The range of applicability of some nonmonotonic logics for strict inertia

Erik Sandewall
Email: erisa@ida.liu.se

Abstract

We present assessments for a number of previously proposed theories of action and change as well as some new ones. Each assessment identifies a class of scenarios for temporal reasoning where the theory is guaranteed to obtain the correct results. For some of the theories we have also obtained upper-bound results on the range of correct applicability. The theories and the their validation criteria use a logic with explicit integer time, and has been generalized to linear and branching metric time. The assessments show how the correctness of each theory depends on factors such as whether actions are deterministic, whether scenarios involve prediction only, or what is the behavior of each type of action within its execution period. The paper describes equally the methodology that has been used for obtaining the assessments, and the assessments themselves.

The final part of the work on this article was done while the author was visiting at LAAS-CNRS, Toulouse, France.

This paper was published in the Journal of Logic and Computation*, Vol. 4, No. 5, 1994, pp. 581-615.*

IDA Technical Report 1994
LiTH-IDA-R-94-14
ISSN-0281-4250

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

1 Theories of action and change: a methodology for assessment

There has been much research in recent years on methods for reasoning about actions and change, and on finding solutions of the so-called "frame problems". New variants of nonmonotonic logics for common-sense reasoning have been proposed, only to be quickly refuted by counterexamples. Unfortunately the results that have been obtained in this fashion are notoriously unreliable. According to the standard research methodology in the area, the evidence in favor of a proposed logic or theory should consist of intuitive plausibility arguments and a small number of scenario examples for which the theory is proven (or claimed) to give the intended conclusions and no others. This prevailing methodology was summarized by Davis [6] 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...*

This methodology has been effective for identifying a range of important problems, such as the qualification and ramification problems. However since it relies on examples, it is not sufficient for obtaining reliable results of a general nature. One also needs to *validate* a proposed approach and prove that it is correct for the intended type of reasoning about a whole class of scenarios.

1.1 Validation of theories of action

The quotation from Davis indicates also how each proposed approach to common-sense reasoning must specify the syntax of its language of logical formulae, the conventions for how to use that language for describing actual scenarios, and the methods for drawing conclusions from a scenario description. I shall use the term *theory* for such a collection of rules and conventions, and reserve the use of the word *logic* for the syntax and the formal semantics which may be used by a particular theory or by several theories jointly.

The validation of such a theory must rely on a formal definition of the function Σ from a *scenario description* Υ, expressed using logical formulae, to the set $\Sigma(\Upsilon)$ of its *intended models*. If $[\![\Upsilon]\!]$ represents the set of classical models, then it is commonly assumed that $\Sigma(\Upsilon) \subseteq [\![\Upsilon]\!]$, reflecting the possibility of conclusions by default. Once the intended models function has been defined, one can analyze a proposed theory in terms of its *selection function* S from scenario descriptions to sets of models. The validation consists of proving that the set $S(\Upsilon)$ of *selected models* is provably equal to $\Sigma(\Upsilon)$ for an interesting and sufficiently large class of scenario descriptions Υ.

The validation criterion can of course be formulated as well in terms of sets of conclusions: if $Cn(\Upsilon)$ is the set of classical conclusions from the given scenario description, $C(\Upsilon)$ is the set of actual conclusions using a proposed theory and $Ci(\Upsilon)$ is the set of intended conclusions according to the chosen formal definition, then usually $Cn(\Upsilon) \subset Ci(\Upsilon)$ especially when a nonmonotonic logic is being used, and one requires from C that $C(\Upsilon) = Ci(\Upsilon)$. In this article I use the model-theoretic formulation since it is the more convenient one for analyzing logics based on *preferential semantics*, where the set of selected models is chosen (in the simplest case) as $Min(\ll, [\![\Upsilon]\!])$, i.e. as the set of those classical models which are minimal with respect to some preference relation $\ll$. Many of the proposed theories of action and change have this character.

The work of validating several theories for the same purpose is facilitated considerably if they use the same logic, which I will then call their *base logic*, as well as the same conventions for scenario description. In this case they can use the same intended-models function, and the validation results can be compared more easily.

In the case of reasoning about actions and change, the scenario description is a *chronicle* containing statements about the state of the world at various points in time (the observations), about specific actions and their temporal order (the schedule of actions), as well as action laws defining the effects of actions. Models represent possible histories of the world; those models which satisfy all the statements but violate the "frame" assumption of persistence or inertia are classical but unintended models.

Lin and Shoham [19] were probably the first to report a validation of a previously proposed theory of action. They showed that chronological minimization is correct for prediction in deterministic situation-calculus theories with strict inertia, i.e. no surprises or qualification, but allowing for ramification. The assumption of determinism applies for the ramification side-effects as well.

A number of later papers have obtained validations of other theories, and often for broader classes of chronicles. The present article, which combines and extends two previous conference papers [32, 33], reports validation of several theories of action for chronicles with *nondeterministic actions* and *discrete metric time*. This includes integer time as a special case, and allows one to represent actions with extended duration, which may be characterized not only in terms of their final effects, but also with respect to changes within their interval of duration. It also allows chronicles where the order of the actions is only partially specified, like in nonlinear planning [29]. Several of the validations generalize trivially for deterministic ramification, but ramification in nondeterministic chronicles is more difficult and has not been addressed.

We have shown in [34] that these results generalize easily to the case of branching metric time as well, but for simplicity the present article addresses the special case of integer time. I share the view on the utility of metric time which is presented by Ferguson and Allen in another article in the present Special Issue.

Other validation results for chronicles in situation calculus with strict inertia have been reported by Lifschitz [17], Reiter [28], and Kartha [14]. Validations for theories with ramification have been reported by Lin and Reiter [18] and by Peppas and Wobcke [25]. Validations for the generalization of situation calculus to theories with concurrent single-timestep actions have been reported by Lin and Shoham [20] and by Baral and Gelfond [3].

The validation results by these other authors are *direct validations*, in the sense that they prove that a certain theory of action is correct for all chronicles in a certain class, given a certain validation criterion. However the validations reported in the present article are *assessments*, which identify the range of correct applicability of each of a number of theories. We initially define a relatively large class of chronicles, and assign a *taxonomy* to it in terms of a number of properties that chronicles may have. The complete assessment of a theory of action consists of both a lower bound and an upper bound. The lower bound is a family of chronicles within the taxonomy for which the theory is provably correct; the upper bound is a family (larger or equal to the lower bound) outside which it is provably impossible that the theory is correct. In this way it has sometimes been possible to identify the exact range of applicability of previously proposed theories of action.

Different researchers have also used different validation criteria. Lin and Shoham [19], Lifschitz [17], and Reiter [28] relate a nonmonotonic theory over a certain set of "common-sense" axioms, to a monotonic theory over a larger set of axioms which is assumed to capture the full intentions. In other words, they define $\Sigma(\Upsilon)$ as $[\![\mathbf{G}(\Upsilon)]\!]$, where $\mathbf{G}$ is an operation which extends or transforms the given set of logical formulae.

Lifschitz and Gelfond have introduced a separate specification language $\mathcal{A}$ for defining the effects of actions, and used it to define the intended conclusions for reasoning about actions [10]. Kartha [14] applied the same criterion to the theories previously proposed by Baker [2], Pednault [24], and Reiter [28]. Baral and Gelfond [3] generalized $\mathcal{A}$ to the case of concurrent single-timestep actions and used it for validations as well.

These two approaches have only been used for direct validation. The work reported here uses an *underlying semantics* which captures basic notions of intelligent agents, and defines intended models in its terms. The underlying semantics can be viewed as a kind of abstract simulator of the interaction between an intelligent robot and its environment. It is hoped that this approach will facilitate its use for cognitive robotics as well as for CSR. The immediate benefit, however, is that the underlying semantics made it easy to define the chronicle taxonomy that is needed for assessments.

Yet another approach is to constrain the set of 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 [7] use the postulates of Katsumo and Mendelzon [15] as the basis, and Peppas and Wobcke [25] use an extended set of postulates for epistemic entrenchment as the basis for the correctness proof for their system. Since postulates are in principle written to constrain the choice of $\Sigma(\Gamma)$ but not to specify it completely, it seems that their use will lead to intrinsically weaker results.

Given this variety of validation criteria one may also consider to what extent they are equivalent and exchangable. A first step towards such an analysis has been reported by Thielscher [38], who identified the relationship between the $\mathcal{A}$ language of Gelfond and Lifschitz and a subset of the underlying semantics that is used in the present paper.

1.2 Intended models and common sense

The systematic methodology introduces a two-step process for going from common sense to logic: first one attempts to capture some characteristics of common sense in terms of a formal definition of $\Sigma(\Upsilon)$, and then one tries to find a theory whose selection function S satisfies $S(\Upsilon) = \Sigma(\Upsilon)$. But what is the evidence that a proposed definition of Σ itself captures common sense reasoning in an adequate manner?

In my view the formalization of common sense can only be achieved as the limiting case in a progression of successively more refined functions Σ_1, Σ_2, and so on. To illustrate the need for several such Σ_i, consider the "stolen car scenario", a very simple example of common-sense reasoning originally proposed by Kautz in [16]. In this chronicle it is assumed that if a certain car is left in the garage overnight then it will still be there the next morning. It is also observed that after the car has been left in the garage for two nights, it is no longer there. What are the common-sense conclusions, or equivalently what should be the set $\Sigma(\Upsilon)$ of intended models for this chronicle?

One possible ontology assumes simple persistence without any exceptions or modifications. I will refer to this case as *strict inertia*. Then the given chronicle is inconsistent, since the car is supposed to remain in the garage after two nights, but it is also stated not to be there. The set of intended models for the stolen car scenario under strict inertia is therefore empty. Another possible ontology allows persistence to be modified by "surprises", and at the same time one makes the *epistemological assumption* that $\Sigma(\Upsilon)$ must be non-empty except in the case where the set $[\![\Upsilon]\!]$ of classical models is itself empty. The set of surprises is minimized subject to the restriction $\Sigma(\Upsilon) \neq \varnothing$. In this case the given chronicle has two intended models: one where the car vanishes surprisingly during the first night, and one where the car vanishes surprisingly during the second night.

The well-known "qualification" and "ramification" problems concern additional ontological characteristics of CSR besides surprises. It appears that full CSR involves a considerable number of such difficulties: not only surprises, qualification, and ramification, but also concurrent events, causation chains of events, creation and termination of objects, unreliable observations, and so forth. All of them represent challenges for a logic of common-sense reasoning. However none of the proposed theories gets even close to addressing all of these phenomena, although several of them propose to address more than one phenomenon by one single technical method. In particular, some of the proposed preference relations on models aspire to establish inertia, to minimize side-effects (ramification), and to minimize surprises all at the same time.

What I propose to do, therefore, is to introduce one definition $\Sigma_{\mathbf{IA}}$ for strict inertia, a modified definition $\Sigma_{\mathbf{IAS}}$ for inertia with surprises but without any of the other characteristics, and so on. Unlike the proposed theories, the definitions of the various Σ_i should represent all the characteristics explicitly, so that one can clearly discern each of them.

This proposal has sometimes evoked impatience: why bother with a very simple case such as $\Sigma_{\mathbf{IA}}$ i.e. strict inertia, when other researchers have already thought about more advanced problems such as qualification and ramification? The answer is that the systematic assessment of a theory of action is considerably more complex than finding a few examples where it works, and the assessment is a non-trivial undertaking even for the case of strict inertia. One ought not have to apologize for the strategy of obtaining a solid analysis of the simple cases before one proceeds to the more difficult ones, even though one impatiently feels that the directly useful results will be found in the later stages of the process.

