On the reasoning of real-world agents: Toward a semantics for active logic

Michael L. Anderson
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

Walid Gomaa
Department of Computer Science
University of Maryland
College Park, MD 20742

John Grant
Department of Computer and Information Sciences, Department of Mathematics
Towson University, Towson, MD 21252

Don Perlis
Department of Computer Science
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

Abstract

The current paper details a restricted semantics for active logic, a time-sensitive, contradiction-tolerant logical reasoning formalism. Central to active logic are special rules controlling the inheritance of beliefs in general, and beliefs about the current time in particular, very tight controls on what can be derived from direct contradictions ($P\&\neg P$), and mechanisms allowing an agent to represent and reason about its own beliefs and past reasoning. Using these ideas, we introduce a new definition of model and of logical consequence, as well as a new definition of soundness such that, when reasoning with consistent premises, all classically sound rules are sound for active logic. However, not *everything* that is classically sound remains sound in our sense, for by classical definitions, all rules with contradictory premises are vacuously sound, whereas in active logic not everything follows from a contradiction.

1 Introduction

Real agents have some important characteristics that we need to take into account when thinking about how they might actually reason logically: (a) their reasoning takes time, meaning that agents always have only a limited, evolving awareness of the consequences of their own beliefs[1], and (b) their knowledge is imperfect, meaning that some of their beliefs will need to be modified or retracted, and they will inevitably face contradictions and other inconsistencies. The challenge from the standpoint of classical logical formalisms is that, if an agent's knowledge base can be inconsistent, then according to classical logic, it is permissible to derive *any* formula from it.

This fact about classical logics is commonly known by the Latin phrase *ex contradictione quodlibet*: from a contradiction everything follows. However, Graham Priest has coined the somewhat more vivid term *explosive logics*: a logic is explosive iff for all formulas A and B, $(A\&\neg A) \rightarrow B$. Priest defines a paraconsistent logic precisely as one which is not explosive [Priest, 2002; Priest *et al.*, 1989; Priest and Tanaka, Summer 2004]. Now, clearly real agents cannot tolerate the promiscuity of belief resulting from explosive logics, and must somehow maintain control over their reasoning, watching for and dealing with contradictions as they arise. The reasoning of real agents, that is, must be paraconsistent. But what *sort* of paraconsistent logic might agents usefully employ, what methods might agents use to control inference and deal with contradictions, and how can these logics (and methods) be modeled in terms of truth and consequence in structures?

In the current paper we are primarily interested in the last of these questions. For some time we have been developing, and have had significant practical success with, a time-sensitive, contradiction-tolerant logical reasoning formalism called active logic [Elgot-Drapkin and Perlis, 1990; Miller and Perlis, 1993; Nirkhe *et al.*, 1997; Purang, 2001]. Here we offer a start on a semantics for such a logic. We hope and expect it will be of interest as a specific model of formal reasoning for real-world agents that have to face both the relentlessness of time, and the inevitability of contradictions.

2 Active logic

One of the original motivations for active logic was the need to design formalisms for reasoning about an approaching deadline; for this use it is crucial that the reasoning take into account the ongoing passage of time as that reasoning proceeds. Thus, active logic reasons one

[1] Levesque's distinction between *explicit* and *implicit* beliefs [Levesque, 1984] points to this same issue; however, our approach is precisely to model the evolving awareness itself, rather than trying to model the full set of (implicit) consequences of a given belief set.

step at a time, updating its belief about the current time at each step, using rules like 1.

$$(1)\quad \begin{array}{ll} \texttt{i} & : \\ \texttt{i+1} & : \end{array} \frac{\texttt{Now(i)}}{\texttt{Now(i+1)}}$$

This step-wise, time-aware approach gives active logic fine control over what it does, and does not, derive and inherit at each step; note, for instance, that $Now(i)$ is not inherited at time step $i+1$.[2] This is accomplished by special inheritance rules like 2, shown below.

$$(2)\quad \begin{array}{ll} \text{i} & : \text{A} \\ \text{i+1} & : \text{A} \end{array}$$

[condition: $\neg$A $\notin$ KB at step i and A $\neq$ Now(i)]

Such step-wise control over inference gives active logic the ability to explicitly track the individual steps of a deduction. Thus, for instance, an inference rule can refer the results of all inferences *up until now*—i.e. thru step i—as it computes the subsequent results (for step $i+1$). This allows an active logic to reason, for example, about its own (past) reasoning; and in particular about what it has *not* yet concluded. Moreover, this can be performed quickly, since it involves little more than a lookup of the current knowledge base.

Active logic's step-wise control over inference, and its built-in ability to refer to individual steps of reasoning, make it a natural formalism for detecting and reasoning about contradictions and their causes. For as soon as a contradiction reveals itself—that is, as soon as P and $\neg P$ are both present in the KB—it is possible to "capture" it, preventing further reasoning using the contradictands as premises (and thereby preventing any explosion of wffs), while at the same time marking their presence, to allow further consideration of the cause of the contradiction. Current implementations of active logic incorporate a "conflict-recognition" inference rule like 3 for this purpose.

$$(3)\quad \begin{array}{ll} \texttt{i} & : \\ \texttt{i+1} & : \end{array} \frac{\texttt{P, } \neg\texttt{P}}{\texttt{contra(P, } \neg\texttt{P, i)}}$$

Through the use of such rules, *direct* contradictions can be recognized as soon as they occur, and further reasoning can be initiated to repair the contradiction, or at least to adopt a strategy with respect to it, such as simply avoiding the use of either of the contradictands for the time being. Unlike in truth maintenance systems [Doyle, 1979; 1980] where a separate process resolves contradictions using justification information, in an active logic the contradiction detection and handling [Miller, 1993] occur in the same reasoning process. Note that the `contra` predicate is a meta-predicate: it is about the course of reasoning itself (and yet is also part of that same evolving history).

