretations $\langle M, R\rangle$. Such an entailment method will be said to be correct for a set $\mathcal{Z} \subseteq \mathcal{K}$-**IA** of chronicles iff $S(\Upsilon) = \Sigma_{\mathbf{IA}}(\Upsilon)$ for every $\Upsilon \in \mathcal{Z}$ where the actions are necessarily sequential.

5.7.1 Some previously proposed entailment methods

Let a chronicle $\Upsilon = \langle \mathcal{O}, \Pi, \text{SCD}, \text{OBS} \rangle \in \mathcal{K}\text{-}\mathbf{IA}$ be given, and let $[\![\Upsilon]\!]$ be the set of those classical models $\langle M, R \rangle$ for $\Pi(\text{SCD}) \cup \text{OBS}$ where $\mathcal{O}$ is the object domain. The set $[\![\Upsilon]\!]$ allows models which do not satisfy inertia and models where there are additional actions besides those specified in SCD, and we have proved

Fact 1 $\Sigma_{\mathbf{IA}}(\Upsilon) \subseteq [\![\Upsilon]\!]$.

Any entailment method which obtains $\Sigma_{\mathbf{IA}}(\Upsilon)$ correctly must therefore further reduce the set of models, and this is of course the reason why non-monotonic logics are used for reasoning about worlds with inertia. Several proposed entailment methods define the set of selected models as

$$Min(\ll, [\![\Upsilon]\!])$$

i.e. as the set of those members of $[\![\Upsilon]\!]$ which are minimal with respect to a preference relation $\ll$. All such approaches are inherently limited due to the following fact.

Fact 2 *There does not exist any preference order* $\ll$ *such that*

$$Min(\ll, [\![\Upsilon]\!]) = \Sigma_{\mathbf{IA}}(\Upsilon)$$

for all $\Upsilon \in \mathcal{K}\text{-}\mathbf{IA}$.

Proof. Consider a world where there is a single propositional feature p and a single action A whose effect is like tossing a coin, i.e. the value of p after the action is nondeterministically `T` or `F`, regardless of its value when the action starts. Let Υ be the chronicle where both SCD and OBS are empty, and let Υ' be the chronicle where OBS is empty and SCD consists of the single formula $[4, 5]\, A$. Then $[\![\Upsilon]\!] = [\![\Upsilon']\!]$ but $\Sigma_{\mathbf{IA}}(\Upsilon) \neq \Sigma_{\mathbf{IA}}(\Upsilon')$, and the proposition follows. ■

However it is anyway meaningful to assess the range of applicability of those entailment methods which have been proposed, both because of their historical interest and since they may be useful in applications that do not require the full range of $\mathcal{K}\text{-}\mathbf{IA}$.

The simplest possible minimization is to use the original proposal by McCarthy [38], which was also used by Hanks and McDermott, but to omit the implication of abnormality from the action laws. The resulting method can be rephrased as follows in our base logic. Let $I = \langle M, R \rangle$ and $I' = \langle M', R' \rangle$ be two members of $[\![\Upsilon]\!]$. The changeset of I is defined as

$$changeset(I) = \{\langle f, t \rangle \mid R(f, t-1) \neq R(f, t)\}$$

i.e. the set of pairs of a feature and a timepoint t where the feature changes value from time $t-1$ to time t. Then $I \ll_{pgm} I'$ iff $M = M'$ and $changeset(I) \subset changeset(I')$. The preference relation $\ll_{pgm}$ defines *prototypical global minimization*, PGM, so that

$$S_{pgm}(\Upsilon) = Min(\ll_{pgm}, [\![\Upsilon]\!]).$$

We will return below to the case where abnormality is explicitly implied. Among the various proposals for solving the Yale shooting problem, the proposal by Kautz [22] can be easily rewritten using our base logic as follows, and will be called *original chronological minimization*, OCM. The breakset of I at time t is defined as

$$breakset(I, t) = \{f \mid R(f, t-1) \neq R(f, t)\}$$

i.e. the set of features which change value from time $t-1$ to time t. Then $I \ll_{ocm} I'$ iff $M = M'$ and there is some timepoint $\mathtt{t}$ such that both

- for all $t < \mathtt{t}$, $breakset(I, t) = breakset(I', t)$
- $breakset(I, \mathtt{t}) \subset breakset(I', \mathtt{t})$.

The preference relation $\ll_{ocm}$ defines the entailment method S_{ocm} as

$$S_{ocm}(\Upsilon) = Min(\ll_{ocm}, [\![\Upsilon]\!]).$$

Actually this definition is slightly different from the one given by Kautz, since changes from false to true and from true to false are not treated symmetrically in his paper. However we assume that this was just an oversight.

Some counterexamples to OCM can be remedied by the following modification of the definition, obtaining *prototypical chronological minimization*, PCM. Let $I \ll_{pcm} I'$ iff $M = M'$ and there is some timepoint $\mathtt{t}$ such that both

- for all $t < \mathtt{t}$, $R(t) = R'(t)$
- $breakset(I, \mathtt{t}) \subset breakset(I', \mathtt{t})$.

The preference relation $\ll_{pcm}$ defines the entailment method S_{pcm} as

$$S_{pcm}(\Upsilon) = Min(\ll_{pcm}, [\![\Upsilon]\!]).$$

We proceed now to the assessments for these methods. The proofs of the main results require a number of auxiliary concepts and lemmas, and can not be included here due to the space constraints. For the full proofs and for some fine points regarding the methods, please refer to the book.

5.7.2 Original chronological minimization

Fact 3 *Original chronological minimization (OCM) is correct for* $\mathcal{K}$sp-**IsAd**.