In this perspective, the reason for using the formal definition $\Sigma_{\mathbf{IA}}$ corresponding to strict inertia is not that it is claimed to correspond exactly to common sense – it clearly does not – but because it is the *first step* on a *well structured path* towards the goal of characterizing common sense reasoning; a goal which is very ambitious and which therefore we can not expect to achieve in the near future. For each choice Σ_i of the intended-models function, one obtains a well-defined assessment problem for one or more proposed theories of action, and this is a problem that can be analyzed in an entirely formal fashion.

Although each intended-models function Σ_i is to be *defined* in a formal fashion, the *choice* of that definition for the purpose of CSR is properly outside the scope of formal methods. The traditional methodology in A.I. would leave it to the discretion of the researcher: everyone presumably knows what common sense is? One may speculate whether methods of cognitive psychology could be used for testing the psychological validity of one or the other intended-models functions, but that is not a topic of the present article.

1.3 A methodology for assessment of range of applicability

Consider a given choice of ontological assumptions, in particular the case of strict inertia represented by its intended-models function $\Sigma_{\mathbf{IA}}$, and consider also a proposed theory viewed as a selection function S. A *direct validation* of S with respect to $\Sigma_{\mathbf{IA}}$ will simply verify whether $S(\Upsilon) = \Sigma_{\mathbf{IA}}(\Upsilon)$ for every chronicle Υ that is correctly formed with respect to the chosen ontological assumptions, and the outcome will state whether S is or is not always correct. However even if S is not always correct one may be interested in its *range of applicability*, i.e. in a family $\mathcal{Z}$ of chronicles such that $\Upsilon \in \mathcal{Z} \Rightarrow S(\Upsilon) = \Sigma_{\mathbf{IA}}(\Upsilon)$. Such results which I will call assessments have a potential practical interest: it is not clear that a broader-range theory will always be preferred, since a theory with a more narrow range of applicability might allow a more efficient implementation. However in order to make use of it, one must have precise knowledge of whether it is correctly applicable for the application one has at hand.

Range of applicability results have a theoretical interest as well: as we proceed to successive new definitions Σ_{i} of the intended models, a selection function that agrees completely with one version Σ_{i} can by definition not agree completely with the next one. An understanding of selection functions that are almost correct for one stage may therefore be useful for the next stage.

The range of applicability may be characterized in terms of various properties of the chronicle, for example whether it contains a nondeterministic action, whether it contains observations for time greater than zero, and so on. I will refer to such properties as *sub-characteristics* since they provide a finer-grain categorization in addition to the one that is obtained from the main characteristics for inertia, surprises, ramification, etc. The collection of characteristics and sub-characteristics together constitute a *taxonomy* of chronicles which may be used as the coordinate system for specifying and comparing the range of applicability of various theories of action.

Notice what is the logical order in which the various definitions must be made:

1. Define the reportoire of main characteristics.
2. Select some class of chronicles $\mathcal{Z}_i$ in terms of the chosen main characteristics.
3. Choose a definition of syntactic structure for chronicles in $\mathcal{Z}_i$.
4. Define the underlying semantics for $\mathcal{Z}_i$, and in particular the selection function Σ_i from well-formed chronicles to their intended model-sets.
5. The underlying semantics is also used for the formal definition of the sub-characteristics which generate a taxonomical structure within $\mathcal{Z}_i$.
6. Proposed theories of action for $\mathcal{Z}_i$ are expressed as selection functions and are assessed with respect to the selection function Σ_i. The results are expressed in terms of the taxonomy defined by the sub-characteristics.
7. Algorithms and other implementation aspects are developed for the theories that have been assessed.

Each definition of Σ_i is merely an approximation to real common sense. When the theories and implementations for one choice of $\mathcal{Z}_i$ is sufficiently well understood, one proceeds to a more refined class and repeats the process. One hopes of course that the analysis of a later class can make use of the results from earlier classes in the progression.

The present article summarizes some recent results in this research, and in particular it identifies the range of applicability of several of the previously proposed theories of actions with respect to strict inertia. In fact we obtain not only lower bounds but also upper bounds on the range of applicability, where the upper bounds identify regions where we can not even hope that a given theory will be correctly applicable. The previously published results are direct validations, and the possibility of range-of-applicability results is one of the contributions of the present approach.

By necessity the present article is fairly brief, and omits most of the proofs as well as some of the details in the definitions. For the full account please refer to the much more detailed presentation in a forthcoming book [34], hereafter referred to as "the book". A preliminary version is available as a departmental technical report.

2 The base logic

Before proceeding to the underlying semantics I shall outline the syntax and semantics of the logic, called *discrete fluent logic*, which will be used as the base logic. We consider first the "propositional" case, and then generalize to the "first-order" case which allows multiple objects each having time-dependent properties. The latter is essentially a two-sorted first-order logic, with "time-points" and "physical objects" as the two sorts. However I make certain modifications relative to the standard, in order to obtain a system which is more natural for knowledge representation, and in order to facilitate the subsequent analysis. This discrete fluent logic is the base logic so that several different theories of action will all use its syntax and classical semantics[1], and only differ with respect to their preference relations or more generally their selection functions on models. Basic choices such as the use of explicit metric time are intrinsic to the base logic, and are the reason why I do not use situation calculus as the base logic instead.

2.1 The propositional main language

Conventional propositional temporal logic allows each proposition symbol to have a truth-value as a function of time. In DFL, proposition symbols are replaced by *feature symbols* each of which is associated with its own, finite set of possible values called its *value domain*. For example, the current direction of movement of a particular automobile may be represented using a value domain with three members, for "forward", "backward", and "standing still".

If cd is the feature for the direction of movement of this particular car, and the value domain for cd is chosen as $\{\mathtt{F}, \mathtt{B}, \mathtt{S}\}$ with the obvious intended meanings, then the logic formula

$$[620]\, cd \mathrel{\hat{=}} \mathtt{B}$$

will represent that the car is moving backward at time 620. Propositional features such as "alive" or "loaded" have the binary value domain $\{\mathtt{T}, \mathtt{F}\}$, and $p \mathrel{\hat{=}} \mathtt{T}$ for a propositional feature is abbreviated simply as p.

The expression to the right of the $\hat{=}$ symbol can only be an explicit feature value; it is not allowed to be a variable or a composite expression. The reason for introducing these multi-valued features is that they are often very natural in knowledge representation, for example in the "discrete value spaces" of qualitative reasoning. They are of course easily intertranslatable with the conventional representation. In particular if each pair of a feature

[1] as distinct from the underlying semantics

and its value is considered as a truth-valued property of the world, then one may write the above expression as $\mathit{Holds}(cd \hat{=} \mathtt{B}, 620)$. However such a reformulation necessitates a considerable number of domain constraints.

The temporal prefix may use an integer timepoint such as 620, a timepoint constant symbol such as $\boldsymbol{t}_2$, or a temporal variable such as t_3. Constant symbols will be written in boldface font, variables in italics. The prefix may also refer to an open or closed interval rather than a single timepoint. For example,

$$[620, \boldsymbol{t}_2]\, cd \hat{=} \mathtt{F}$$

represents the statement that the car is moving forward at each individual timepoint between times 620 and $\boldsymbol{t}_2$ inclusive.

Ordering relations between timepoints are expressed in the usual way, for example as in the formula $\boldsymbol{t}_2 > 620$. Composition of formulae using propositional connectives and quantifiers works as usual.

An *interpretation* I is a pair $\langle M, R\rangle$ where M is a *valuation* that assigns values to the temporal constants, and R is a *history* that assigns a value to each combination of a feature and a timepoint. Thus $R(f, t)$ is a member of the the value domain of the feature f, and expresses the value of the feature at time t.

The *surface semantics* defines the truth of a logical formula in an interpretation $\langle M, R\rangle$ in the usual Tarskian style. The set $[\![\Gamma]\!]$ of classical models for a given set of logic formulae in the main language is likewise defined in the standard way as the set of all interpretations where all members of Γ are true.

The logic whose syntax and semantics has been outlined here will be referred to as the *main language*. One notices that the main language is not able to express statements about actions. Such statements are made in a separate *action language*.

2.2 The propositional action language

Actions are supposed to operate over intervals and are expressed using action symbols. The following is a simple example of an *action statement* in the propositional case.

$$[530, 560]\,\mathsf{Open\text{-}door}.$$

The action characterized by this statement does *not* apply over subintervals or superintervals of $[530, 560]$. In general an action statement is formed by an expression for a closed temporal interval in the main language, followed by an action expression which in the present case is just an action symbol. Actions are not represented in the interpretations, and we shall return to their semantics in a while.

2.3 The first-order main language

Conventional treatments of first-order logic make a point of strictly separating objects and constant symbols, so that objects may be values of constant symbols but may not themselves occur in the formulae. The point is of course to allow several constant symbols to refer to the same object, and to allow objects that are not named by any constant symbol. However many applications in A.I. proceed quickly to eliminating this variability of the object domain by introducing unique-names axioms and domain-closure axioms to guarantee that the object domain is isomorphic with the set of constant symbols. This is like paying for the cake and then throwing it away.

In DFL I use two disjoint sets of constant symbols for objects, called "object names" and "object constants", respectively. Unique-names axioms and domain-closure axioms (or the equivalent restrictions formulated in semantical terms) apply for object names but not for object constants. Thus if a particular chronicle uses the object names `#1`, `#2`, and `#3` and the object constants $\boldsymbol{o}_1$ through $\boldsymbol{o}_6$, then one assumes that `#1`, `#2`, and `#3` are different objects, and that each member of the object domain is one of them. In particular, each of the $\boldsymbol{o}_i$ must be equal to exactly one of `#1`, `#2`, and `#3`. The reason for this convention is that some temporal reasoning problems, for example TWEAK-style planning, require the choice of object to be postponed for as long as possible during the reasoning process, and formally such a choice of object will be represented as an object constant whereas its equality to one or the other object name is held open. (The symbol for the object-to-be-determined is called a "variable" in Chapman's article [4], but in the terms of logic is is better viewed as a constant). Object variables are written as o_i for example o_4, and are used for quantification as usual.

In the first-order case a *feature expression* is a feature symbol followed by a (possibly empty) list of arguments, each of which must be an object name, an object constant, or an object variable. For example if *dir* is a feature with one argument indicating the direction of movement of its argument object, then

$$[620]\,dir(\texttt{\#3}) \hat{=} \texttt{B}$$

would express that object `#3` is moving backwards at time 620. The signature (similarity type) for each use of the logic must therefore indicate the arity and the value domain for each feature symbol.

Since the set of object names is required to be isomorphic with the object domain, it is convenient to identify those two, and to assume that each object name *is* an object. The M component of an interpretation must therefore map temporal constants $\boldsymbol{t}_i$ to non-negative integers, and object constants $\boldsymbol{o}_i$ to object names `#`k. For the purpose of the R component, a *feature* is defined as an expression formed by a feature symbol followed by a possibly empty list of arguments, each of which must be an object *name*. Thus $dir(\texttt{\#3})$ is a feature but $dir(\boldsymbol{o}_3)$ is not.

A signature specifies the arity (number of arguments) and the set of possible values for each feature symbol. A signature and an object domain determine together a corresponding set of features, which will be denoted $\mathcal{F}$. Thus the R component of the interpretation is always a mapping from $\mathcal{F}$ times non-negative integers, to corresponding feature values. The rule for determination of the truth of a formula in a given interpretation is straightforward.

Since we are using a strictly semantical approach, it is natural to bypass the use of unique-names and domain-closure axioms, and to use instead an index on the function for forming the model set. Thus if Γ is a set of logic formulae and $\mathcal{O}$ is a set of object names taken as objects, then $[\![\Gamma]\!]_{\mathcal{O}}$ will represent the set of all models where all members of Γ are true and where furthermore the object domain (used in the construction of $\mathcal{F}$ and for the values of object constants in M) is exactly $\mathcal{O}$. The notation $[\![\Gamma]\!]_{\mathcal{O}}$ is of course only defined when all object names occurring in Γ are members of $\mathcal{O}$.