Thus, speaking somewhat more broadly, active logic is a paraconsistent logic that *achieves* its paraconsistency in virtue of possessing two simultaneously active (and interactive) modes of reasoning, which might be called *circumspective* and *literal*. In literal mode, the reasoning agent is simply working with, and deriving the consequences of, its current beliefs. In circumspective mode, the reasoning agent is reasoning *about* its beliefs, noting, for instance, that it has derived a contradiction, and deciding what to do about that. It is important to active logic that these are not separate, isolated modes, but interactive and part of the same overall reasoning process.

3 A semantics for real-world reasoning

In this section we propose a semantics for a time-sensitive, contradiction-tolerant reasoning formalism, based heavily on the basic features of active logic detailed above.

3.1 Starting assumptions

In order to make the problem tractable for our first specification of the semantics, we will work under the following assumptions concerning the agent, the world, and their interactions:

- There is only one agent a.
- The agent starts its life at time $t = 0$ and runs indefinitely.
- The world is *deterministic* and *stationary* for $t \geq 0$. Thus, changes occur only in the beliefs of the agent a.

3.2 The language $\mathcal{L}$

In order to express theories about such an agent-and-world, we define a sorted first-order language $\mathcal{L}$. We define it in two parts: the language $\mathcal{L}_w$, a propositional language in which will be expressed facts about the world, and the language $\mathcal{L}_a$, a first order language used to express facts about the agent and about the agent's beliefs. We write Sn_{Lan} to mean the sentences of any language Lan.

Definition 1 Let $\mathcal{L}_w$ be a propositional language consisting of the following symbols:

- a set S of sentence symbols (propositional or sentential variables) $S = \{S_i : i \in N\}$ (N is the set of natural numbers).
- the propositional connectives $\neg$ and $\rightarrow$

Definition 2 Let $\mathcal{L}_a$ be a sorted first-order language that does not contain variables or quantifiers. It has three sorts $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$. Sort $\mathcal{S}_1$ is the sort of sentences in language $\mathcal{L}_w$, $\mathcal{S}_2$ is the sort of times, and $\mathcal{S}_3$ is the sort of sentences in language $\mathcal{L} = \mathcal{L}_a \cup \mathcal{L}_w$. $\mathcal{L}_a$ consists of the following symbols:

[2] To "inherit" P is, roughly speaking, to assert P at step $i+1$ just in case it was believed at step i. However, in a temporal, non-monotonic formalism what is justified *now* may not be justified *later*. Thus, although inheriting is a reasonable default behavior, there will be conditions and limits. Inheritance and disinheritance are directly related to belief revision [Gärdenfors, 1988] and to the frame problem [McCarthy and Hayes, 1969; Brown, 1987]; see [Nirkhe *et al.*, 1997] for further discussion.

- the propositional connective $\neg$
- a set of constant symbols $C = \{c_i : i \in N\}$, each c_i is of sort $\mathcal{S}_2$ (time)
- a set of constant symbols $D = \{d_{\sigma_i} : i \in N\}$, each d_{σ_i} is of sort $\mathcal{S}_1$ ($Sn_{\mathcal{L}_w}$)
- a set of constant symbols $E = \{e_{\theta_i} : i \in N\}$, each e_{θ_i} is of sort $\mathcal{S}_3$ ($Sn_{\mathcal{L}}$)
- the unary relation symbol **Now** of sort $\mathcal{S}_2$
- the ternary relation symbol **contra** of sort $(\mathcal{S}_1 \times \mathcal{S}_1 \times \mathcal{S}_2)$
- the binary relation symbol **bel** of sort $(\mathcal{S}_3 \times \mathcal{S}_2)$

In $\mathcal{L}_a$, *Now* is used to express the agent's time, the symbol *contra* is used to indicate the existence of a direct contradiction in its beliefs, and the symbol *bel* expresses the fact that the agent had a particular belief at a given time. These will be defined formally in definition 5. All the agent's knowledge is expressed in $Sn_{\mathcal{L}}$. The agent's knowledge may be incomplete, incorrect, or contradictory.

3.3 The semantics of $\mathcal{L}$

In the following several definitions, we define the semantics of the formalism given above, in the standard way.

Definition 3 An $\mathcal{L}_w$-truth assignment is a function $h : S \rightarrow \{T, F\}$ defined over the set S of sentence symbols in $\mathcal{L}_w$.

Definition 4 An $\mathcal{L}_w$ interpretation h (we keep the same notation) is a function $h : Sn_{\mathcal{L}_w} \rightarrow \{T, F\}$ over $Sn_{\mathcal{L}_w}$ that extends an $\mathcal{L}_w$-truth assignment h as follows:
$h(\neg\varphi) = T \iff h(\varphi) = F$
$h(\varphi \rightarrow \psi) = F \iff (h(\varphi) = T \text{ and } h(\psi) = F)$

We also stipulate a standard definition of consistency for $\mathcal{L}_w$: a set of $\mathcal{L}_w$ sentences is *consistent* iff there is some interpretation in which all the sentences are true. Notationally we write the usual $h \models \Sigma$, to mean all the sentences of Σ are assigned T by h.

4 A model of the agent's $\mathcal{L}_a$ beliefs

First of all it is important to note that, even in the case where the agent's beliefs are incomplete, incorrect, or inconsistent, there is always a complete and consistent theory *of* those beliefs at the meta level, and this theory can be expressed using the language $\mathcal{L}_a$. For instance, if the agent believes both S_i and $\neg S_i$, the two sentences "the agent believes that S_i" and "the agent believes that $\neg S_i$" can both be true at the same time.