Proof outline. Consider a $\mathcal{K}$sp-**IsAd** chronicle $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle$, and a valuation M with which the actions in SCD are sequential. For given Υ and M, exactly one state r satisfies the observations for time zero, according to the epistemological assumptions, and since the actions are deterministic there is at most one intended model for the given M. Compare it to the set of classical models (members of $[\![\Upsilon]\!]$) containing the given M. Call this set S. For each non-intended member of S one can construct a corresponding intended model which is OCM-preferred over the given one. Also, if two members of S have equal breaksets at all timepoints they are equal. Finally, if there is no intended model for a certain M, then $[\![\Upsilon]\!]$ also does not contain any member with that M. This concludes the proof. ■

In the $\mathcal{K}$sp-**IsAd** family the initial state must be completely specified, there must not be any observations for times later than the initial one, and all actions must take a single timestep and be deterministic. It is fairly easy to find counterexamples where any of these restrictions is violated.

5.7.3 Prototypical chronological minimization

For the assessment of prototypical chronological minimization (PCM) we define additional sub-characteristics as follows. Let Υ be a chronicle and let $\langle \texttt{Infl}, \texttt{Trajs} \rangle$ be the trajectory-semantics description of the action laws in Υ. Let

$$\texttt{Pdur}(E, r) = \{length(v) \mid v \in \texttt{Trajs}(E, r)\}$$

i.e. the set of possible durations (expressed as the number of elements in the trajectory) when the action E is performed from the starting state r. An action E is *equidurational* iff $\texttt{Pdur}(E, r)$ is independent of r. Υ is equidurational, and has the subcharacteristic **IAe**, iff all actions in it are equidurational.

For given E and r and a given integer k, let W_k be

$$\{r; y \mid y \in \texttt{Trajs}(E, r) \wedge length(y) = k\}.$$

i.e. it is obtained by taking the subset of $\texttt{Trajs}(E, r)$ consisting of all those members which have length k, and "consing" r itself at the front. Υ has the subcharacteristic **IAx** iff $Min(\ll_{pcm}, W_k) = W_k$ for all E, r, and k.

Fact 4 *Prototypical chronological minimization (PCM) is correct for* $\mathcal{K}$p-**IAex**.

For the proof, please refer to the book. In $\mathcal{K}$p-**IAex** the initial state does not have to be completely specified, but still there must not be any observations for times later than the initial one. Notice that **IsAd** $\subset$ **IAex** since single-timestep actions are certainly equidurational, and since W_k for a deterministic single-step action is either a singleton or empty. Therefore PCM applies correctly for all members of $\mathcal{K}$p-**IsAd**, which is a generalization over OCM.

A generalization of deterministic actions can be defined as follows. An action is said to exhibit *necessary change* if it is allowed to be non-deterministic, but only in such a way that at each step in time, there is only a choice between alternative new values, and never a choice between having or not having a change. For example a roulette wheel which will never show the same number twice in a row, exhibits necessary change. A world in **IA** is said to have the **IAn** subcharacteristic if all its actions exhibit necessary change. It follows that **IsAd** $\subset$ **IsAn** $\subset$ **IAex**.

Using another methodology, Lin and Shoham [33] have previously analyzed the $\mathcal{K}$p-**IsAd** family of reasoning problems, and proved that PCM is correct for a family of reasoning tasks that appears equivalent to $\mathcal{K}$p-**IsAd** plus deterministic ramification. The present result for PCM confirms theirs and generalizes it considerably.

At this point the proposition refers only to a lower bound on the range of applicability for PCM: it is correct for all members of $\mathcal{K}$p-**IAex**. We return to the question of upper bounds in a later subsection.

5.7.4 Prototypical global minimization

The assessment of prototypical global minimization (PGM) requires some additional sub-characteristics. A world in **IsA** described by $\langle \mathtt{Infl}, \mathtt{Trajs} \rangle$ is said to be *minimal-change-perfect* and have the subcharacteristic **IsAz** iff, whenever $\langle \bar{r} \rangle \in \mathtt{Trajs}(E, r)$ and $\langle \bar{r}' \rangle \in \mathtt{Trajs}(E, r')$, it holds that

$$\{f \mid \bar{r}[f] \neq r[f]\} = \{f \mid \bar{r}'[f] \neq r'[f]\}.$$

In other words, the set of features that actually change as the result of an action E is always the same.

A chronicle is said to be *pointwise consistency-retaining* iff it is the case that, for every valuation M, either there is no classical model containing M, or there is some intended model containing M. This will be represented by the epistemological subcharacteristic $\mathcal{K}$r.

Fact 5 *Prototypical global minimization is correct for chronicles in* $\mathcal{K}$r-**IsAz**.

The **IsAz** requirement generalizes in a straightforward fashion to the case of extended duration of actions, but it is still quite restrictive.

5.8 Assessments of logics based on occlusion and filtering

The logics that were assessed in the previous section were already known to fail fairly easily, and the new results give a precise form to that insight. In order to obtain a logic that gives correct results for a larger class of systems while still retaining the preferential character, one must find alternatives to the straightforward preferential entailment methods of the form

$$Min(\ll, [\![\Upsilon]\!])$$

which can be rewritten equivalently as

$$Min(\ll, [\![\Pi(\text{SCD}) \cup \text{OBS}]\!])$$

where the model sets $[\![.]\!]$ are restricted to those models having $\mathcal{O}$ as the object domain. *Filtering* and *occlusion* were described in subsection 3.3.