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. Following McCarthy, a *fluent* is a mapping from timepoints to corresponding values in a particular feature-value domain. Thus $\lambda t.R(f,t)$ is a fluent for each choice of the feature f.

2.4 The first-order action language

The action language is generalized to the first-order case by allowing arguments for action symbols in the obvious ways. The following is an example of an action statement

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

and may represent the action of painting the object $\boldsymbol{o}_5$ with green color during the time interval from 530 to 560. 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 $\boldsymbol{o}_5$.

The letter A will be used as a metavariable for action names, and E for action designators. Expressions will be quoted using Quine quotes [27]. For example we might have $E = \ulcorner\mathsf{Paint}(\texttt{\#5}, \texttt{G})\urcorner$. The full details of this syntax and semantics are given in the book.

3 Chronicles

Recall that a chronicle is a particular kind of scenario description which is used in the context of reasoning about actions and change. A chronicle will not be defined as a single set of logic formulae, but as a tuple whose structure depends on the ontological family, and whose elements are sets of formulae or other syntactic entities. In the case of strict inertia, a chronicle is defined as a fourtupe $\langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle$, where $\mathcal{O}$ is a finite set of objects/object names $\{\texttt{\#1}, \texttt{\#2}, ...\texttt{\#}k\}$ to be used in interpretations e.g. when constructing features. Π is a set of "laws" characterizing the effects of actions. $\textsc{scd}$ is a schedule i.e. a finite set of formulae characterizing the actions, chosen among the following two simple kinds:

- An action statement, i.e. a formula $[s,t]\, A(...)$,
- A *timing statement* i.e. a disjunction of one or more formulae of the form $s < t$ or $s \leq t$.

The expressions s and t are variable-free terms for timepoints. The arguments of A may be object names, object constants, 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 $\textsc{obs}$ is a set of observation statements, which can be chosen as any formulae in the main language. This means in particular that observations may not contain action statements.

These constraints may seem disturbing from the point of view of classical logic: why not allow any kind of formulae anywhere? However the constraints are actually used for the assessments of applicability. This does not mean that the same constraints will be there forever, but only that one has to relax them gradually. Consider for example the case of composite action statements of the form

$$\exists u [\, [s,u]\, A_1 \wedge [u,t]\, A_2]$$

or

$$(\, [s]\, \varphi \Rightarrow [s,t]\, A_1) \wedge (\neg [s]\, \varphi \Rightarrow [s,t]\, A_2),$$

which correspond to sequential composition and conditional expressions for actions. In order to account for chronicles containing such action statements one has to extend the underlying semantics and generalize the assessments of the entailment methods that have been proposed. The work has been done, in that particular case, but it was a nontrivial extension and it is not considered in the present article. Anyway, the examples of common-sense reasoning that have been used in the previous research are easily accomodated within the stated constraints.

Since the schedule contains both action statements and temporal constraints it can be viewed as a plan which specifies what to do and in what order. However its use is not restricted to planning future actions; it may also be used for describing past actions, which is why the term "schedule" was chosen. Notice that since the schedule is a finite set, it will only specify a finite number of actions. Since each action will in addition be assumed to have finite duration, it follows that in every intended model of the chronicle the changes are limited to a finite interval of time.

Notice also that although each chronicle commits to a particular choice of object domain, one will presumably wish to write the set of action laws so that it can be used for many different chronicles which may have different object domains. For example the action laws for moving blocks and for other actions in the blocks world are the same regardless of the number of blocks in the world. The binding of the object domain across models which was described above, is appropriate when reasoning about one particular chronicle, but not if one wishes to draw conclusions merely from the action laws and without reference to any particular chronicle. Consequently the object names in $\mathcal{O}$ may be used in the formulae in SCD and OBS, but not in the action laws Π.

Mnemonically, both the letters s and t will be used for timepoints, representing starting time and termination time of an interval or an action. This allows a set of action statements to be represented as $[\boldsymbol{s}_i, \boldsymbol{t}_i]\, E_i$ for varying i. (Consequently I use r rather than s to represent states).

I will write Π(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:

$$[s,t]\,\mathsf{Load} \Rrightarrow [t]\, l \hat{=} \mathtt{T}$$

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 ($\mathtt{T}$) at time t. If SCD has the member $[4,6]\,\mathsf{Load}$, Π(SCD) would then in its place have the member $[6]\, l \hat{=} \mathtt{T}$. 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 Π(SCD) is defined.

This treatment of the action laws has the advantage of simplifying the semantics and the preference relations on models. If actions were to be explicit objects in the logic then one would have to minimize both the set of actions and the set of changes resulting from the actions (Morgenstern and Stein [22]). Such a more powerful theory may in fact be necessary for expressing causation chains between events, or the credibility of actions or events, but that again would be a later development.

4 Underlying semantics

The underlying semantics has two related purposes: to define the function $\Sigma_{\mathbf{IA}}$, and to define the subcharacteristics mentioned in subsection 1.3 and thereby the taxonomy of chronicles. For each chronicle $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle$, $\Sigma_{\mathbf{IA}}(\Upsilon)$ shall be a set of interpretations $\langle M, R \rangle$ which represent those developments of the world which are in accordance with Υ and with the assumption of strict inertia: if $R(t)$ is considered as a function from time-points to states, then there shall be no changes over time in R except those which are explicitly specified by the action statements in SCD and the corresponding action laws in Π. The definition of $\Sigma_{\mathbf{IA}}$ should furthermore satisfy the following informal criteria:

- It should be intuitively natural from the point of view of common-sense reasoning.
- It should be possible to relate it to the underlying, physical and quantitative world description that may be available in cognitive-robotics applications.
- It should be straight-forward to generalize it later on to more complex characteristics of CSR, such as those involving surprises, qualification, and ramification.

We first define the underlying semantics for a broader ontological family **IAD** which allows certain types of ramification, and then obtain the case of strict inertia by imposing a particular constraint on it.

4.1 Trajectory structures

Technically, the stated goals are achieved using constructs on two levels. On the more specific level, each action is characterized using a set of *trajectories*. Each trajectory is a sequence of partial states representing the successive feature values in the affected part of the world, during successive timepoints while the action is being executed. For example, the trajectory

$$\langle \{p : \mathtt{T}, w : \mathtt{F}\}, \{p : \mathtt{T}, w : \mathtt{T}\} \rangle$$

represents performing an action in two timesteps, where after the first timestep p is true and w is false, and after the second timestep both p and w are true.

This arises for example if the phone starts ringing (p) in the first time-step and the person in the room becomes awake (w) in the second time-step. All other features in $\mathcal{F}$ besides those that are defined in the trajectory, are assumed to stay constant.

For each combination of an action designator E and a state r, the semantics identifies a set of trajectories, representing all possible ways of performing E if it starts at a time when the world in state r. This allows the outcome of the action to depend on the state of the world when the action starts, and it allows the action to be non-deterministic with respect to its duration, its final outcome, and the intermediate states while it is being performed. For a given action and state of the world, all partial states in all the corresponding trajectories are required to be defined over the same subset of $\mathcal{F}$.

The construction of a history R, for a given chronicle Υ and a given valuation M, is defined using a *game* between an "ego" (corresponding to the "knowledge level" of an intelligent agent proposed by Newell in [23]) and a "world" that comprises its physical environment and the sensori-motoric system of the robot itself. This game represents the higher level construct above the level of the trajectories. It is reminiscent of the "agent model" of Genesereth and Nilsson [11].

4.2 The ego-world game

It is commonplace [1, 5, 36] to propose that intelligent robots should be designed with a layered architecture. Lower software layers receive input from the sensors, recognize patterns and properties of the world on a discrete level, and send them to the knowledge level, where they are used as the basis for prediction, planning, and other types of reasoning. The knowledge level is then able to invoke actions by emitting a request that a particular action is to be performed by lower layers of the architecture, which naturally use their sensory input for this purpose as well. In the simplest case the action is just treated as a subroutine that is allowed to run until it terminates, and in this case the knowledge level will not be engaged again until the action has been completed. The resulting history of actions will therefore be *sequential*. In less trivial cases the knowledge level may also interact with the execution of the action at some other specific times, and supervise the execution of several concurrent actions. It should however be isolated from the continuous details of the actions. When actions are viewed in this way they have *duration* and *internal structure*, and this is the reason for our choice of base logic.

The "game" between ego and world in the underlying semantics is an abstract version of this architecture. The ego corresponds to the knowledge level; the world corresponds to the combination of the lower levels of the robot and its physical environment. Thus the interaction between ego and world consists only of (a) the world informing the ego about what has recently happened in the world, in particular as the result of actions that have been invoked by the ego; (b) the ego requesting the world to perform certain actions.

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) which accumulates the set of interaction times between the ego and the world. Its largest member n is referred to as the "*now*" of the ongoing game. 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 $s \leq n$.

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 $\mathcal{J} \rightarrow \mathcal{J}$ 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 ego-world game offers a general framework which can be constrained in various ways in order to obtain particular types of systems. To begin with there are some minimal restrictions on $\mathbf{W}$ and $\mathbf{K}$ which assure that they are not able for example to run time backwards. In particular, the ego must leave the first and third component of the development unchanged, so it can not change the setting of n ("now") at all. The moves of the ego are to add elements $\langle n, E \rangle$ to $\mathcal{C}$, meaning to invoke actions starting at the present now-time n, or to do nothing. The moves of the world are to increase n by adding more members to $\mathcal{B}$, 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 $n' > n$ by removing $\langle n, E \rangle$ from $\mathcal{C}$, adding n' to $\mathcal{B}$, and adding $\langle n, E, n' \rangle$ as an additional member of $\mathcal{A}$. If the ego has left $\mathcal{C}$ empty i.e. chosen to do nothing, then the world adds $n{+}1$ as a new member of $\mathcal{B}$, causing "now" to be increased by one, and extends R according to $R(n{+}1) = R(n)$.

It is straightforward to constrain the rules of the game so that the actions are sequential in all resulting developments: the agent is allowed to add at most one member to $\mathcal{C}$, and the world must always remove that member in its moves. Concurrent actions are obtained if $\mathcal{C}$ is allowed to contain several members at the same time, and the ego and the world are allowed to add and delete members asynchronously. Only sequential actions will be considered here.

In this article the set $\mathcal{B}$ is only used for its largest member. However the full set plays a role in certain assessment proofs, which are induction proofs over the members of $\mathcal{B}$, as well as in a generalization to branching time.

4.3 The use of trajectories in the ego-world game

The use of the trajectory semantics constitutes an additional specialization in the framework of the ego-world game. We proceed to the precise definition of a world $\mathbf{W}$ with strict inertia, using the trajectory structures defined in subsection 4.1. 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 and no others. A *world description* for a world with strict inertia is a pair $\langle \mathtt{Infl}, \mathtt{Trajs}\rangle$, where $\mathtt{Infl}$ (influenced features) and $\mathtt{Trajs}$ (trajectories) are functions such that if E is an action designator and r is a state, then $\mathtt{Infl}(E, r) \subseteq \mathcal{F}$, and $\mathtt{Trajs}(E, r)$ is a non-empty set of trajectories for $\mathtt{Infl}(E, r)$. $\mathtt{Trajs}(E, r)$ is the set of possible ways that the action can be realized in the world. $\mathtt{Trajs}(\ulcorner A(...)\urcorner, r)$ for a fixed action symbol A but varying its arguments and the starting state r, corresponds to the action law for the action named A.

A world description $\langle \mathtt{Infl}, \mathtt{Trajs}\rangle$ specifies a world $\mathbf{W}$ with the following behavior in the game with sequential actions. If the ego has invoked the action designator E at the 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 $\mathtt{Trajs}(E, R(s))$. The world's move is to modify the current finite 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 $s{+}k$. The case where $\mathcal{C} = \varnothing$ when the world has the move, has already been defined.