Thus, we define the $\mathcal{L}_a$-structure at time t that models the theory about the agent's beliefs at the meta level, given KB_t^a (the agent's knowledge base at time t).

Definition 5 Let $H_t^r = (\mathcal{S}_1 = Sn_{\mathcal{L}_w}, \mathcal{S}_2 = N, \mathcal{S}_3 = Sn_{\mathcal{L}}, \{c_i\}_{i\in N}, \{d_{\sigma_i}\}_{i\in N}, \{e_{\theta_i}\}_{i\in N}, Now, contra, bel)$ where:

- the sort $\mathcal{S}_1$ is the sort of sentences in the language $\mathcal{L}_w$; the sort $\mathcal{S}_2$ is the sort of the times; and the sort $\mathcal{S}_3$ is the sort of sentences in the language $\mathcal{L}$
- $\forall i \in N, c_i$ names the time index i
- Since $\mathcal{L}_w$ is a countable language then $Sn_{\mathcal{L}_w}$ is countable, so can be enumerated as $Sn_{\mathcal{L}_w} = \{\sigma_i : i \in N\}$. The d-constants name every element in this set, that is, $\forall i \in N, d_{\sigma_i}$ names σ_i
- Since $\mathcal{L}$ is a countable language then $Sn_{\mathcal{L}}$ is countable, so it can be enumerated as $Sn_{\mathcal{L}} = \{\theta_i : i \in N\}$. The e-constants name every element in this set, that is, $\forall i \in N, e_{\theta_i}$ names θ_i
- The relation symbol Now has the following semantics: $H_t^r \models Now(c_s) \iff s = t$
  Now keeps track of the time, and indicates the current time of the agent's internal clock. *Now* is a logical symbol so at every time step it has the same interpretation in all structures.
- The relation symbol contra has the following semantics: $H_t^r \models contra(d_{\sigma_i}, d_{\sigma_j}, c_k)$ all $i, j, k \in N \iff k \leq t$, $\sigma_i; \sigma_j \in KB_k^a$; and either $\sigma_i = \neg\sigma_j$ or $\sigma_j = \neg\sigma_i$ for some σ_i and $\sigma_j \in Sn_{\mathcal{L}_w}$
  contra indicates that σ_i and σ_j are in direct contradiction, and that both were in the agent's KB at some time c_k where k is less than or equal to t. For simplicity of expression we will typically write $contra(d_\alpha, d_{\neg\alpha}, c_t)$.
- The relation symbol bel has the following semantics: $H_t^r \models bel(e_{\theta_i}, c_k) \iff k \leq t$ and $\theta_i \in KB_k^a$

bel has the rough meaning "believes that", and simply states that a given sentence from $\mathcal{L}$ was in the agent's KB at a time c_k where k is less that or equal to t. We will typically write $bel(e_\alpha, c_t)$.

Finally we define an $\mathcal{L}$-structure that models the theory of the agent-and-world.

Definition 6 An active structure at time t, shortly a-structure, is an $\mathcal{L}$-structure defined as follows: $M_t =< h_t, H_t^r >$

5 A model of the agent's $\mathcal{L}_w$ beliefs

Now we turn to the challenging problem of how to model, at the object level, the agent's beliefs about the world, given that these beliefs are not just evolving from moment to moment, but that at any given time, they may be inconsistent.

First, we define two notions of *temporal consistency* relative to the language $\mathcal{L}$. We will not mention the language when it is clear from the context.

Definition 7 A set of sentences $\Sigma \subseteq Sn_{\mathcal{L}}$ is said to be temporally-strongly consistent at time t (t-strongly consistent) if and only if $\exists(M_t)[M_t \models \Sigma]$.

Definition 8 A set of sentences $\Sigma \subseteq Sn_{\mathcal{L}}$ is said to be temporally-weakly consistent at time t (t-weakly consistent) if and only if $\exists(M_t)[M_t \models (\Sigma \bigcap Sn_{\mathcal{L}_a})]$.

From now on, we will assume that the agent's KB is t-weakly consistent, and therefore that the agent's knowledge at the meta-level is t-strongly consistent.

Definition 9 Let $\Sigma_t^\omega = \{\Sigma \in \mathcal{P}(Sn_\mathcal{L}) : \Sigma$ is t-weakly consistent and Σ is finite$\}$

Next we define a new propositional language $\mathcal{L}'_w$ to express the awareness of the agent of its knowledge about the world (the notion of awareness will become clearer as we proceed with our discussion).

Definition 10 The propositional language $\mathcal{L}'_w$ derived from $\mathcal{L}_w$ consists of the following symbols:

- a set of sentence symbols $S' = \{S_i^j : j \in N$ and $S_i \in \mathcal{L}_w\}$. Thus for every sentence symbol in $\mathcal{L}_w$, there is a corresponding infinite pool of sentence symbols in $\mathcal{L}'_w$.
- the propositional connectives $\neg, \rightarrow$

Definition 11 Let $\mathcal{L}' = \mathcal{L}'_w \bigcup \mathcal{L}_a$. Let $Sn_{\mathcal{L}'} = Sn_{\mathcal{L}'_w} \bigcup Sn_{\mathcal{L}_a}$. The language $\mathcal{L}'$ is used to express the agent's awareness of its agent-and-world knowledge.