5.8.1 Filtering

Filtering [58] treats the premise sets SCD and OBS separately, and defines the selected set of models as

$$Min(\ll, [\![\Pi(\text{SCD})]\!]) \cap [\![\text{OBS}]\!]$$

so that the minimization is imposed "before" the observations. If the preference relation $\ll$ is chosen as $\ll_{pcm}$ (which was defined for PCM) then one obtains *prototypical chronological minimization with filtering* (PCMF), defined by

$$S_{pcmf}(\langle \mathcal{O}, \Pi, \text{SCD}, \text{OBS} \rangle) = Min(\ll_{pcm}, [\![\Pi(\text{SCD})]\!]) \cap [\![\text{OBS}]\!]$$

We have proved

Fact 6 *PCMF is correct for the* $\mathcal{K}$-**IAex** *family of chronicles.*

PCMF therefore allows observations at arbitrary timepoints, but the other two restrictions of PCM are retained.

5.8.2 Occlusion

Occlusion [58] is a formal device for identifying when a feature is liable or "authorized" to change, in particular as the result of the action, and regardless of whether it actually changes or not. In this way one can eliminate the restriction to "necessary change" that was obtained for PCM and PCMF. The authorization to change is expressed using a property of occlusion. The occlusion information is the same as is expressed by the `Infl` component in the trajectory semantics. Interpretations are extended from $\langle M, R \rangle$ to $\langle M, R, X \rangle$ where X is a relation over features times timepoints. An action law saying that a feature f possibly changes its value over an interval of time $[s, t]$, is now expressed as saying that f is occluded throughout $(s, t]$. In addition there is usually some information about its value within or at the end of the interval. The minimization methods are changed so that they minimize *unoccluded* change rather than any change. Also of course occlusion itself has to be minimized. Observations are not allowed to use the occlusion operator.

The syntax for referring to the occlusion predicate X in action laws will be indicated through an example; the general formulation can be found in the book. Suppose the action A always takes three steps of time, and it affects the propositional features p and q. If both p and q are initially true, A causes p to become false somewhere during the interval, and q may possibly become false in the last time-step but is certainly true before then. The relevant part of the action law for A would then be:

$$[s,t]A \Rrightarrow t = s+3 \wedge ([s]p \wedge q \Rightarrow$$

$$[s,t]p{:=}\mathtt{F} \wedge (s,t]\mathbb{X}q \wedge [s,t-1]q)$$

Here $[s,t]p{:=}\mathtt{F}$ is considered as an abbreviation for $(s,t]\mathbb{X}p \wedge [t]p \mathrel{\hat{=}} \mathtt{F}$. Informally, it is read as saying that the feature p changes its value to become $\mathtt{T}$ some time during the interval $[s,t]$. Formally, $(s,t]\mathbb{X}p$ states that p is occluded during the interval $(s,t]$, i.e. at time t and in all inner points of the interval, but not at time s.

The use of occlusion together with default frame axioms and global (non-chronological) minimization will obtain the original method proposed by McCarthy, where abnormality is explicitly implied in the action laws. The resulting method, *global minimization of occlusion with nochange premises*, GMON, is correct for chronicles in $\mathcal{K}$r-**IsA** subject to an additional requirement which is satisfied if the set of occluded features for an action is independent of the starting state. For the full assessment, please refer to the book.

A preference relation that chronologically minimizes occlusion and change together, and which also uses filtering, is defined as follows. It will be referred to as *chronological minimization of occlusion and change*, CMOC. Let $\langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle$ be a chronicle in $\mathcal{K}$-**IA**, and let $I = \langle M, R, X \rangle$ and $I' = \langle M', R', X' \rangle$ be two models for $\Pi(\textsc{scd})$, both having $\mathcal{O}$ as their object domains. Define a modified breakset function as

$$brs(I,t) = \{f \mid R(f,t-1) \neq R(f,t) \wedge \neg X(f,t)\}$$

i.e. as the set of features having unoccluded changes at time t. Then the preference relation $\ll_{cmoc}$ is defined so that $I \ll_{cmoc} I'$ iff $M = M'$ and there is some timepoint $\mathtt{t}$ such that both of the following hold:

- For all $t < \mathtt{t}$, $R(t) = R'(t)$ and $X(t) = X'(t)$.
- Either of the following holds:
  - $X(\mathtt{t}) \subset X'(\mathtt{t})$
  - $X(\mathtt{t}) = X'(\mathtt{t}) \wedge brs(I,\mathtt{t}) \subset brs(I',\mathtt{t})$.

The set of selected models is defined using filtering, as

$$S_{cmoc} = Min(\ll_{cmoc}, [\![\Pi(\textsc{scd})]\!]) \cap [\![\textsc{obs}]\!].$$

We have proved

Fact 7 *CMOC is correct for the* $\mathcal{K}$-**IAe** *family of chronicles.*

The CMOC method was tested against a fairly extensive set of test examples, which contains or subsumes most problems that have been used by other authors in the field, and it passed all the tests. However when attempting to prove the correctness of this method for $\mathcal{K}$-**IA** it was discovered that it is in fact not completely general, and the correct assessment is that CMOC is correct for $\mathcal{K}$-**IAe** chronicles.

The use of occlusion and filtering together therefore allows to eliminate two of the three constraints on the applicability of PCM within the $\mathcal{K}$-**IA** family of chronicles. The remaining constraint, equidurationality, is unfortunately a quite damaging one: certainly if you allow actions to have conditional effects depending on the starting state and you allow them to have a duration different from 1, you would expect that the duration of the action may depend on the starting state. Once you know of this restriction, it is easy to construct counterexamples outside $\mathcal{K}$-**IAe** where CMOC gives the wrong results.

5.8.3 Chronological assignment and chronological minimization of occlusion and change

Fortunately it is possible to correct the remaining problem of CMOC and obtain a method which gives the correct results throughout $\mathcal{K}$-**IA**. The revised entailment method, CAMOC, is characterized by an additional syntactic restriction and a modification of the preference relation.