Several world descriptions may exhibit the same behavior, for the following reasons. Let $\langle \mathtt{Infl}, \mathtt{Trajs}\rangle$ be given and choose some E and some r such that $\mathtt{Infl}(E, r) \subset \mathcal{F}$ (strict subset relation). Add one more member f to $\mathtt{Infl}(E, r)$, and add the maplet $f \mapsto r(f)$ to every partial state in every trajectory in $\mathtt{Trajs}(E, r)$. It is easily seen that the new world description exhibits exactly the same behavior as the given one – it just specifies the persistence of f explicitly rather than by default. In the limiting case where $\texttt{Infl}(E, r) = \mathcal{F}$ for every E and r one does not rely on inertia at all.

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.

4.4 Locality of change

Locality of change is a restriction that avoids certain types of ramification. It is also useful for the assessments since it makes it easier to relate a world description $\langle \texttt{Infl}, \texttt{Trajs} \rangle$ to a set of action laws. It is only applicable in the "first-order" case where features and actions may have objects as arguments; in the propositional case it is trivially satisfied. This means in particular that it is trivially satisfied in all the standard examples of common sense reasoning in the "frame problem" literature.

The idea is as follows. Consider a world description $\langle \texttt{Infl}, \texttt{Trajs} \rangle$ where $E = \ulcorner A(o) \urcorner$ is an action designator for an object o, and where some of the features in $\mathcal{F}$ have the structure $\ulcorner f(o) \urcorner$. Both actions and features may have several arguments, but we use the simplest case for the example. For example when o is chosen as `#4`, $\texttt{Infl}(\ulcorner A(\texttt{\#4}) \urcorner, r)$ must be defined for every r and shall be a set of features.

Locality of change requires first of all that all members of $\texttt{Infl}(\ulcorner A(\texttt{\#4}) \urcorner, r)$ are features which either have no arguments, or only have `#4` as an argument. In general, we define the *range of influence* of an action designator E as the set of features all of whose arguments occur also as arguments of E, and we require for locality of change that $\texttt{Infl}(E, r)$ shall be a subset of the range of influence of E.

Secondly it is required that the effects of the action do not depend on features outside its range of influence. In other words if r and r' are two states whose restrictions to the range of influence of E are equal, then $\texttt{Infl}(E, r) = \texttt{Infl}(E, r')$ and $\texttt{Trajs}(E, r) = \texttt{Trajs}(E, r')$.

Thirdly it is required that the effects of an action are independent of the naming of the objects. Thus if one selects an arbitrary permutation π of the object domain and applies this permutation throughout $\langle \texttt{Infl}, \texttt{Trajs} \rangle$ then one obtains back the same world description. This means concretely speaking that if one should wish the effects of an action to depend on features of an object with a particular role (such as "the king") then one must not just set aside one object `#i` for that role. It becomes necessary to introduce a constant symbol for it, and to include that constant as an argument in one particular argument position of the action.

The combined effect of these restrictions is to guarantee that $\texttt{Trajs}(E, r)$

can be characterized using a formula of finite size, which in turn is important for the assessment proofs.

Locality of change is related to the informal concept of ramification in the "frame problem" literature. Some of the frequently cited examples of ramification are "if *alive* ceases to be true, then *dead* ceases to be false", and "if *alive* ceases to be true, then *walking* ceases to be true". If several objects are being considered at once then one would of course write, in DFL notation, that if $alive(o)$ changes its value then $dead(o)$ changes its value. Such constraints between features of one and the same object are compatible with locality of change. 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 can not be directly accomodated when locality of change is assumed.

4.5 Full trajectory normal form

The purpose of the world description $\langle \mathtt{Infl}, \mathtt{Trajs} \rangle$ is to represent the behavior of the world in a way which can be directly applied in the course of the ego-world game. It therefore corresponds to the action-law component Π of a chronicle. For a given chronicle $\langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle$ one wishes to consider the (unique) world corresponding to Π, and the set of games that can be played by that world.

One possible approach would have been to define the truth of a set Π of action laws in a world description $\langle \mathtt{Infl}, \mathtt{Trajs} \rangle$. Unfortunately such an approach seems to require a fairly complex definition of action-law syntax. Therefore we have chosen instead to define a particular format for the action laws, called *full trajectory normal form* or FTNF, which corresponds directly to the world description. Analogous to full disjunctive normal form in propositional logic, the FTNF expresses the action laws on a case by case basis. For example, if A is an action without arguments then its FTNF action law has the form

$$[s,t]\, A \Rrightarrow \bigwedge_r \Theta_r$$

where r ranges over the set $\mathcal{R}$ of all states, and Θ_r has the form

$$[s]\, \varphi_r \Rightarrow (\psi_{r,1} \vee \psi_{r,2} ... \vee \psi_{r,k}),$$

where φ_r is a formula of the form

$$f_1 \mathrel{\hat{=}} x_1 \wedge f_2 \mathrel{\hat{=}} x_2 ... \wedge f_n \mathrel{\hat{=}} x_n,$$

where again $x_i = r(f_i)$ for each i. Therefore φ_r is a formula that exactly characterizes r as a possible starting state of the action. Similarly there is one expression $\psi_{r,i}$ for each possible ending state r_i of the action when r is the starting state. The expression $\psi_{r,i}$ characterizes r_i and the set of all trajectories in $\mathtt{Trajs}(A, r)$ that end in r_i. For example if the action is deterministic for the given r then there is only one $\psi_{r,i}$. Constraints on the duration of the action and on the timing of various changes within its duration must be expressed by the various $\psi_{r,i}$.

The direct transformation between trajectory sets and logic formulae is the starting point for the assessments of theories of actions in this framework, and for the purpose of those proofs it is very convenient to have the direct structural relationship between trajectory sets and action laws. On the other hand, the FTNF is useless from a practical point of view due to its bulkyness, so the resulting assessments would be vain if they only applied to those chronicles where the action laws are in fact expressed in FTNF. Fortunately there is no such restriction, because the action laws Π are only used for obtaining $[\![\Pi(\textsc{scd})]\!]$ i.e. for their models as applied to a particular schedule. The assessments are therefore equally applicable if the action laws are replaced by another version as long as the expressions to the right of the $\Rightarrow$ operator is classically logically equivalent to the FTNF version.

When actions with object arguments are used then $\mathtt{Trajs}(E, r)$ will be defined for an infinite number of action designators E, and possibly for an infinite number of states r, since the world description must apply for all possible choices of object domain. The locality of change assumption is important in this context for insuring that $\mathtt{Trajs}(E, r)$ for an infinite number of argument combinations can be described by a finite formula.

There are a number of additional technical details which will not be adressed here, such as the existence of the formula $\psi_{r,i}$ when the corresponding set of trajectories has infinitely many members. Each trajectory is finite, but there need not be any upper bound on their lengths. Those topics are made precise in the book.

4.6 Intended models under the assumption of strong inertia

The definition of the set $\Sigma_{\mathbf{IA}}(\Upsilon)$ of the intended models of Υ with respect to strict inertia can now be finalized. It is assumed that $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle$ is a correct chronicle for strict inertia, and the definition proceeds as follows. Select an arbitrary world $\mathbf{W}$ which is exactly described by Π. This is certainly possible since the action-law component Π was required to be equal to or logically equivalent to an FTNF expression, which in turn was defined to be constructed from some world description $\langle \mathtt{Infl}, \mathtt{Trajs} \rangle$. 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 where object constants are mapped into $\mathcal{O}$. Generate all possible developments which can be obtained in games between

$\mathbf{W}$ and $\mathbf{K}$, 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 formulae in $\Pi(\textsc{scd}) \cup \textsc{obs}$ 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 that satisfy the restrictions. Then $\Sigma_{\mathbf{IA}}(\Upsilon)$ is defined as the set of those interpretations $\langle M, R\rangle$ for which $\langle \mathcal{B}, M, R, \mathcal{A}, \mathcal{C}\rangle \in W$ for some $\mathcal{B}$, $\mathcal{A}$, and $\mathcal{C}$.

The relationship between $\mathcal{A}$ and SCD in this definition is made precise as follows. If M is a valuation and τ is a variable-free temporal term then $M[\tau]$ is the value of τ using M, and similarly for object terms. $M[[\varsigma,\tau]\, A(\varphi_1, \ldots \varphi_k)]$ is defined as the triple $\langle M[\varsigma], \ulcorner A(M[\varphi_1], \ldots M[\varphi_k])\urcorner, M[\tau]\rangle$. $M[\textsc{scd}]$ is obtained as the set consisting of $M[[\varsigma,\tau]\, A(\varphi_1, \ldots \varphi_k)]$ for every action statement in SCD. The requirement for intended models is that $M[\textsc{scd}] = \mathcal{A}$.

One consequence of this definition is that if SCD contains a certain number of action statements $[s_i, t_i]\, E$ for different i but the same E, then the development must contain similar actions in one-to-one correspondence. It is not possible for one action statement to have several realizations in $\mathcal{A}$, or for several statements to "share" one realization.

We notice that although this definition uses the trajectory mechanism for defining the behavior of the "world" side of the game, it does not directly define or restrict the "ego" side. There is only the posterior restriction that the action set $\mathcal{A}$ resulting from the game shall equal what is stipulated in the schedule part of the chronicle. In future extensions, some of the more complex characteristics of chronicles may require prior restrictions on the ego behavior as well. This may arise for example when one uses statements about the beliefs and behavior patterns of the ego, and is a topic for continued research.

5 The ontological and epistemological taxonomy

In order to assess and compare the range of applicability of various theories of action, one needs a common "coordinate system": a taxonomy of chronicles whereby one can specify where a theory is applicable and where it is not. It was proposed above to organize such a taxonomy using a distinction between main characteristics and sub-characteristics. Orthogonally we need a distinction between ontological and epistemological characteristics. Ontological characteristics refer to properties of the world which is described in the chronicle and which participates in the ego-world game. An example of such a property is whether actions are non-deterministic or not. Epistemological characteristics refer to assumptions about the description, for example whether there are any observations for time later than zero.

Ontological characteristics constrain the action laws and their interpretation. Epistemological characteristics constrain similarly the other three components of the chronicle – in themselves or in conjunction with the action laws. The term "characteristics" is used since the word "feature" is already being used for a category in the logic.

5.1 Main characteristics

The ontological characterization of classes of worlds uses sets of bold-face letters, where each letter indicates the presence of some characteristics. Strict inertia is denoted by **IA**, where **I** stands for inertia as such, and **A** for "alternative results": the results of an action are conditional on the starting state, for example when firing the gun in the Yale Shooting Problem. **C** represents that concurrent actions are allowed, **L** that actions may have delayed effects (resulting in changes that occur after the period of the action itself), and so on.

The trajectory-based ego-world game that was defined above without the restriction of locality will be denoted as **IAD**. The narrower ontological family **IA** corresponding to strict inertia is obtained by also imposing strict locality. The case without alternatives will be denoted by **I**.

Epistemological characteristics will use script letters. I 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, namely that each action has a known precondition, and in case the precondition is satisfied there is full knowledge about the possible effects of the action, otherwise not.

The family of chronicles with specified epistemological and ontological characteristics will be represented by hyphen-joining the two characteristics, as in $\mathcal{K}$-**IA**. Strict inertia in the frame problem literature is concerned with the $\mathcal{K}$-**IA** family of reasoning problems. Classical planning algorithms such as TWEAK [4] 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.

5.2 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. **IAd** denotes the subfamily of **IA** where all actions are deterministic, and so on.