Note that definitions 7 and 8 can be extended to $\mathcal{L}'$

We next define a *perception (awareness) function* for an agent. The notion of a perception function is intended to help capture, at least roughly, how the world might seem to an agent with a given belief set KB. For a real agent, only some logical consequences of its KB are believed at any given time, since it cannot manage to infer all the infinitely many consequences in a finite time, let alone in the present moment. Moreover, even if the KB has contradictory beliefs, the agent still has a view of the world, and there will be limits on what the agent will and won't infer. This is in sharp distinction to the classical notion of a model, where (i) inconsistent beliefs are ruled out of bounds, since then there are no models, and (ii) all logical consequences of KB are true in all models.

The task we are addressing, then, is that of finding a notion of model based somehow on semantic-like concepts, yet that avoids both (i) and (ii) above. Our idea—via perception functions—is to suppose that an agent's limited resources apply also to its ability to inspect its own KB. Thus, if S_i and $\neg S_i$ are both in the KB, the agent might not realize, at first, that the two letters in question are the same. Thus it might seem to the agent that the world is one in which, say, S_i^1 is true, and so is $\neg S_i^2$. Only later might the agent realize the two letters are one and the same.

This allows the agent to have inconsistent beliefs while still having a consistent world model. Later, when S_i^1 and S_i^2 are unified into S_i—i.e., when the agent realizes there is a conflict—it will take some remedial action such as doubting one or both beliefs. Moreover, it allows us to see how an agent with inconsistent beliefs could avoid vacuously concluding *any* wff, and also reason in a directed way, by applying inference rules only to an appropriately perceived sub-set of its beliefs. We hope that this approach can shed some light on focused, step-wise, resource-bounded reasoning more generally.

In our definition we start with a t-weakly consistent set Σ. The perception function can make changes only to $\Sigma \bigcap Sn_{\mathcal{L}_w}$, which we call Γ. A perception function does not change $\Sigma - \Gamma$. We assume that the elements of Γ are ordered alphabetically. We use the same notation *per* when the perception function is applied to a sentence symbol, a sentence, or a set of sentences.

Definition 12 A perception (awareness) function at time t is a mapping: $per_t : \Sigma_t^\omega \rightarrow \mathcal{P}(Sn_{\mathcal{L}'})$ defined by a finite sequence of positive integers $< i_1, \ldots, i_n >$ with the following effect:

1. Let S_k be the k^{th} sentence symbol in Γ. For $k \leq n$, $per_t(S_k) = S_k^{i_k}$. Then for $\phi \in \Gamma$, $per_t(\phi) = \phi'$ where ϕ' is obtained by applying per_t to all the applicable symbols in ϕ.
2. Let the set of contradictory pairs be defined as follows: $CP = \{< \phi, \psi > | per_t(\phi) = \neg per_t(\psi)$ *or* $per_t(\psi) = \neg per_t(\phi)\}$.
3. $per_t(\Sigma) = (\Sigma - \Gamma) \bigcup \{per_t(\phi) | \phi \in \Gamma\} \bigcup \{contra(d_\phi, d_\psi, c_t) | < \phi, \psi > \in CP\} - \{per_t(\phi) | < \phi, \psi > \in CP\} - \{per_t(\psi) | < \phi, \psi > \in CP\}$

Note that in the above definition of the perception function we assumed that the agent can become aware of the *direct contradictions* in its knowledge base. And since we assume instantaneous awareness then this means that the agent can capture these types of contradictions immediately after they appear in its knowledge base.

Definition 13 Let PER_t be the class of all perception functions at time t.

Note that there are infinitely many ways of assigning superscripts to sentences, yielding infinitely many perception functions at any given time.

Theorem 1 If KB_t^a is t-weakly consistent, then there is some $per_t \in PER_t$ such that $per_t(KB_t^a)$ is t-strongly consistent (in $\mathcal{L}'$).

Proof Assume that KB_t^a is finite. Consider the subset $KB_t^a \bigcap Sn_{\mathcal{L}_w} = \Gamma$, and the ordered set of all sentence symbol tokens from S_1 to S_n in Γ. Apply to this set of sentence symbols the perception function $< 1, 2, 3 \ldots, n >$ according to the procedure described in definition 12, to obtain Γ'. That is, replace S_1 with S_1^1, S_2 with S_2^2, etc. The result makes every sentence symbol appearing in Γ' unique. Now, if an even number of the negation symbol is applied to a sentence symbol x in a formula φ, we say x occurs positively in φ, otherwise we say x occurs negatively in φ. Consider the following truth assignment in Γ': for each symbol that occurs positively, assign a true value, otherwise assign a false value. By this assignment every $\varphi \in \Sigma'$ would be true and hence Γ' is consistent. Since the remaining sentences in KB_t^a are consistent by assumption, $per_t(KB_t^a)$ is t-strongly consistent.

Definition 14 Let $per_t \in PER_t$ be a perception function at time t. We define $KB_{per_t}^a = per_t(KB_t^a)$ as the agent's perception of its knowledge base at time t. We

also define $W^a_{per_t} = (Sn_{\mathcal{L}'_w} \cap KB^a_{per_t})$ as the agent's perception of the part of its own knowledge base concerning the external world.

From the above definition and our prior assumptions, $W^a_{per_t}$ is t-strongly consistent, so there exists a set of $\mathcal{L}'_w$ interpretations G_{per_t} such that $h_{per_t} \models W^a_{per_t}$ for every $h_{per_t} \in G_{per_t}$.

Definition 15 Let $per_t \in PER_t$ be a perception function at time t. Define G_{per_t} to be the class of $\mathcal{L}'_w$interpretations determined by $W^a_{per_t}$.

Now we define an $\mathcal{L}'$-structure that models the agent's KB after a perception function has been applied to it. This is meant to capture the way the world might seem to the agent at a given time.