The syntactic restriction is as follows. All action statements must be written in the form $[\boldsymbol{s}_i, \boldsymbol{t}_i] A_i$ where $\boldsymbol{s}_i$ and $\boldsymbol{t}_i$ are constant symbols, and the same constant symbol is only used in one single action statement. The schedule SCD may only contain such action statements, temporal ordering statements of the form $\boldsymbol{s}_i < \boldsymbol{t}_i$ or $\boldsymbol{t}_i \leq \boldsymbol{s}_j$ which serve to order the actions, and disjunctions between the latter kind of statements. All other information about the order and distance between timepoints must be placed in the set OBS of observations. These syntactic constraints are harmless since they can easily be accomodated by moving premises from the SCD to the OBS partition of the chronicle, which of course does not affect $[\![\Upsilon]\!]$. However it may be necessary to introduce additional temporal constant symbols in order to replace composite expressions in the time interval for an action.

The preference relation is modified as follows. Given a chronicle and two models I and I' like for CMOC, we first define $M_{0:\mathtt{t}}$ as a partial function which is obtained from M by restricting it to object constants and to those temporal constants whose value is $\leq \mathtt{t}$. (Remember that M is a mapping from constant symbols to corresponding values). If there are no object constants, the definition of $M_{0:\mathtt{t}}$ is therefore

$$M_{0:\mathtt{t}} = \{(\tau \mapsto t) \mid (\tau \mapsto t) \in M \wedge t \leq \mathtt{t}\}$$

where as usual a valuation is a set of pairs $(\tau \mapsto t)$ consisting of a constant symbol τ and a timepoint (integer) t. This operation is therefore a restriction on a function to a restricted value range, not to a restricted argument range.

The preference relation $\ll_{camoc}$ is then defined as follows: $I \ll_{camoc} I'$ iff there is some timepoint $\mathtt{t}$ such that all of the following hold:

- $M_{0:\mathtt{t}} = M'_{0:\mathtt{t}}$.
- for all $t < \mathtt{t}$, $R(t) = R'(t)$ and $X(t) = X'(t)$.
- Either of the following holds:
  - $X(\mathtt{t}) \subset X'(\mathtt{t})$
  - $X(\mathtt{t}) = X'(\mathtt{t}) \wedge brs(I, \mathtt{t}) \subset brs(I', \mathtt{t})$.

The set of selected models is defined using filtering, as

$$S_{camoc} = Min(\ll_{camoc}, [\![\Pi(\text{SCD})]\!]) \cap [\![\text{OBS}]\!].$$

We have proved

Fact 8 *CAMOC is correct for all chronicles in* $\mathcal{K}$-**IA** *that satisfy the syntactic constraints mentioned above.*

The intuition behind the preference relation $\ll_{camoc}$ is that the values of timepoints should not be assigned all at once, for example in the initial valuation M_0 of the game between ego and world. Instead, if a schedule specifies an action as $[\boldsymbol{s}_i, \boldsymbol{t}_i] E_i$, then the constant symbol $\boldsymbol{s}_i$ should be assigned the current value of *now* when the ego decides to invoke the action, and the constant symbol $\boldsymbol{t}_i$ should be assigned the current value of *now* when the world has concluded the action. This makes sense since the actual duration of the action is a consequence of the choice of trajectory within $\mathtt{Trajs}(E, r)$; a choice which is made by the world in the ego-world game.

5.9 Assessments using other methods

The $\mathcal{A}$ language of Gelfond and Lifschitz is able to express scenarios with binary-valued properties, deterministic actions and without a notion of duration, i.e. the $\mathcal{K}$-**IbsAd** family in our notation. Kartha [20] has verified that the state-space approach of Baker, the ADL-based method of Pednault, and the explanation closure method of Reiter are correct within this framework. If the semantics of $\mathcal{A}$ is identical to the $\mathcal{K}$-**IbsAd** specialization of trajectory semantics, which seems obvious but has not been proven formally, then one has a verified range of applicability also for these approaches.

The $\mathcal{A}$ language used by Kartha does not allow a larger class of reasoning problems to be analyzed, but one may speculate whether the range of applicability of these three methods is in fact broader than has been proved. We consider the case of the persistence assumption (no surprises, ramification, etc), where the possible generalizations would be in the direction of multi-valued features, non-deterministic actions, and actions with extended duration. The actual presentations of these three methods do not include these more general cases. The generalization to multi-valued features appears to be trivial for all three. It is probably straightforward to generalize Pednault's scheme to allow non-deterministic actions, and Reiter's method can probably accomodate all three extensions.

5.10 Upper bounds

The assessments in the previous sections obtain lower bounds on the range of applicability. It is natural to also look for corresponding upper bounds for the same entailment methods. Ideally one would like to know that a certain chronicle family is both a lower bound and an upper bound, so that no additional improvement of the result is possible.

The concept of upper bound however deserves some consideration. The simplest definition would be that $\mathcal{Z} \subseteq \mathcal{K}$-$\mathbf{IA}$ is an upper bound on the range of applicability for an entailment method S iff $\Sigma_{\mathbf{IA}}(\Upsilon) \neq S(\Upsilon)$ for any $\Upsilon \in \mathcal{K}$-$\mathbf{IA} - \mathcal{Z}$. This is however an inconvenient and fairly uninformative definition, since there are always a number of cases where an entailment method so to say accidentally obtains the intended models. A more useful definition can be obtained along the following lines.

Consider first the method CMOC, which has been proven correct for $\mathcal{K}$-$\mathbf{IAe}$.

Fact 9 *If* $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle \in \mathcal{K}$-$\mathbf{IA} - \mathcal{K}$-$\mathbf{IAe}$*, then there is some modified chronicle* $\Upsilon' = \langle \mathcal{O}, \Pi, \textsc{scd}', \textsc{obs}' \rangle$ *such that* $\Sigma_{\mathbf{IA}}(\Upsilon') \neq S_{cmoc}(\Upsilon')$.