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

The assessment of a proposed theory of action is always done within the framework of a given (main) characteristics, according to the sequence of steps defined in section 1.3. For example, 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 any chronicle $\Upsilon \in \mathcal{K}\text{-}\mathbf{IA}$, and given also a selection function S, one wishes to find a set $\mathcal{Z} \subseteq \mathcal{K}\text{-}\mathbf{IA}$ for which one can prove that every $\Upsilon \in \mathcal{Z}$ satisfies that $\Sigma_{\mathbf{IA}}(\Upsilon) = S(\Upsilon)$. If it is known that the theory 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 for each choice of M the initial state $R(0)$ is completely determined by OBS. Similarly $\mathcal{K}\mathsf{p}$ represents that there are no observations about any timepoint after the initial one. The combination of the last two conditions is written as $\mathcal{K}\mathsf{sp}$.

Any given chronicle determines its own epistemological subcharacteristics. The particular epistemological subcharacteristics 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 characteristics and not a sub-characteristics.

6 Assessments of some simple entailment methods

The following are the assessments of some previously proposed theories of action. Each of them is first reformulated using the base logic i.e. the syntax and the surface semantics that were defined above. After this reduction the theory is only characterized by its selection function S which maps chronicles to model sets. As has already been discussed, such a selection function 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 $\Upsilon \in \mathcal{Z}$. A theory of action is therefore correct for chronicles with strict inertia without restrictions if this equality holds for every $\Upsilon \in \mathcal{K}$-**IA**.

Throughout this section, Υ will represent a chronicle $\langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle \in \mathcal{K}$-**IA** which satisfies the syntactic rules imposed on chronicles for strict inertia. $[\![\Upsilon]\!]$ will represent the set of classical models of Υ, already defined as $[\![\Pi(\textsc{scd}) \cup \textsc{obs}]\!]_{\mathcal{O}}$ i.e. the set of all interpretations $\langle M, R \rangle$ where all statements in Υ are satisfied, no inertia assumption is imposed, and the object domain is $\mathcal{O}$.

6.1 Safe sequentiality and the notion of entailment method

The underlying semantics was defined so that all actions must occur sequentially, i.e. no intended model can contain two actions $\langle s, E, t \rangle$ and $\langle s', E', t' \rangle$ where $(s, t] \cap (s', t']$ is nonempty. A chronicle containing a schedule with for example the two action statements $[5, 15]\,A_1$ and $[10, 20]\,A_2$ will therefore not have any intended models. However it may have some classical models, and any selection function which is defined in terms of a preference ordering will then produce a nonempty set of selected models for it. A chronicle Υ will be said to be *safely sequential* iff the actions in its schedule $\textsc{scd}$ occur sequentially according to any valuation M that satisfies the timing statements in $\textsc{scd}$.

It is therefore reasonable to arrange that preference orderings are only applied to models of safely sequential chronicles. We observe that if Υ is an arbitrary chronicle, it is easy to construct a safely sequential chronicle Υ' such that $[\![\Upsilon']\!]$ consists exactly of the sequential models in $[\![\Upsilon]\!]$. The construction is obtained by adding appropriate temporal constraints to the schedule component of Υ. It is hard to think of a way of achieving the same purpose using a preference relation on models.

It follows that the semantical method of model-preference is not suitable for the entire task of replicating $\Sigma_{\mathbf{IA}}$. An *entailment method* is consequently defined as a pair $\langle \mathbf{G}, S \rangle$ where $\mathbf{G}$ is a syntactic transformation such as the one mapping Υ to its safely sequential counterpart Υ', and S is a selection function i.e. a mapping from a chronicle to a set of models such that $S(\Upsilon) \subseteq [\![\Upsilon]\!]$. We shall define and assess a number of such entailment methods, where the syntactic part in several cases only serves to obtain safely sequential chronicles, but in one case it performs other tasks as well. The base logic, $\mathbf{G}$, and S are together a theory of action in the sense defined initially.

If $\langle \mathbf{G}, S \rangle$ is an entailment method then S will be called its *essential* selection function and $\lambda\Upsilon[S(\mathbf{G}(\Upsilon))]$ will be called its *full* selection function. The entailment method $\langle \mathbf{G}, S \rangle$ is correct for $\mathcal{Z}$ iff its full selection function is correct for $\mathcal{Z}$. To prove this it is sufficient to show that S is correct for the projection of $\mathcal{Z}$ by $\mathbf{G}$, and that $\Sigma_{\mathbf{IA}}(\Upsilon) = \Sigma_{\mathbf{IA}}(\mathbf{G}(\Upsilon))$. In particular if $\mathbf{G}$ is the transformation that ensures safe sequentiality then it is sufficient to show that S is correct for all safely sequential members of $\mathcal{Z}$. This will be the form of all the assessment results that follow.

6.2 Some previously proposed entailment methods

The following proposition is a direct consequence of the definitions of $\Sigma_{\mathbf{IA}}(\Upsilon)$ and $[\![\Upsilon]\!]$, and confirms our expectation with respect to classical and intended models. Anyway it has obtained its proper proof.

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

Any correct selection function must therefore produce a subset of $[\![\Upsilon]\!] = [\![\Pi(\textsc{scd}) \cup \textsc{obs}]\!]_{\mathcal{O}}$. In retrospect, the reason why non-monotonic logics are used for this purpose is that if $\textsc{scd} \subseteq \textsc{scd}'$ it does not necessarily follow that

$$\Sigma_{\mathbf{IA}}(\langle \mathcal{O}, \Pi, \textsc{scd}', \textsc{obs} \rangle) \subseteq \Sigma_{\mathbf{IA}}(\langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle)$$

as would be the case if $\Sigma_{\mathbf{IA}}$ were monotonic. Several proposed theories of action define their (essential) selection function 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.

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

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

for all safely sequential $\Upsilon \in \mathcal{K}$-**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 $\textsc{scd}$ and $\textsc{obs}$ are empty, and let Υ' be the chronicle where $\textsc{obs}$ is empty and $\textsc{scd}$ 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.

An anonymous reviewer suggested that this observation might be due to the fact that the information about the action itself has been removed from the models, which would indicate a possible way of remedying the problem.

Unfortunately this is not the correct diagnosis, for the following reason. We revise the example so that it contains two propositional features p and q corresponding to two tossable coins. Υ and Υ' are constructed with the same schedule consisting of the single formula $[4,5]A$. However the action laws in Υ and Υ' are different: in Υ the effect of A is to toss p, in Υ' it is to toss q. This means that the same set of actions $\{\langle 4, A, 5\rangle\}$ is obtained for both schedules. Still they have the same set of classical models and different sets of intended models.

The proposition generalizes immediately to saying that no function S on model sets satisfies $S([\![\Upsilon]\!]) = \Sigma_{\mathbf{IA}}(\Upsilon)$ for all $\Upsilon \in \mathcal{K}$-**IA**. But it is anyway meaningful to assess the range of previously proposed theories, both because of their historical interest and since they may be useful in applications that do not require the full range of $\mathcal{K}$-**IA**.

The original proposal by McCarthy [21] was rephrased by Hanks and McDermott and used for a counterexample in their Yale Shooting Problem paper [12]. This method uses global (non-chronological) minimization of abnormality. If one assumes as in e.g. [26] or [18] that abnormality *only* arises when change actually occurs ("strong abnormality"), then the abnormality is completely determined by $\langle M, R\rangle$ and the preference relation can be expressed as follows. Let $I = \langle M, R\rangle$ and $I' = \langle M', R'\rangle$ be two members of $[\![\Upsilon]\!]$. The changeset of I is defined as

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

i.e. the set of pairs of a feature f 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, whose essential selection function is

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

The complementary case of "weak abnormality" will be defined and assessed in the next section. Among the various proposals for solving the Yale Shooting Problem, the proposal by Kautz [16] can easily be rewritten using our base logic as follows[2], and will be called *original chronological minimization*, OCM. The breakset of I at time t is defined as

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

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

[2]The present formulation treats positive and negated propositions symmetrically, and in that respect it is a modification of Kautz's actual proposal. I take this liberty since the proposal was only expressed in terms of an example.

- 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 essential selection function S_{ocm} as

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

Some counterexamples to OCM can be remedied by the following modification of the definition, in line with the original approach of Shoham [37], and 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 essential selection function S_{pcm} as

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

The requirement $M = M'$ is essential in these selection functions, and counterexamples are easily constructed for the variant where it is omitted.

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 conditions, please refer again to the book.

6.3 Assessment of original chronological minimization

The following is a simple example of assessment of applicability.

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

Recall that $\mathcal{K}$sp-**IsAd** is the set of chronicles where the initial state is completely specified by OBS for each choice of M, there are no observations for times other than 0, and all actions are deterministic and take a single step of time.

Proof Consider a $\mathcal{K}$sp-**IsAd** chronicle $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle$ and some $\langle M, R \rangle \in [\![\Upsilon]\!]$. For this choice of M, $R(0)$ must be the only state that satisfies the observations for time zero, according to the epistemological assumptions. Since the actions are deterministic and take a single timestep there is exactly one intended model for the given M. Let $I' = \langle M, R' \rangle$ be that intended model. By proposition 1 it is a member of $[\![\Upsilon]\!]$. It is easily verified that if another classical model with the same M has a smaller ($\subset$) breakset for some t, then it has a larger one ($\supset$) for some $t' < t$. Therefore $I' \in S_{ocm}(\Upsilon)$, so that all intended models are selected.

To prove that all selected models are intended, suppose there is also some other member $I'' = \langle M, R'' \rangle \in S_{ocm}(\Upsilon)$ which contains the same valuation M and which is not a member of $\Sigma_{\mathbf{IA}}(\Upsilon)$. Let $\mathtt{t}$ be the smallest timepoint where $R'(\mathtt{t}) \neq R''(\mathtt{t})$. Clearly $\mathtt{t} > 0$. The states $R'(\mathtt{t})$ and $R''(\mathtt{t})$ can not differ in any of the features affected by an action ending in $\mathtt{t}$, because then I'' would not belong to $[\![\Upsilon]\!]$ since actions are assumed deterministic. But I', being an intended model, does not change in any feature except those necessarily changed by an action. Therefore $breakset(I', \mathtt{t}) \subset breakset(I'', \mathtt{t})$ from which it follows $I' \ll_{ocm} I''$. This contradicts the assumption that $I'' \in S_{ocm}(\Upsilon)$, which concludes the proof.

It is fairly easy to find examples where the stated restrictions are violated, and where OCM does not obtain the intended set of models. However it is possible to strengthen this result by introducing another kind of subcharacteristics.

6.4 Preference-compatibility

Let $\langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle$ be a chronicle, let $\langle \mathtt{Infl}, \mathtt{Trajs} \rangle$ be the trajectory-structure counterpart of the action laws Π, and consider the set of trajectories $\mathtt{Trajs}(E, r)$ which characterizes all the possible ways of executing E from the starting state r. Let k be an arbitrary integer, and construct the set V_k of all sequences

$$\langle r, r \oplus r'_1, r \oplus r'_2, \ldots r \oplus r'_k \rangle$$

where $\langle r'_1, r'_2, \ldots r'_k \rangle \in \mathtt{Trajs}(E, r)$ is a trajectory of length k. Every such sequence characterizes a fragment of history over an interval where E is performed, including the starting time with the starting state r as well as the ending time. The trajectory structure is said to be *OCM-compatible* iff $V_k = Min(\ll_{ocm}, V_k)$, for any V_k that is constructed by a choice of E, r, and k. This requirement means that the preference relation $\ll_{ocm}$ will not prefer any member of V_k over any other. The analogous compatibility properties for PCM and PGM are defined as well.