Definition 16 Let $per_t \in PER_t$ be a perception function at time t. Then a perceived temporal structure at time t, shortly pt-structure, is an $\mathcal{L}'$-structure defined as follows: $M_{per_t} = \langle h_{per_t}, H^r_t \rangle$ for some $h_{per_t} \in G_{per_t}$. We use $\mathbb{M}_{per_t}$ for the set of all M_{per_t}s.

6 Active consequence

At this point we are ready to define the notion of *active consequence* at time t—the active logic equivalent of logical consequence.

Definition 17 Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$ such that $\Sigma = KB^a_t$. Then Θ is said to be a 1-step active consequence of Σ at time t, written $\Sigma \models_1 \Theta$, if and only if the following holds:

$(\exists per_t \in PER_t)(\exists per_{t+1} \in PER_{t+1})(\forall M_{per_t} \in \mathbb{M}_{per_t})[H^r_{t+1} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a}) \& M_{per_t} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w})]$

Roughly speaking, *if* for the $\mathcal{L}_w$ sentences, the set of conclusions—as perceived by the agent at time $t+1$ according to the restrictions set out in definition 12—are yielded by the antecedents as perceived by the agent at time t according to those same restrictions, *and if* for the $\mathcal{L}_a$ sentences, the $\mathcal{L}_a$ structure H^r_{t+1} models—according to definition 5—the agent's perception of the conclusions at time $t+1$, *then* it can be said that the conclusions are one-step active consequents of the antecedents.

Examples:

1 Let $\Sigma = \{\varphi, \neg\varphi\} = KB^a_t$, let $\Theta = \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$. Let $per_t \in PER_t$, such that $per_t(\Sigma) = \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$, and $per_{t+1} \in PER_{t+1}$ such that $per_{t+1}(\Theta) = \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$. We have $H^r_{t+1} \models \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$, hence $\Sigma \models_1 \Theta$.

This case is very straightforward. Since the consequents contain only $\mathcal{L}_a$ sentences, we only need to determine if the agent's perception of the consequent, namely $contra(d_\varphi, d_{\neg\varphi}, c_t)$, is modeled by H^r_{t+1}, which clearly by definition 5 it is, since φ and $\neg\varphi$ are contradictory, and both $\in KB$ at t.

2 Let $\Sigma = \{Now(t), S_1, S_1 \rightarrow S_4, S_{12}\} = KB^a_t$. Let $\Theta = \{Now(t+1), S_4, S_{12}\}$. Let $per_t \in PER_t$ such that $per_t(\Sigma) = \{Now(t), S^a_1, S^a_1 \rightarrow S^b_4, S^c_{12}\}$ for some $a, b, c \in N$. Then $\forall(h_{per_t}) \in G_{per_t}$, $h_{per_t}(S^a_1) = h_{per_t}(S^b_4) = h_{per_t}(S^c_{12}) = T$. Let $per_{t+1} \in PER_{t+1}$ where $per_{t+1}(\Theta) = \{Now(t+1), S^b_4, S^c_{12}\}$. Clearly, $H^r_{t+1} \models per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a} = \{Now(t+1)\}$ and $h_{per_t} \models per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w} = \{S^b_4, S^c_{12}\}$ for every h_{per_t}. Hence $\Sigma \models_1 \Theta$.

This is also relatively straightforward, since, once the perception function has been applied, determining whether the $\mathcal{L}_w$ sentences in Θ are active consequents of Σ is similar to determining this classically; and for the $\mathcal{L}_a$ sentences it is just a matter of being sure that H^r_{t+1} models those sentences according to definition 5.

3 Let Σ, Θ be as in the previous example with $bel(e_{S_5}, c_t)$ added to Θ. Since $S_5 \notin \Sigma$, then $H^r_{t+1} \not\models bel(e_{S_5}, c_t)$ (see definition 5). So $\Sigma \not\models_1 \Theta$.

4 Let $\Sigma = \{Now(t)\} = KB^a_t$. Let $\Theta = \{Now(t+5)\}$ The perception function can only give these same sentences: $per_t(\Sigma) = \{Now(t)\}$ and $per_{t+1}(\Theta) = \{Now(t+5)\}$. But $H^r_{t+1} \not\models \{Now(t+5)\}$ at time $t+1$ (see definition 5), and so $\Sigma \not\models_1 \Theta$.

A more general notion of active consequence, called an n-step active consequence, is defined recursively from 1-step active consequence.

Definition 18 Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$ such that $\Sigma = KB^a_t$. Then Θ is said to be an n-step active consequence of Σ at time t, written $\Sigma \models_n \Theta$, if and only if the following holds:

$\exists \Gamma \subseteq Sn_{\mathcal{L}}: \Sigma \models_{n-1} \Gamma$ and $\Gamma \models_1 \Theta$

Finally, we define active consequence, written $\models_a$, in terms of n-step active consequence.

Definition 19 Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$ such that $\Sigma = KB^a_t$. Θ is said to be an active consequence of Σ, that is, $\Sigma \models_a \Theta$, iff $\Sigma \models_n \Theta$ for some $n \in N$.

Example:

1. Let $\Sigma = KB^a_t = \{S_1, S_2, S_2 \rightarrow \neg S_1\}$ and $\Theta = \{contra(d_{S_1}, d_{\neg S_1}, t+1)\}$. Then $\Sigma \models_a \Theta$ as follows.

Let $\Gamma = \{S_1, \neg S_1\}$. Let $per_t \in PER_t$ such that $per_t(\Sigma) = \{S^a_1, S^b_2, S^b_2 \rightarrow \neg S^c_1\}$ for some $a, b, c \in N$. Then for every $h_{per_t} \in G_{per_t}$ the following must hold: $h_{per_t}(S^a_1) = h_{per_t}(S^b_2) = T, h_{per_t}(S^c_1) = F$.