Proof. In this case there is some action designator E, some states r and r', and some trajectory $v \in \texttt{Trajs}(E, r)$ of length k such that no trajectory $v' \in \texttt{Trajs}(E, r')$ has length k. If there are several r with this property, for the same E, k, and r', then select an r which minimizes the set of features where r and r' differ. Construct $\textsc{scd}'$ as $\texttt{[}1, k+1\texttt{]}\, E$ and $\textsc{obs}'$ as $\texttt{[0]}\, \phi$, where ϕ is a formula that precisely characterizes the state r'. Then Υ' does not have any intended models according to $\mathcal{K}$-$\mathbf{IA}$. However the interpretation $\langle M, R \rangle$ where $R(0) = r'$, $R(1) = r$, and R then continues according to v is a preferred CMOC model. ■

Of course if $\Upsilon \in \mathcal{K}$-$\mathbf{IA} - \mathcal{K}$-$\mathbf{IAe}$ then $\Upsilon' \in \mathcal{K}$-$\mathbf{IA} - \mathcal{K}$-$\mathbf{IAe}$ as well, since the ontological characterization is only based on the properties of the world, which is specified by the action laws Π. In other words, if we are building a robot that is to operate in an equidurational world, then the entailment method CMOC will obtain the correct conclusions in all the chronicles that can arise, and if the world is not equidurational then there are some scenarios in the world where CMOC will obtain the wrong conclusions. In that sense of upper bound on range of applicability, the lower bound and the upper bound that we have obtained are equal.

For PCMF the proven lower bound is $\mathcal{K}$-**IAex**. Here the situation is similar:

Fact 10 *If* $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle \in \mathcal{K}\text{-}\mathbf{IA} - \mathcal{K}\text{-}\mathbf{IAex}$*, then there is some modified chronicle* $\Upsilon' = \langle \mathcal{O}, \Pi, \textsc{scd}', \textsc{obs}' \rangle$ *such that* $\Sigma_{\mathbf{IA}}(\Upsilon') \neq S_{pcmf}(\Upsilon')$.

Proof. Since $\Upsilon \in \mathcal{K}\text{-}\mathbf{IA} - \mathcal{K}\text{-}\mathbf{IAex}$, it is either not equidurational or does not satisfy the **Ax** subcharacteristic. In the former case the proof of fact 9 applies. In the latter case there is some action designator E, some state r, and some positive integer k such that the set W_k of trajectories in $\mathtt{Trajs}(E, r)$ of length k satisfies $Min(\ll_{pcm}, W_k) \neq W_k$. Choose $\textsc{scd}'$ with $[0, k]E$ as its only member, and $\textsc{obs}'$ as $[0]\phi$ where ϕ completely characterizes r. The result follows immediately. ■

For PCM the proven lower bound is $\mathcal{K}\mathsf{p}$-**IAex**, which contains both ontological and epistemological constraints. The ontological constraint can be analyzed in similar terms as the previous cases:

Fact 11 *If* $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle \in \mathcal{K}\mathsf{p}\text{-}\mathbf{IA} - \mathcal{K}\mathsf{p}\text{-}\mathbf{IAex}$*, then there is some modified chronicle* $\Upsilon' = \langle \mathcal{O}, \Pi, \textsc{scd}', \textsc{obs}' \rangle$ *which also has the epistemological sub-characteristic* p*, and such that* $\Sigma_{\mathbf{IA}}(\Upsilon') \neq S_{pcm}(\Upsilon')$.

Proof. The proof for fact 10 applies for this case also. ■

.

Therefore, as long as we stay within the epistemological constraint $\mathcal{K}\mathsf{p}$-**IA**, the assessment $\mathcal{K}\mathsf{p}$-**IAex** is both lower and upper bound. However the epistemological requirement p can be improved upon. It is easy to find single examples where a relaxation is possible, for example a chronicle containing observations of the form $[t]\phi$ where $t > 0$ but ϕ is a tautology. One would therefore be interested in an upper bound in the epistemological dimension as well, if such a concept can be defined. The following would seem to be an appropriate framework. Let S be an entailment method, for example PCM, let $\mathcal{K}\xi\text{-}\Psi$ be a chronicle family formed using the ontological family Ψ and the epistemological sub-characteristics ξ, and where S is correctly applicable for all chronicles in $\mathcal{K}\xi\text{-}\Psi$. Then:

- $\mathcal{K}\xi\text{-}\Psi$ is an *ontologically maximal* assessment iff the following condition holds. If $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle \in \mathcal{K}\xi\text{-}\mathbf{IA} - \mathcal{K}\xi\text{-}\Psi$, then there is some modified chronicle $\Upsilon' = \langle \mathcal{O}, \Pi, \textsc{scd}', \textsc{obs}' \rangle$ that also has the epistemological sub-characteristic ξ, and such that $\Sigma_{\mathbf{IA}}(\Upsilon') \neq S(\Upsilon')$.