Preference-compatibility plays a natural role in the assessment of a preferential-entailment method: since the set of trajectories is intended to characterize the set of possible ways of performing the action, we certainly do not wish a preference relation to prefer one of them over another. Lacking preference-compatibility one loses not only completeness but also soundness, since some of the intended models may be lost. We therefore consider the different variants of preference-compatibility as subcharacteristics, on a par with "single-timestep", "deterministic", etc which were introduced above. The notation [OCM] will be used for OCM-compatibility, so for example $\mathcal{K}$-**IsA**[OCM] is the family of OCM-compatible, single-timestep chronicles with strict inertia.

Another subcharacteristics, "equidurationality", is defined as follows. Let Υ be a chronicle and let $\langle \mathtt{Infl}, \mathtt{Trajs} \rangle$ be the trajectory-semantics description of the action laws in Υ. Let

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

i.e. the set of possible durations (expressed as the number of elements in the trajectory) when the action designator E is invoked from the starting state r. A world description is *equidurational* for an action designator E iff $\mathtt{Pdur}(E, r)$ is independent of r. Υ is equidurational, and has the subcharacteristics **IAe**, iff its world description is equidurational for all E.

The new taxonomical concepts make it possible to obtain stronger assessments than the previous ones. It is clear for example that

$$\mathcal{K}\textsf{sp-}\mathbf{IsAd} \subseteq \mathcal{K}\textsf{sp-}\mathbf{IA}[\textsc{ocm}],$$

since the set V_k used above will have at most one member if E is deterministic and takes a single timestep. It may be difficult to get a good intuition for what preference-compatibility means, but the following behavior, called *orchestration*, is one way of looking at it. Let $\mathtt{Infl}$, $\mathtt{Trajs}$, E, r, and $k > 0$ be given. For every feature f in $\mathtt{Infl}(E, r)$ there shall be a set $N_f \subseteq [1, k]$ i.e. a set of integers such that, in any trajectory in $\mathtt{Trajs}(E, r)$ of length k, the value of f changes exactly at the times indicated by N_f. This means that the time(s) of change for a given feature within the action are uniquely determined by the chosen duration of the action, so it is the same in all members of V_k. It means in particular that the action may nondeterministically change the feature f to one value or another, but it may not make a nondeterministic choice between changing or not changing the value of f. Also it may not be nondeterministic about when the value changes. A concrete special case as an example is if one feature f_1 always changes its value in the first timestep of the action while another one, f_2, always changes its value in the last timestep of the action.

If $\langle \mathtt{Infl}, \mathtt{Trajs} \rangle$ satisfies these conditions for each choice of E, r, and k then we shall say that it is *locally orchestrated.* Local orchestration is a sufficient condition for OCM-compatibility, but not a necessary one.

We state without proof an improved assessment for OCM using these concepts.

Proposition 4 Original chronological minimization (OCM) is correct for $\mathcal{K}$sp-**IAde**[OCM].

The proof is found in the book.

6.5 Assessment of prototypical chronological minimization

The assessment of prototypical chronological minimization (PCM) is expressed immediately in terms of preference-compatibility.

Proposition 5 Prototypical chronological minimization (PCM) is correct for $\mathcal{K}$p-**IAe**[PCM].

For the proof, please refer to the book. Comparing with OCM, we have been able to drop the requirement of determinism and that the initial state must be completely determined. It is also easy to verify that any OCM-compatible chronicle is PCM-compatible. At this point the proven range of applicability of OCM is therefore a subset of the range for PCM. However these assessments are only lower bounds. We return to upper bounds in a later section.

The analysis of Lin and Shoham, which used another methodology (compare section 1), considered the $\mathcal{K}$p-**IsAd** family of reasoning problems, and they proved that PCM is correct for that family. The present result for PCM confirms and generalizes theirs.

6.6 Assessment of prototypical global minimization

The assessment of prototypical global minimization (PGM) requires some additional sub-characteristics. A world described by $\langle \texttt{Infl}, \texttt{Trajs}\rangle$ is said to be *PGM-perfect* iff it satisfies the following requirement. Let $v \in \texttt{Trajs}(E, r)$ and $\bar{v} \in \texttt{Trajs}(E, \bar{r})$ be trajectories of the same length and for the same E but not necessarily the same starting state. If $v = \langle r'_1, r'_2, ... r'_k\rangle$ then construct

$$v' = \langle r, r \oplus r'_1, r \oplus r'_2, ... r \oplus r'_k\rangle$$

like in the definition of preference-compatible, and similarly for $\bar{v}$ starting from $\bar{r}$. Then it is required that $changeset(v') = changeset(\bar{v}')$.

This condition means that delay actions (having no effect except the passage of time) and toggle-type actions are allowed. A simple deterministic action such as loading the gun (if the gun was unloaded it becomes loaded, if it was already loaded then nothing happens) is not PGM-perfect. PGM-perfection requires the kind of orchestration that was described for preference-compatibility above, but it must be applied globally across varying r, and not only locally for each specific choice of r.

Secondly we introduce an additional epistemological subcharacteristics as follows. A chronicle Υ is said to be *pointwise consistency-retaining* iff

$$\langle M, R\rangle \in [\![\Upsilon]\!] \Rightarrow \exists R'[\langle M, R'\rangle \in \Sigma_{\mathbf{IA}}(\Upsilon)].$$

This means that if the chronicle has a classical model then it has an intended model, and this condition holds individually for each choice of M. The stolen car scenario is an example of a chronicle that is not consistency-retaining, neither pointwise nor globally. The family of consistency-retaining chronicles in $\mathcal{K}$-**IA** will be written $\mathcal{K}$r-**IA**. We have proved:

Proposition 6 Prototypical global minimization is correct for PGM-perfect chronicles in $\mathcal{K}$r-**IA**.

The requirement of PGM-perfection is a very strong one, but assessment of upper bounds has shown that it is not possible to improve very much on it.

7 Theories of action using occlusion and filtering

The theories of action 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 theory that gives correct results for a larger class of chronicles while still retaining the preferential character, one must find alternatives to the straightforward selection functions the form

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

or equivalently

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

where the notation $[\![\Gamma]\!]_{\mathcal{O}}$ was defined to mean the restriction of the set of classical models for the set Γ, to those members where the object domain is $\mathcal{O}$. *Filtering* and *occlusion* are two such improvements which were introduced by the present author in [30].

7.1 Filtering

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

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

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

$$S_{pcmf}(\langle\mathcal{O},\Pi,\textsc{scd},\textsc{obs}\rangle) = Min(\ll_{pcm}, [\![\Pi(\textsc{scd})]\!]_{\mathcal{O}}) \cap [\![\textsc{obs}]\!].$$

We have proved

Proposition 7 PCMF is correct for $\mathcal{K}$-**IAe**[PCM].

Recall that PCM is correct for $\mathcal{K}$**p-IAe**[PCM], i.e. for equidurational and PCM-compatible chronicles where all observations are for time 0. The introduction of filtering in PCMF removes the third one of those restrictions. In the case of PGM, filtering does not particularly improve the range of applicability.

7.2 Occlusion

Occlusion 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 of preference-compatibility.

Occlusion can be understood as a refinement of the usual abnormality concept as applied to the frame problem. Already McCarthy in his 1984 article proposed to use an axiom for "no change without abnormality", which can be written as follows in situation-calculus notation

$$\neg Ab(f,A,s) \Rightarrow (Holds(f,s) \Leftrightarrow Holds(f,Result(A,s))),$$

and which says that if a feature is not abnormal (with respect to inertia) in a situation then its value must be unchanged there. Correspondingly in our logic, abnormality would be associated with time-points, and the nochange axiom takes the form of the axiom schema

$$\neg\,[s]\,Ab(f) \Rightarrow ([s\text{-}1]\,f \mathrel{\hat{=}} \xi \Leftrightarrow [s]\,f \mathrel{\hat{=}} \xi),$$

where ξ is a meta-variable. This axiom schema is used in conjunction with the minimization of Ab. Since it applies for any ξ one can replace $\Leftrightarrow$ by $\Rightarrow$.

Now suppose an action has the effect of changing the value of a specific feature f. The action law must specify the new value of f, but various proposals in the literature differ with respect to whether the abnormality itself shall be explicitly implied in the action law. The following is an example of an action law where the implication is explicit

$$[s,t]\,A \Rightarrow s{+}1 = t \wedge [t]\,f \mathrel{\hat{=}} \mathtt{G} \wedge [t]\,Ab(f),$$

and the following is a law where abnormality is not explicitly specified:

$$[s,t]\,A \Rightarrow s{+}1 = t \wedge [t]\,f \mathrel{\hat{=}} \mathtt{G}.$$

The McCarthy proposal of 1984 used explicit implication of abnormality, and Hanks and McDermott followed his proposal, but already Kautz in 1986 made the opposite choice. This does not make any difference if the world and the formulation of the action laws are such that abnormality only arises when change actually occurs. In this case the nochange axiom schema may be replaced by

$$\neg\,[s]\,Ab(f) \Leftrightarrow (\,[s\text{-}1]\,f \mathrel{\hat{=}} \xi \Leftrightarrow [s]\,f \mathrel{\hat{=}} \xi)$$

without affecting the set of selected models. This case has been called strong abnormality. The opposite case using explicit implication of weak persistence-abnormality has other properties, and this is what I call *occlusion* in order to have a shorter name. Note by the way that abnormality can be used for other purposes besides for inertia.

One of the interesting aspects of occlusion is that it makes it possible to have two levels of abnormality. Whereas the nochange axiom stipulates that there can be no unoccluded (abnormal) changes, one may instead use a selection function where unoccluded changes are *minimized* rather than forbidden. Unoccluded changes are then the changes that are not stated by the action laws. This alternative is of interest for broader classes than $\mathcal{K}$-**IA**, since different types of changes require different treatment. For examples, changes due to the regular effects of actions may be minimized chronologically, whereas "surprises" should be minimized non-chronologically. However within $\mathcal{K}$-**IA** it does not matter much whether unoccluded changes are minimized or restricted by nochange axioms.

In brief, the occlusion concept provides a way of characterizing and reasoning about those combinations of features and temporal intervals where strict inertia *does not* apply.

From the point of view of the underlying semantics, the occlusion information is the same as is expressed by the `Infl` component in the trajectory semantics. It requires the following modification of the base logic. Interpretations are extended from $\langle M, R\rangle$ to $\langle M, R, X\rangle$ where X representing occlusion 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 $(s,t]$. For example the action law for tossing the coin as a single-timestep action will just state that the action takes a single timestep, and that the position of the coin is occluded at the end of the action. For actions where the behavior of occluded features is not entirely random, the action law specifies constraints on their development as well.

The occlusion operator is therefore allowed to be used in the action laws, whereby it also appears in $\Pi(\textsc{scd})$, but not in the set of observations $\textsc{obs}$. The definition of the classical model operator $[\![.]\!]$ is modified so that $[\![\Upsilon]\!]$ is now a set of triples $\langle M, R, X\rangle$, and the definition of correctness for a selection function is changed accordingly so that it requires $\mathsf{Ci}(S([\![\Upsilon]\!])) = \Sigma_{\mathbf{IA}}(\Upsilon)$, where $\mathsf{Ci}(W) = \{\langle M, R\rangle \mid \langle M, R, X\rangle \in W\}$. In this way the occlusion component X can be used by the preference relations, but not in the intended conclusions from the chronicle. Note that proposition 2 does not apply after these redefinitions.

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:

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

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

Here $[s,t]\,p := \mathtt{F}$ is an often-useful 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.

7.3 Chronological minimization of occlusion and change

An entailment method 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 $\mathcal{K}$-$\mathbf{IA}$, 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

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

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:

- $X(\mathtt{t}) \subset X'(\mathtt{t})$
- $X(\mathtt{t}) = X'(\mathtt{t}) \wedge brs(I, \mathtt{t}) \subset brs(I', \mathtt{t})$,

where $X(t) = \{f \mid X(f,t)\}$. The set of selected models is defined using filtering, with the essential selection function

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

We have proved

Proposition 8 CMOC is correct for $\mathcal{K}$-**IAe**.