Let $per_{t+1} \in PER_{t+1}$ where $per_{t+1}(\Gamma) = \{S^a_1, \neg S^c_1\}$. Clearly, $\forall(h_{per_t}) \in G_{per_t}, h_{per_t} \models per_{t+1}(\Gamma)$, hence $\Sigma \models_1 \Gamma$. Notice that Γ is potentially part of KB^a_{t+1}, and that once in the KB, the superscripts a and c are dropped, so that the sentences would appear in KB^a_{t+1} as $\{S_1, \neg S_1\}$.

Next, let $per_{t+2} \in PER_{t+2}$ where $per_{t+2}(\Theta) = contra(d_{S_1}, d_{\neg S_1}, t+1)$. Since $S_1, \neg S_1 \in \Gamma$ (potentially part of KB^a_{t+1}), then $H^r_{t+2} \models per_{t+2}(\Theta)$, hence $\Gamma \models_1 \Theta$.

So we have $\Sigma \models_1 \Gamma$ and $\Gamma \models_1 \Theta$; this proves $\Sigma \models_2 \Theta$, and thus $\Sigma \models_a \Theta$.

The point of this example is that, in active logic, it can take time for particular sentences to appear in the

KB. So, for instance, because the contradiction in Σ is indirect, it will not become a direct contradiction until $t+1$—that is, time $t+1$ is the first time that both S_1 and $\neg S_1$ are actually in the KB. This is important because one of the conditions of $contra(d_{\sigma_i}, d_{\sigma_j}, c_s)$ is that σ_i and σ_j are in the KB at time s, and this does not happen in our example until $t+1$. It is at this point that the contradiction can be recognized, and $\{contra(d_{S_1}, d_{\neg S_1}, c_{t+1})\}$ can be asserted.

Note that this approach to logical consequence allows one to define possible valid paths of reasoning, and, in the case of 1-step active consequence, the shortest possible valid path. However, a given agent may or may not, in practice, take the shortest possible valid path to reach a given conclusion. Any given agent, reasoning validly, may still reason more or less efficiently, or more or less directly to a particular conclusion, depending on the way it perceives its KB, and on the inference rules it in fact employs.

6.1 The relation between logical consequence and active consequence

The following theorem gives a key result regarding the relationship between classical propositional logical consequence and active consequence (restricted to sentences in $\mathcal{L}_w$). It says that for a consistent $KB = \Sigma$, Θ is a classical logical consequence of Σ, iff it is an active consequence. Intuitively this should make sense. For consider that every given set of consistent sentences has a certain definite set of conclusions—call this the "inferential power" of the set. We would expect this same set in active logic to have at least as much, but not more, inferential power as it has under classical logic. "At least as much" because one possible perception function, by assigning the same number to each sentence in Σ essentially leaves the set of sentences, and therefore its inferential power, unchanged. "Not more than" because there is no perception function that *increases* the inferential power of a given set—a perception function either leaves the inferential power the same, or reduces the number of things that can be inferred.

Theorem 2 Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$. If Σ is consistent, then the following holds:

$$\Sigma \models \Theta \Longleftrightarrow \Sigma \models_a \Theta$$

Proof Let $A = \{S_{i_1}, \cdots, S_{i_n}\}$ be the set of all sentence symbols appearing in Σ, and let $B = \{S_{j_1}, \cdots, S_{j_m}\}$ be the set of all sentence symbols appearing in Θ.

$\Rightarrow$ If $\Sigma \models \Theta$, then $\forall h\colon h \models \Sigma \Rightarrow h \models \Theta$. Consider $per_t \in PER_t$ where all instances of every sentence symbol S_k in its input gets mapped to S_k^1. Then the following holds: $(\forall h \models \Sigma)(\forall h')\colon h' \models per_t(\Sigma) \iff (\forall S_k \in A\colon h'(S_k^1) = h(S_k))$, where h is an $\mathcal{L}_w$-truth assignment and h' is an $\mathcal{L}'_w$-truth assignment. Consider $per_{t+n} = per_t$, then the previous sentence holds when replacing Σ by Θ, A by B, and per_t by per_{t+n}. Hence $\Sigma \models_a \Theta$.

$\Leftarrow$ If $\Sigma \not\models \Theta$, then there exists an $\mathcal{L}_w$-truth assignment h such that $h \models \Sigma$ and $h \not\models \Theta$. Then $h \not\models \theta$ for some $\theta \in \Theta$. Let per_t be some perception function at time t; let $\Sigma' = per_t(\Sigma)$. Let per_{t+n} be some perception function at time $t+n$; let $\Theta' = per_{t+n}(\Theta)$; let $\theta' \in \Theta'$ be the mapping of θ under this function. Consider the following $\mathcal{L}'_w$-truth assignment h' where $h'(S_{i_1}^{j_1}) = h(S_{i_1}), \cdots, h'(S_{i_n}^{j_n}) = h(S_{i_n})$ for all $j_1, \cdots, j_n \in N$. Clearly, $h' \models \Sigma'$ and $h' \not\models \theta'$, so $\Sigma \not\models_a \Theta$.

Note that the above theorem *doesn't* hold for all $\mathcal{L}_a$ sentences—that is, a given set of $\mathcal{L}_w$ sentences Σ might, in active logic, yield some $\mathcal{L}_a$ sentences that would not be yielded by classical logic.[3] Consider for instance $\Sigma = \{S_1\}$ at t and $\Theta = \{bel(e_{S_1}, c_t)\}$ at $t+1$. The sentence $bel(e_{S_1}, c_t)$ is an active consequence of S_1 (at t), but is not a classical logical consequence. So the inferential power of a given set *is* increased in active logic, but only in its yield of $\mathcal{L}_a$ sentences.