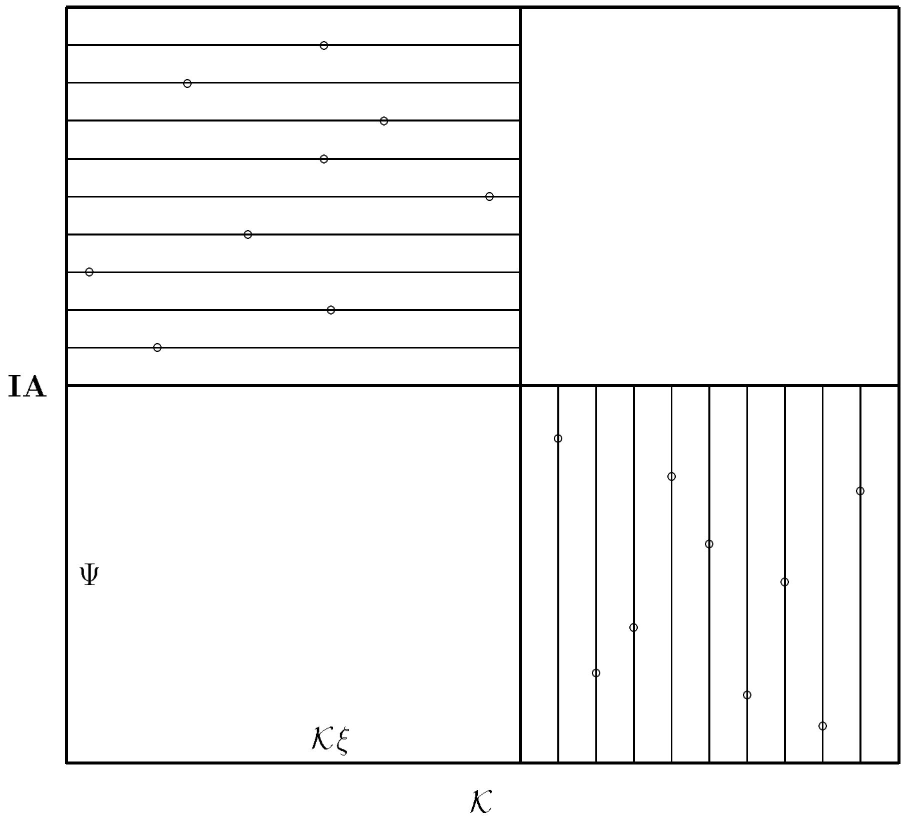

Figure 1: `Maximal assessments`

- $\mathcal{K}\xi$-Ψ is an *epistemologically maximal* assessment iff the following condition holds. If $\Upsilon = \langle \mathcal{O}, \Pi, \text{SCD}, \text{OBS} \rangle \in \mathcal{K}\text{-}\Psi - \mathcal{K}\xi\text{-}\Psi$, then there is some modified chronicle $\Upsilon' = \langle \mathcal{O}, \Pi', \text{SCD}, \text{OBS} \rangle \in \mathcal{K}\text{-}\Psi - \mathcal{K}\xi\text{-}\Psi$ (so that the world described by Π' is also in Ψ), such that $\Sigma_{\mathbf{IA}}(\Upsilon') \neq S(\Upsilon')$.

Figure 1 illustrates this definition, with the (slightly simplified) assumption that the epistemological constraints only refer to the SCD and OBS components of the chronicle. The horizontally striped rectangle to the upper-left indicates how the ontologically maximal assessment works. For every chronicle Υ (arbitrary position on any of the lines) there is some chronicle Υ' with the same action laws and therefore the same world (other position on the same horizontal line) where the entailment method obtains the wrong result (dot on the line). The vertically striped rectangle on the lower-right indicates similarly how the epistemologically maximal assessment works. For every chronicle Υ (arbitrary position on any of the lines) there is some chronicle Υ' with the same schedule and observations but not necessarily the same world (other position on the same vertical line) which satisfies the same epistemological subcharacteristics, and where the entailment method obtains the wrong result (dot on the line).

Therefore, although some chronicles Υ in the striped areas of the figure may satisfy $S(\Upsilon) = \Sigma_{\mathbf{IA}}(\Upsilon)$, it is not possible to extend the assessed region,

Entailment method	Current assessment	Ontologically maximal?	Epistemologically maximal?
PGM	$\mathcal{K}$r-**IsAz**	No	Yes
OCM	$\mathcal{K}$sp-**IsAd**	No	No
PCM	$\mathcal{K}$p-**IAex**	Yes	No
PCMF	$\mathcal{K}$-**IAex**	Yes	Yes
CMOC	$\mathcal{K}$-**IAe**	Yes	Yes
CAMOC	$\mathcal{K}$-**IA**	Yes	Yes
Baker	$\mathcal{K}$-**IbsAd**	No?	Yes
Pednault	$\mathcal{K}$-**IbsAd**	No?	Yes
Reiter	$\mathcal{K}$-**IbsAd**	No?	Yes

Table 1: `Summary of assessments`

represented by the lower-left rectangle, either upwards while retaining its right side, or rightwards while retaining its upper side.

For CAMOC, CMOC and PCMF the epistemological upper bound is not an issue since the assessment does not include any epistemological constraint besides $\mathcal{K}$ itself. For PCM, $\mathcal{K}$p-**IAex** is ontologically maximal but not epistemologically maximal. The full epistemological range for PCM relative to **IAex** allows for example chronicles of the form

$$[0]\,\phi$$

$$[0, \boldsymbol{t}_1]\,A$$

$$[\boldsymbol{t}_1]\,\psi$$

where $\phi \models \psi$. It remains to determine the exact upper bound in the epistemological dimension.

5.11 Summary of applicability results

The assessment results for these entailment methods are summarized in the table 1. For each entailment method it indicates the assessment that has been stated here, whether that assessment is ontologically maximal, and whether it is epistemologically maximal. The book contains stronger assessments for PGM and OCM, as well as assessments for a number of additional entailment methods.

The last three lines refer to the assessments by Kartha which were reviewed in subsection 5.9. The probable generalizations that were mentioned there have not been included in the table attending a precise proof.

6 A unified logic of common sense?

The classical research paradigm for the frame problem was driven by the assumption that one could find a logic of common sense reasoning which were able to produce common-sense conclusions from *bona fide* common-sense scenario descriptions. Although it was recognized right from the start that such a logic would have some non-classical properties, such as nonmonotonicity, there was still an expectation that only one or a few basic principles would have to be added, for example minimization of abnormality, and that those new principles could be formulated in reasonably concise form. Reasoning about actions and change was perceived as one important part of such a logic of common sense reasoning.