In fact the minimization will result in empty breaksets at all points in time. Comparing to PCM and PCMF, we have now eliminated the requirement of PCM-compatibility, and only the requirement of equidurationality remains. For an example of why equidurationality is required, see the proof of proposition 12.

The CMOC method was tested against a 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 entailment 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. Of course if the duration of actions is not being described, as in situation calculus, then this restriction is trivially satisfied.

7.4 Chronological assignment and chronological minimization of occlusion and change

It is possible to correct the remaining problem of CMOC and to obtain a chronological entailment 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 can be met using a slightly more complex syntactic transformation before the essential selection function is applied.

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 temporal restrictions from the SCD to the OBS partition of the chronicle, which of course does not affect $[\![\Upsilon]\!]$ or $\Sigma_{\mathbf{IA}}(\Upsilon)$. SCD may keep those temporal restrictions that are needed for safe sequentiality. However additional constant symbols may need to be introduced in the course of this syntactic transformation. We assume some technical arrangement so that they are not "externally visible".

A schedule which satisfies this syntactic restriction will be called *simply executable*. For each simply executable schedule there exists an ego such that the schedule is true in all games played by that ego against any world. The same does not hold for non-executable schedules.

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 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 with the essential selection function

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

We have proved

Proposition 9 CAMOC is correct for $\mathcal{K}$-**IA**.

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.

7.5 Global minimization of occlusion with no unoccluded change

Before concluding we shall also return to the original McCarthy-Hanks-McDermott method. If that method is used with weak abnormality then it corresponds in our terms to *global minimization of occlusion with no unoccluded change* or GMON which has the following essential selection function

$$S_{gmon}(\Upsilon) = Min(\ll_{gmon}, [\![\Pi(\textsc{scd}) \cup \textsc{obs} \cup \textsc{nch}]\!]_{\mathcal{O}}),$$

where $\ll_{gmon}$ is defined so that $\langle M, R, X\rangle \ll_{gmon} \langle M', R', X'\rangle$ iff $M = M'$ and $X \subseteq X'$. Also $\textsc{nch}$ is the set of all instances of the nochange axiom schema, i.e. all axioms of the form

$$[t\text{-}1]\, f \mathrel{\hat{=}} \xi \wedge \neg [t] \mathbb{X} f \Rightarrow [t]\, f \mathrel{\hat{=}} \xi.$$

The assessment of GMON uses the following subcharacteristics. A chronicle Υ is said to have *uniform change* and belongs to $\mathcal{K}$-**IAu** iff $\mathtt{Infl}(E, r) = \mathtt{Infl}(E, r')$ for all E, r, and r'. This condition means that the set of features for which an action law specifies the changes, is independent of the starting state. This characteristics is not only a property of the world itself (the player in the ego-world game), but also of how the action laws are written. For example in the Yale Shooting Problem, Hanks and McDermott characterized the effects of firing the gun so that "if the gun is fired when loaded then in the resulting situation Fred is dead and the aliveness of Fred is abnormal". (I omit the effect on the loadedness of the gun). This formulation of the action law does not satisfy uniform change. However the law can also be expressed as follows: "If the gun is fired then the aliveness of Fred is abnormal; if the gun was loaded then Fred results being dead, and if the gun was not loaded then the aliveness of Fred is retained". This formulation of the action law satisfies uniformity.

We have proved

Proposition 10 GMON is correct for $\mathcal{K}$r-**IAu**.

The actual result in the book uses a concept of GMON-perfection similar to PGM-perfection, and the uniformity requirement is a special case.

The corresponding method of *filtered global minimization of occlusion with no unoccluded change* or GMONF is defined by the following selection function

$$S_{gmonf}(\Upsilon) = Min(\ll_{gmon}, [\![\Pi(\textsc{scd})]\!]_{\mathcal{O}}) \cap [\![\textsc{obs} \cup \textsc{nch}]\!],$$

where $\textsc{nch}$ is as before. We have proved

Proposition 11 GMONF is correct for $\mathcal{K}$-**IAu**.

Clearly any action law can be rewritten in uniform form, and often it does not grow very much in size as a result. The approach does however presume weak abnormality. It follows that the nonmonotonic entailment method that was proposed by McCarthy and used with negative results by Hanks and McDermott, is actually correct for the Yale Shooting Problem if the action laws are written in uniform form. It follows also that with the simple modification of filtering, the same method is correctly applicable for all chronicles with strict inertia provided that action laws are written in uniform form. These observations are somewhat surprising in view of the great attention that has been given to the YSP over the years.

8 Upper bounds on range of applicability

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 $\mathcal{Z} \subseteq \mathcal{K}$-**IA** is an upper bound on the range of applicability for an entailment method whose main selection function is S iff $\Sigma_{\mathbf{IA}}(\Upsilon) \neq S(\Upsilon)$ for any $\Upsilon \in \mathcal{K}$-**IA** $- \mathcal{Z}$. This is however a fairly useless definition, since there are often a number of cases where an entailment method so to say accidentally obtains the intended models. The only exception so far is for pointwise consistency-retention: entailment methods such as PCM which use straight preferential minimization (without filtering), and where the preference relation only compares models with equal M, will fail to obtain the intended model set for any schedule that does not have pointwise consistency-retention. For other restrictions, a more useful definition can be obtained along the following lines.

Consider first the method CMOC, which has been proven correct for $\mathcal{K}$-**IAe** i.e. for equidurational chronicles.

Proposition 12 If $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle \in \mathcal{K}\text{-}\mathbf{IA} - \mathcal{K}\text{-}\mathbf{IAe}$, then there is some modified safely sequential chronicle $\Upsilon' = \langle \mathcal{O}, \Pi, \textsc{scd}', \textsc{obs}' \rangle$ such that $\Sigma_{\mathbf{IA}}(\Upsilon') \neq S_{cmoc}(\Upsilon')$.

Proof Let the set Π of action laws correspond to the world description $\langle \mathtt{Infl}, \mathtt{Trajs} \rangle$. In this case there is some action designator E, some states r and r', and some trajectory $v \in \mathtt{Trajs}(E, r)$ of length k such that no trajectory $v' \in \mathtt{Trajs}(E, r')$ 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 $\mathtt{[}1, k{+}1\mathtt{]}\, E$ and $\textsc{obs}'$ as $\mathtt{[}0\mathtt{]}\, \phi$, where ϕ is a formula that precisely characterizes the state r'. Then Υ' does not have any intended models according to $\mathcal{K}$-**IA**. However the interpretation $\langle M, R \rangle$ where M is arbitrary, $R(0) = r'$, $R(1) = r$, and R then continues according to v is a CMOC-preferred model.

Of course if $\Upsilon \in \mathcal{K}\text{-}\mathbf{IA} - \mathcal{K}\text{-}\mathbf{IAe}$ then $\Upsilon' \in \mathcal{K}\text{-}\mathbf{IA} - \mathcal{K}\text{-}\mathbf{IAe}$ 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 chronicles 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 $\mathcal{K}$-**IAe**[PCM]. Here the situation is similar:

Proposition 13 If $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle \in \mathcal{K}\text{-}\mathbf{IA} - \mathcal{K}\text{-}\mathbf{IAe}[\textsc{pcm}]$, then there is some modified safely sequential chronicle $\Upsilon' = \langle \mathcal{O}, \Pi, \textsc{scd}', \textsc{obs}' \rangle$ such that $\Sigma_{\mathbf{IA}}(\Upsilon') \neq S_{pcmf}(\Upsilon')$.

Proof Since $\Upsilon \in \mathcal{K}\text{-}\mathbf{IA} - \mathcal{K}\text{-}\mathbf{IAe}[\textsc{pcm}]$, it is either not equidurational or not PCM-compatible. In the former case the proof of proposition 12 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 $\texttt{[}0,k\texttt{]}E$ as its only member, and $\textsc{obs}'$ as $\texttt{[}0\texttt{]}\phi$ where ϕ completely characterizes r. The result follows immediately.

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

Proposition 14 If $\Upsilon = \langle \mathcal{O}, \Pi, \textsc{scd}, \textsc{obs} \rangle \in \mathcal{K}\mathsf{p}$-$\mathbf{IA} - \mathcal{K}\mathsf{p}$-$\mathbf{IAe}[\textsc{pcm}]$, then there is some modified safely sequential chronicle $\Upsilon' = \langle \mathcal{O}, \Pi, \textsc{scd}', \textsc{obs}' \rangle$ which also has the epistemological sub-characteristics p, and such that $\Sigma_{\mathbf{IA}}(\Upsilon') \neq S_{pcm}(\Upsilon')$.

Proof The proof for proposition 13 applies for this case also.

Therefore, as long as we stay within the epistemological constraint $\mathcal{K}\mathsf{p}$-$\mathbf{IA}$, the assessment $\mathcal{K}\mathsf{p}$-$\mathbf{IAe}[\textsc{pcm}]$ 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 $\texttt{[}t\texttt{]}\phi$ where $t > 0$ but ϕ is a tautology. One would therefore be interested in an upper bound in the epistemological dimension, if such a concept can be defined. The following would seem to be an appropriate framework. Let S be a selection function, let $\mathcal{K}\xi$-Ψ be a chronicle family formed using the ontological family Ψ and the epistemological subcharacteristics ξ, and where S is correct for $\mathcal{K}\xi$-Ψ. Then:

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

Figure 1 illustrates these definitions, with the simplifying assumption that the epistemological constraints only refer to the $\textsc{scd}$ and $\textsc{obs}$ components of the chronicle. The horizontally striped rectangle in the upper-left indicates how the ontologically maximal assessment works. For every chronicle Υ that satisfies $\mathcal{K}\xi$ but not Ψ (arbitrary position on any of the horizontal lines) there is some chronicle Υ' with the same action laws and therefore the same world (other position on the same line and still within $\mathcal{K}\xi$) where the entailment condition obtains the wrong result (dot on the line). The vertically striped

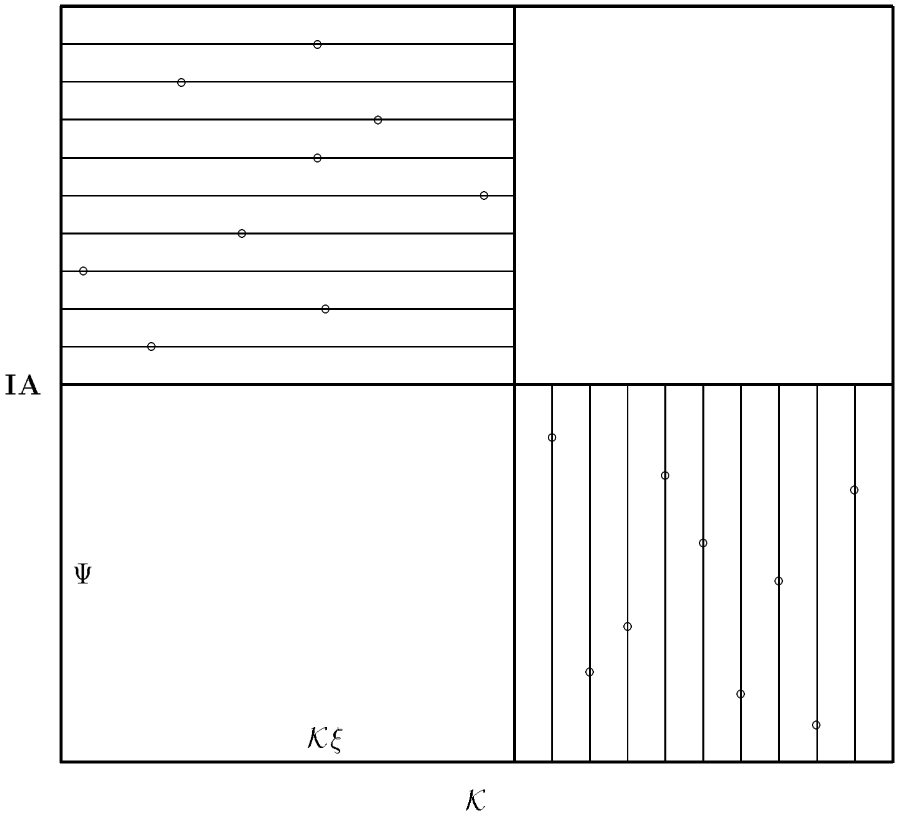

Figure 1: `Maximal assessments` [gfh]

rectangle in the lower-right indicates similarly how the epistemologically maximal assessment works. For every chronicle Υ that satisfies Ψ but not $\mathcal{K}\xi$ (arbitrary position on any of the vertical lines) there is some chronicle Υ' with the same schedule and observations but not necessarily the same world (other position on the same line) where anyway the world in Υ' is still in Ψ, and where the entailment condition 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 positively assessed region which is represented by the lower-left rectangle, neither upwards while retaining its right rectangular edge, or rightwards while retaining its upper edge.