Of course, it is precisely the fact that active logic permits the inference of certain additional $\mathcal{L}_a$ sentences that allows it to *reduce* the inferential power of *inconsistent* sets. This is crucial because (as previously noted) in classical logic the inferential power of an inconsistent set is indefinitely large. For active logic, however, there are only two possibilities for inconsistent sets: either (1) the set is made consistent by a perception function that assigns different superscript numbers to the relevant sentences, in which case nothing will follow from the contradiction (as there will be no contradiction), or (2) the contradiction will be recognized and $contra(d_\varphi, d_{\neg\varphi}, c_t)$ (and only this) will follow from the contradiction (see definition 26 and theorem 8, below).

This brings us to our notion of sound inference, which we define in terms of n-step active consequence.

Definition 20 An active sound (a-sound) inference is one in which the consequent is an active consequence of the antecedent.

7 Sound and unsound inferences in active logic

At this point we are in a position to define some inference rules, beginning with the rules most central to active logic.

7.1 Some active-sound inference rules

First we define the timing inference rule.

Definition 21 If $now(t) \in KB_t^a$ (remember KB_t^a is t-weakly consistent), then the timing inference rule is defined as follows:

$$\frac{t : Now(c_t)}{t+1 : Now(c_{t+1})}$$

Theorem 3 The timing inference rule is a-sound.

[3] This is in addition to the obvious fact that by classical inference $\{Now(1)\}$ would not yield $\{Now(2)\}$.

Proof We need to show that $\{Now(c_t)\} \models_1 \{Now(c_{t+1})\}$ at time t. This holds by definition 12 of the perception functions: $\forall per_{t+1} \in PER_{t+1}: per_{t+1}(\{Now(c_{t+1})\}) = \{Now(c_{t+1})\}$ and by definition 5 of $\mathcal{L}$-structures: $H^r_{t+1} \models \{Now(c_{t+1})\}$.

We also define the direct contradiction rule.

Definition 22 If $\varphi, \neg\varphi \in KB^a_t$, where $\varphi \in Sn_{\mathcal{L}_w}$ and $\neg\varphi \in Sn_{\mathcal{L}_w}$, then the direct contradiction inference rule is defined as follows:

$$\frac{t : \varphi, \neg\varphi}{t+1 : contra(d_\varphi, d_{\neg\varphi}, c_t)}$$

Theorem 4 The direct contradiction inference rule is a-sound.

Proof We need to show that $\{\varphi, \neg\varphi\} \models_1 \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$ at time t. This holds by definition 12 of the perception functions: $\forall per_{t+1} \in PER_{t+1}: per_{t+1}(\{contra(d_\varphi, d_{\neg\varphi}, c_t)\}) = \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$ and by definition 5 of $\mathcal{L}$-structures: $H^r_{t+1} \models \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$ (since $\varphi, \neg\varphi$ are in the antecedents and hence in KB^a_t).

We define the introspection inference rule as follows.

Definition 23 If $\varphi \in KB^a_t$, where $\varphi \in Sn_{\mathcal{L}}$, then the introspection inference rule is defined as follows:

$$\frac{t : \varphi}{t+1 : bel(e_\varphi, c_t)}$$

Theorem 5 The introspection rule is a-sound.

Proof We need to show $\varphi \models_1 bel(e_\varphi, c_t)$ at time t. This holds by definition 12 of the perception functions: $\forall per_{t+1} \in PER_{t+1}: per_{t+1}(\{bel(e_\varphi, c_t)\}) = \{bel(e_\varphi, c_t)\}$ and by definition 5 of $\mathcal{L}$-structures: $H^r_{t+1} \models \{bel(e_\varphi, c_t)\}$ (since φ is in the antecedents and hence in KB^a_t).

We define the negative introspection inference rule as follows.

Definition 24 If $\varphi \notin KB^a_t$, where $\varphi \in Sn_{\mathcal{L}}$ for some $i \in N$, then the negative introspection inference rule is defined as follows:

$$\frac{t : KB^a_t}{t+1 : \neg bel(e_\varphi, c_t)}$$

Theorem 6 The negative introspection rule is a-sound.

Proof We need to show $KB^a_t \models_1 \neg bel(e_\varphi, c_t)$ at time t. This holds by definition 12 of the perception functions: $\forall per_{t+1} \in PER_{t+1}$: $per_{t+1}(\{\neg bel(e_\varphi, c_t)\}) = \{\neg bel(e_\varphi, c_t)\}$ and by definition 5 of $\mathcal{L}$-structures: $H^r_{t+1} \models \{\neg bel(e_\varphi, c_t)\}$ (since $\varphi \notin KB^a_t$).

We can define the equivalent of the modus ponens inference rule—active modus ponens, or AMP—as follows.

Definition 25 If $\varphi, \varphi \to \psi \in KB^a_t$, then the AMP inference rule is defined as follows:

$$\frac{t : \varphi, \varphi \to \psi}{t+1 : \psi}$$

Theorem 7 The AMP inference rule is a-sound.

Proof We need to show $\{\varphi, \varphi \to \psi\} \models_1 \{\psi\}$ at time t. Let $per_t \in PER_t$, following the same procedure as described in the proof of theorem 1, except that both instances of φ are mapped to the same sentence φ^1. So we have $per_t(\{\varphi, \varphi \to \psi\}) = \{\varphi^1, \varphi^1 \to \psi^1\}$. From the proof of theorem 1, we can see that this latter set is consistent and any interpretation must satisfy the following: $\forall(h_{per_t}) \in G_{per_t}: h_{per_t}(\varphi^1) = h_{per_t}(\psi^1) = T$. Let $per_{t+1} \in PER_{t+1}$ such that $per_{t+1}(\{\psi\}) = \{\psi^1\}$, then $\forall(h_{per_t}) \in G_{per_t}: h_{per_t} \models \{\psi^1\}$.