This vision is now in serious doubt. The methodology of representative examples, which has been used for a large part of the research, has been useful for finding problems but not for analyzing them with precision. Results that have been obtained so far using a systematic methodology, and which are reviewed in Part II of this handbook chapter, suggest that fairly complex logical operations are needed to obtain any level of generality. Even for the basic case where only persistence is assumed, one needs nontrivial syntactic operations on the axioms, as in explanation closure, or nontrivial preference relations on models as in the CAMOC method. In view of the complexity of the most general methods, one can also suspect that simpler methods with a more limited range of applicability may continue to be useful if they have better computational properties.

The broader range of frame problems includes surprises, ramifications, and qualifications. Eventually we will also need to address topics that have been little studied so far, such as cause-and-effect chains, continuous change, and delayed effects of actions. It now seems likely that several of these topics may require their own preference mechanisms. Perhaps the logic of common-sense reasoning, rather than being unified and concise, will have the character of a Swiss army knife and contain one tool for each purpose. In any case it will be increasingly important to use a methodology where the range of applicability of each method is established in precise terms. The search for the unique and completely general-purpose logic of common-sense reasoning is no longer the main objective of the research.

References

[1] Andrew B. Baker. A simple solution to the Yale shooting problem. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 11–20, 1989.

[2] Andrew B. Baker and Matthew L. Ginsberg. Temporal projection and explanation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1989.

[3] Chitta Baral and Michael Gelfond. Representing concurrent actions in extended logic programming. In *International Joint Conference on Artificial Intelligence*, pages 866–871, 1993.

[4] John Bell. Why the frame problem is not a problem. *Artificial Intelligence Communications*, 3(1):3–10, 1990.

[5] G. Brewka and J. Hertzberg. How to do things with worlds: On formalizing actions and plans. *Journal of Logic and Computation*, 1993.

[6] David Chapman. Planning for conjunctive goals. *Artificial Intelligence*, 32:333–377, 1987.

[7] Ernest Davis. *Representation of Commonsense Knowledge*. Morgan-Kaufmann Publishing Co., 1990.

[8] Alvaro del Val and Yoav Shoham. Deriving properties of belief update from theories of actions. In *National (U.S.) Conference on Artificial Intelligence*, pages 584–589, 1992.

[9] Richard E. Fikes and Nils J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2:189–208, 1971.

[10] J. J. Finger. *Exploiting Constraints in Design Synthesis*. PhD thesis, Stanford University, Stanford, CA, 1987.

[11] Michael Gelfond. Autoepistemic logic and formalization of commonsense reasoning. In *Non-Monotonic Reasoning: Second International Workshop (Lecture Notes in Artificial Intelligence 346)*, pages 176–186. Springer-Verlag, 1989.

[12] Michael Gelfond and Vladimir Lifschitz. Describing action and change by logic programs. In *Proc. Joint International Conference and Symposium on Logic Programming*, 1992.

[13] Matthew L. Ginsberg and David E. Smith. Reasoning about action I: A possible worlds approach. *Artificial Intelligence*, 35:165–195, 1988.

[14] Matthew L. Ginsberg and David E. Smith. Reasoning about action II: The qualification problem. *Artificial Intelligence*, 35:311–342, 1988.

[15] Andrew R. Haas. The case for domain-specific frame axioms. In Frank Brown, editor, *Proceedings of the 1987 Workshop on the Frame Problem in Artificial Intelligence*. Morgan Kaufmann, Lawrence, Kansas, 1987.

[16] Steve Hanks and Drew McDermott. Nonmonotonic logics and temporal projection. *Artificial Intelligence*, 33:379–412, 1987.

[17] Brian A. Haugh. Simple causal minimizations for temporal persistence and projection. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 218–223, 1987.

[18] Carl Hewitt. Description and theoretical analysis (using schemata) of PLANNER: a language for proving theorems and manipulating models in a robot. Technical Report A.I. memo 251, MIT Project MAC, 1972.

[19] David Israel. What's wrong with nonmonotonic logic? In *National (U.S.) Conference on Artificial Intelligence*, pages 99–101, 1980.

[20] G. Neelakantan Kartha. Soundness and completeness theorems for three formalizations of action. In *International Joint Conference on Artificial Intelligence*, pages 724–729, 1993.

[21] Hirofumi Katsuno and Alberto O. Mendelzon. A unified view of propositional knowledge base update. In *International Joint Conference on Artificial Intelligence*, 1989.

[22] Henry A. Kautz. The logic of persistence. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 401–405, 1986.

[23] Kurt Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35:343–382, 1988.

[24] I. Kramosil. A note on deduction rules with negative premises. In *International Joint Conference on Artificial Intelligence*, pages 53–56, 1975.

[25] Vladimir Lifschitz. Computing circumscription. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 121–127, 1985.

[26] Vladimir Lifschitz. On the semantics of STRIPS. In Michael P. Georgeff and Amy L. Lansky, editors, *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop*. Morgan Kaufmann Publishers, Los Altos, CA, 1986.

[27] Vladimir Lifschitz. Formal theories of action. In F. Brown, editor, *The Frame Problem in Artificial Intelligence: Proceedings of the 1987 Workshop*, pages 35–58. Morgan Kaufmann Publishers, Los Altos, CA, 1987.

[28] Vladimir Lifschitz. Pointwise circumscription. In Matthew L. Ginsberg, editor, *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann Publishers, Los Altos, CA, 1987.