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, $\mathcal{K}\mathsf{p}$-$\mathbf{IAe}$[PCM] is ontologically maximal but not epistemologically maximal. The full epistemological range for PCM relative to $\mathbf{IAe}$[PCM] allows for example chronicles of the form

$$[0]\,\phi$$

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

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

where $\phi \models \psi$ and with some additional constraints. The exact upper bound in the epistemological dimension may be the topic of further research.

The assessment results for these entailment methods are summarized in table 1[assum]. For each entailment method the table indicates the assessment that has been stated here, whether that assessment is ontologically maximal, and whether it is epistemologically maximal. The code [PGM*] stands for PGM-perfect.

The first four entailment methods in the table represent methods that do not use occlusion whereas the last four use occlusion. Since the methods in the latter group clearly have a broader range of applicability, one may wonder whether the first group has any advantages besides the hypothetical efficiency advantage in implementations. One possible advantage occurs in the context of ramification. As long as all effects of an action are directly stated in its action law, it is no disadvantage to have to state occlusion explicitly, and often one can easily accept to write the uniform action laws required by GMON. However if effects are also stated implicitly using domain constraints, then it becomes much less straightforward to also express how occlusion can be implied by those domain constraints. On the other hand if the domain constraints do not imply occlusion as well, then the range

Entailment criterion	Current assessment	Ontologically maximal?	Epistemologically maximal?
PGM	$\mathcal{K}$r-**IA**[PGM*]	No	Yes
OCM	$\mathcal{K}$sp-**IAde**[OCM]	No	No
PCM	$\mathcal{K}$p-**IAe**[PCM]	Yes	No
PCMF	$\mathcal{K}$-**IAe**[PCM]	Yes	Yes
CMOC	$\mathcal{K}$-**IAe**	Yes	Yes
CAMOC	$\mathcal{K}$-**IA**	Yes	Yes
GMON	$\mathcal{K}$r-**IAu**	No	Yes
GMONF	$\mathcal{K}$-**IAu**	No	Yes

Table 1: `Summary of assessments` [assum]

of applicability reverts presumably to what has been obtained in the first group of entailment methods. Note however that this discussion is merely speculative, and that at this point we do not even have a precise definition of intended models in the presence of ramification.

9 Additional results

Section 1.3 proposed a strategy for systematic research on ontology-specific entailment methods. The present paper has described the general approach and the first generation of results that have been obtained with it. It is intended to use these results as the basis for the continued work. The following results are already complete and are described in the book [34]:

- Lower-bound and upper-bound assessments for additional entailment methods, including one method (PMON) using the action schema NCH which is correctly applicable throughout $\mathcal{K}$-**IA** without requiring action laws to be written in uniform form.
- The assessments in this paper have been concerned with a completeness criterion, requiring equality between intended models and selected models. In some cases one may be content with a soundness criterion, only requiring the $\subseteq$ relation. Soundness assessments for a number of entailment methods are obtained as a by-product of the completeness assessments.
- A generalization of the time domain from integer time to metric branching time. It means in particular that actions theories based on the situation calculus are subsumed by these results.

- A generalization to the case where action statements in schedules may be formed using sequential composition, conditional expressions, and while-loops.

In a related project, Doherty and Łukaszewicz [8, 9] have provided a syntactic characterization of most of these entailment methods in terms of generalized-pointwise and second-order circumscription. It appears that the resulting representations are further translatable to logic programs using the method of Gelfond and Lifschitz [10]. This could be an essential step towards efficient implementations of these entailment methods. Karlsson [13] has modified the underlying semantics to apply for the $\mathcal{Q}$ epistemological assumption, where the effects of an action are only known when its precondition is satisfied, and using it he has verified the TWEAK planning algorithm [4]. Work on other extensions is in progress.

10 Related work

In section 1.1 we mentioned some other approaches to the validation of theories of action and change, and related them to the work reported here. Many other theories of action have been proposed besides those that have been assessed at present. Several of them have difficulty with what may loosely be called disjunctions, i.e. when the initial state is incompletely specified or when non-deterministic actions are allowed. A summary and analysis of a number of proposals is found in [31, 35]. An assessment of the range of applicability of such previously proposed approaches would be of considerable interest.

Many of the previously proposed entailment methods claim to solve a broader range of common-sense reasoning problems, particularly the ramification problem and/or the qualification problem. Correct handling of surprises, as in the stolen car problem, is also often considered to be a part of the general "frame problem". The Motivated Action Theory of Morgenstern and Stein [22] addresses causal chains between actions or events. In order to assess entailment methods of these kinds, it would be necessary first to define the underlying semantics and the set of intended models for the additional ontological characteristics that are being introduced. It appears, not surprisingly, that it is much easier to propose new and promising theories of action, than to analyze their properties in a systematic fashion.

Finally a note on the competitive viewpoint which has dominated the present field of research at least since 1985. Each author has advanced his or her own theory of action, and defended it by pointing out its virtues and the limitations of other, "competing" approaches. This was a natural consequence of the prevailing methodology, but it does not seem to have moved the field very much forward.

The systematic methodology changes the perspective. Once a theory of action has been studied carefully enough so that its range of applicability is known, there are two ways of using the results. One may use it as a basis for continued theoretical study, but then there is no reason to force a choice. If two theories both obtain exactly the intended set of models for a given class of chronicles, then they are equivalent and therefore exchangable. Both can be used as tools for continued study.

On the implementation side, it is anyway unlikely that a general purpose theorem-prover for any of the proposed theories would have adequate efficiency. Regardless of which model preference relation or which additional axioms (for example in explanation closure) are used in the theory, one must find special-purpose data structures or other computational devices which capture the essence of time, change, and inertia, and which contribute to good computational properties. It is an open question whether those computational constructs can best be seen as implementations of model minimization, or of explanation closure axioms, or both.

11 Summary and acknowledgements

The results in this paper are on two levels. A new, systematic methodology was defined in the first half of the article. The second part of the paper presented a number of hard results which have been obtained using this methodology. In particular we have shown that although several widely referenced theories of action are not correct for unrestricted use, for certain classes of chronicles they *are* provably correct.

The systematic methodology has made it possible to identify and assess the CAMOC entailment method. It has been proven correct for all $\mathcal{K}$-**IA** problems i.e. all problems with strict inertia, also allowing for actions with extended duration in time, constraints on the timing of changes during the action's duration, and non-deterministic actions.

Several previously proposed theories of action have turned out to be more restricted than was previously known or intuitively obvious; others turned out to be more widely applicable than was previously recognized. It has been possible to identify upper bounds as well as lower bounds on their range of applicability, and in several cases the proven upper and lower bounds are equal so that the range of correct applicability has been identified completely within the family of chronicles with strict inertia.

The work reported here has benefited greatly from discussions and comments from colleagues and students working on the same topic, namely

Patrick Doherty, Witold Lukaszewicz, Lars Karlsson, Hua Shu, and Choong-ho Yi. The group's research is supported by the funding agencies TFR and NUTEK. Our participation in the Esprit projects "DRUMS II" and "Logic and Change" has also contributed to the work described here. The constructive comments of the Journal's reviewers are likewise acknowledged. The later part of the work on this article was done during a stay at LAAS-CNRS in Toulouse, France, thanks to a grant from the CNRS.

References

[1] P.J. Antsaklis and K.M. Passino, editors. *An Introduction to Intelligent and Autonomous Control*. Kluwer, 1992.

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

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

[5] J.H. Connell. SSS: a hybrid architecture applied to robot navigation. In *International Joint Conference on Artificial Intelligence*, pages 2719–2724, 1991.

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

[7] 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.

[8] Patrick Doherty. Reasoning about actions and change using occlusion. In *European Conference on Artificial Intelligence*, pages 401–405, 1994.

[9] Patrick Doherty and Witold Lukaszewicz. Circumscribing features and fluents pointwise. In *International Conference on Temporal Logic*, 1994.

[10] Michael Gelfond and Vladimir Lifschitz. Representing actions in extended logic programs. In *International Conference on Logic Programming*, pages 559–573, 1992.

[11] Michael R. Genesereth and Nils J. Nilsson. *Logical Foundations of Artificial Intelligence*. Morgan-Kaufmann Publishing Co., 1987.

[12] Steve Hanks and Drew McDermott. Nonmonotonic logics and temporal projection. *Artificial Intelligence*, 33(3):379–412, 1987.

[13] Lars Karlsson. DFLEAK, planning with features and fluents. Master's thesis, Linköping University, Sweden, 1993.

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

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

[16] Henry Kautz. The logic of persistence. In *Proc. AAAI 1986*, pages 401–405, 1986.

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

[18] Fangzhen Lin and Ray Reiter. State constraints revisited. *Manuscript*, To appear.

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

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

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

[22] Leora Morgenstern and Lynn Andrea Stein. Why things go wrong: a formal theory of causal reasoning. In *National (U.S.) Conference on Artificial Intelligence*, pages 518–523, 1988.

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

[24] 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.

[25] 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.

[26] Javier Andrés Pinto. *Temporal Reasoning in the Situation Calculus.* PhD thesis, University of Toronto, February 1994.

[27] W.V. Quine. *Mathematical Logic, Revised Edition.* Harvard University Press, 1981.

[28] 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.

[29] Earl Sacerdoti. The non-linear nature of plans. In *International Joint Conference on Artificial Intelligence*, pages 206–214, 1975.

[30] Erik Sandewall. Filter preferential entailment for the logic of action in almost continuous worlds. In *Proc. International Joint Conf. on Artificial Intelligence, Detroit, USA*, 1989. .

[31] Erik Sandewall. Features and fluents. Review version of 1991. Technical Report LiTH-IDA-R-91-29, IDA, 1991.

[32] Erik Sandewall. Nonmonotonic temporal logics and autonomous agents: each contributes to the rigorous basis of the other. In O. Herzog et al, editor, *Grundlagen und Anwendungen der Künstlichen Intelligenz. Proc. German A.I. conference.* Springer Verlag, 1993.

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

[34] Erik Sandewall. *Features and Fluents.* Oxford University Press (to appear), 1994.

[35] Erik Sandewall and Yoav Shoham. Nonmonotonic temporal reasoning. In Dov Gabbay, editor, *Handbook of Artificial Intelligence and Logic Programming (to appear).* Oxford University Press, 1994.

[36] G.N. Saridis. Intelligent robotic control. *IEEE Transactions on Automatic Control*, 28(5):547–557, 1983.

[37] Yoav Shoham. *Reasoning about Change.* MIT Press, 1988.

[38] Michael Thielscher. An analysis of systematic approaches to reasoning about actions and change. In *Manuscript, Fachbereich Informatik, TH Darmstadt*, 1994.