[29] Vladimir Lifschitz. Benchmark problems for formal nonmonotonic reasoning, version 2.00. In Michael Reinfrank, Johan de Kleer, Matthew L. Ginsberg, and Erik Sandewall, editors, *Non-Monotonic Reasoning. Proceedings of the Second International Workshop*, number 346 in Lecture Notes in Artificial Intelligence, pages 202–219. Springer Verlag, 1988.

[30] Vladimir Lifschitz. Toward a metatheory of action. In *International Conf. on Knowledge Representation and Reasoning*, pages 376–386, 1991.

[31] Vladimir Lifschitz and Arkady Rabinov. Miracles in formal theories of action. *Artificial Intelligence*, 38:225–237, 1989.

[32] Fangzhen Lin and Ray Reiter. State constraints revisited. *Journal of Logic and Computation*, 4(5) (to appear), 1994.

[33] Fangzhen Lin and Yoav Shoham. Provably correct theories of action (preliminary report). In *National (U.S.) Conference on Artificial Intelligence*, pages 349–354, 1991.

[34] Fangzhen Lin and Yoav Shoham. Concurrent actions in the situation calculus. In *National (U.S.) Conference on Artificial Intelligence*, pages 590–595, 1992.

[35] John McCarthy. Situations, actions, and causal laws. Technical Report Memo nr 2, Stanford Artificial Intelligence Project, 1963.

[36] John McCarthy. Epistemological problems of artificial intelligence. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 1038–1044, 1977.

[37] John McCarthy. Circumscription — a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[38] John McCarthy. Applications of circumscription to formalizing common sense knowledge. In *Proc. of the Nonmonotonic reasoning workshop*, pages 295–324, October 1984.

[39] John McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28:89–116, 1986.

[40] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Mitchie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.

[41] Drew McDermott. A critique of pure reason. *Computational Intelligence*, 3:151–160, 1987. (Various responses are contained on pages 161–237 of the same isssue of *Computational Intelligence*.).

[42] Robert C. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25:75–94, 1985.

[43] Leora Morgenstern and Lynn Andrea Stein. Why things go wrong: A formal theory of causal reasoning. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 518–523, 1988.

[44] Paul H. Morris. The anomalous extension problem in default reasoning. *Artificial Intelligence*, 35:383–399, 1988.

[45] Allen Newell. The knowledge level. *Artificial Intelligence*, 18(1):87–127, 1982.

[46] Judea Pearl. On logic and probability. *Computational Intelligence*, 4:99–103, 1988.

[47] Edwin P.D. Pednault. Formulating multi-agent, dynamic-world problems in the classical planning framework. In Michael P. Georgeff and Amy L. Lansky, editors, *Reasoning about Actions and Plans*, pages 47–82. Morgan Kaufmann, 1986.

[48] Edwin P.D. Pednault. Adl: Exploring the middle ground between STRIPS and the situation calculus. In *International Conf. on Knowledge Representation and Reasoning*, pages 324–332, 1989.

[49] Pavlos Peppas and Wayne Wobcke. On the use of epistemic entrenchment in reasoning about action. In *European Conference on Artificial Intelligence*, pages 403–407, 1992.

[50] Zenon W. Pylyshyn, editor. *The Robot's Dilemma: The Frame Problem in Artificial Intelligence.* Ablex, Norwood, NJ, 1987.

[51] Anand S. Rao and Norman Y. Foo. Minimal change and maximal coherence: A basis for belief revision and reasoning about actions. In *International Joint Conference on Artificial Intelligence*, pages 966–971, 1989.

[52] Manny Rayner. Did Newton solve the "extended prediction problem"? In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 381–385, 1989.

[53] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

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

[55] Ray Reiter. Proving properties of states in the situation calculus. *Artificial Intelligence*, 64(2):337–351, 1993.

[56] Erik Sandewall. An approach to the frame problem, and its implementation. In *Machine Intelligence, Vol. 7*. Edinburgh University Press, 1972.

[57] Erik Sandewall. Combining logic and differential equations for describing real-world systems. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 412–420, 1989.

[58] Erik Sandewall. Filter preferential entailment for the logic of action in almost continuous worlds. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1989.

[59] Erik Sandewall. *Features and Fluents*. To appear, 1993.

[60] Erik Sandewall. The range of applicability of nonmonotonic logics for the inertia problem. In *International Joint Conference on Artificial Intelligence*, 1993.

[61] Lenhart Schubert. Monotonic solution of the frame problem in situation calculus. In Henry E. Kyburg, Ronald P. Loui, and Greg N. Carlson, editors, *Knowledge Representation and Defeasible Reasoning*, pages 23–67. Kluwer, 1990.

[62] Yoav Shoham. Chronological ignorance: Experiments in nonmonotonic temporal reasoning. *Artificial Intelligence*, 36:279–331, 1988.

[63] Yoav Shoham and Drew McDermott. Problems in formal temporal reasoning. *Artificial Intelligence*, 36:49–61, 1988.

[64] Peter Struss and Oskar Dressler. Physical negation – integrating fault models in the general diagnostic engine. In *International Joint Conference on Artificial Intelligence*, 1989.

[65] David S. Touretzky, John F. Horty, and Richmond H. Thomason. A clash of intuitions: The current state of nonmonotonic multiple inheritance systems. In *International Joint Conference on Artificial Intelligence*, pages 476–482, 1987.

[66] David E. Wilkins. *Practical Planning: Extending the Classical AI Planning Paradigm.* Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[67] Marianne Winslett. Reasoning about actions using a possible models approach. In *National (U.S.) Conference on Artificial Intelligence*, pages 89–93, 1988.

[68] Yan Zhang and Norman Y. Foo. Reasoning about persistence: A theory of actions. In *International Joint Conference on Artificial Intelligence*, pages 718–723, 1993.