7.2 Active-unsound inference rules

We have examined a number of instances of classically unsound inference rules, and get the expected intuitive results that these inferences are also active-unsound. However, one rule that is classically sound, but active-unsound, is the explosive rule. This shows that active logic is a paraconsistent logic, something we consider one of its advantages over classical formalisms.

Definition 26 We'll call the rule where (A and $\neg A$) implies B the explosive rule:

$$\frac{t : \varphi, \neg\varphi}{t+1 : \psi}$$

Theorem 8 The explosive inference rule is a-unsound.

Proof There are two general cases to consider, one in which the perception function treats the sentences as *different*, i.e. assigns them different subscripts, and another in which the perception function at time t treats the sentences in the antecedent as the *same*, i.e. as contradictory, giving $contra(d_\varphi, d_{\neg\varphi}, c_t)$.

1. For three numbers $i, j, k \in N$, such that $i \neq j$, and k may equal either i or j or neither,[4] the perception function gives $per_t(\{\varphi, \neg\varphi\}) = \{\varphi^i, \neg\varphi^j\}$, and $per_{t+1}(\{\psi\}) = \{\psi^k\}$. However, for every one of the possible numeric assignments to i, j, k, (that is, for every $per_t \in PER_t$ except that discussed in clause (2)) there is at least one interpretation h_{per_t} such that $\{\varphi^i, \neg\varphi^j\} \not\models \{\psi^k\}$, namely the one which assigns both φ^i and $\neg\varphi^j$ to T and ψ^k to F.
2. For the only remaining $per_t \in PER_t$, that gives $per_t(\{\varphi, \neg\varphi\}) = \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$ and $per_{t+1}(\{\psi\}) = \{\psi^k\}$, there is also at least one interpretation h_{per_t} such that $\{contra(d_\varphi, d_{\neg\varphi}, c_t)\} \not\models \{\psi^k\}$, that assigns $contra(d_\varphi, d_{\neg\varphi}, c_t)$ to T and ψ^k to F.

[4] If $i = j$ the perception function must produce $contra(d_{\varphi^i}, d_{\neg\varphi^j}, d_t)$, in which case see clause(2).

8 Conclusion and Future Work

In this paper we have outlined a semantics for a time-sensitive, contradiction-tolerant logical reasoning formalism designed for on-board use by real-world agents. Central to the semantics is the notion of a perception function, inspired by the idea that, until an agent *notices* that a set of beliefs is inconsistent, that set *seems* consistent—and that when a contradiction *is* noticed, that fact can be explicitly registered by the agent, and further reasoning with the contradictory beliefs can be curtailed.

To keep this initial presentation relatively simple, we made a number of assumptions that in future work we will discard. The most important of these assumptions is that the world is stationary, and thus all facts about the world are timelessly true. It should be noted that there is no problem in principle with applying active logic to the case of reasoning about a changing world—after all, the facts that beliefs are held at times, that the KB changes over time, and that inference is itself a temporal phenomenon, are all already explicitly modeled by the formalism. To handle a changing world, we would also have to model the additional facts that beliefs can be held not just *at* times, but *about* facts-at-times, and even about the *durations* of facts—e.g. that it rained yesterday, or that it rained yesterday for 1 hour between noon and one. Such modification is straightforward. There are some tricky aspects to modeling proper *reasoning* with temporally relative beliefs in a changing world, but for this we can avail ourselves of the extensive literature on default reasoning and non-monotonic temporal logics.

Future work will also consider multiple agents, reasoning both about the world and about one another's beliefs, and extending the semantics to include predicates.

Finally, we acknowledge that fuller understanding of this work will require comparison and contrast with related efforts, e.g., [Gabbay, 1999; Ismail and Shapiro, 2000; Lespérance and Levesque, 1995; Martins and Shapiro, 1988]. We regret that space contraints made such discussion impossible in this case.

9 Acknowledgments

This work is supported in part by a grant from AFOSR.

References

[Brown, 1987] F. Brown, editor. *The Frame Problem in Artificial Intelligence*. Morgan Kaufmann, 1987.

[Doyle, 1979] Jon Doyle. A Truth Maintenance System. *Artificial Intelligence*, 12(3):231–272, 1979.

[Doyle, 1980] Jon Doyle. *A Model for Deliberation, Action, and Introspection*. PhD thesis, Massachusetts Institute of Technology, 1980.

[Elgot-Drapkin and Perlis, 1990] J. Elgot-Drapkin and D. Perlis. Reasoning situated in time I: Basic concepts. *Journal of Experimental and Theoretical Artificial Intelligence*, 2(1):75–98, 1990.

[Gabbay, 1999] Dov Gabbay. Action, time and default. In H. Levesque and F. Pirri, editors, *Logical Foundations for Cognitive Agents*, pages 151–154. Springer-Verlag, 1999.

[Gärdenfors, 1988] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, MA, 1988.

[Ismail and Shapiro, 2000] Haythem O. Ismail and Stuart C. Shapiro. Two problems with reasoning and acting in time. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference*, 2000.

[Lespérance and Levesque, 1995] Yves Lespérance and Hector J. Levesque. Indexical knowledge and robot action: A logical account. *Artificial Intelligence*, 73(1-2):69–115, 1995.

[Levesque, 1984] H. Levesque. A logic of implicit and explicit belief. In *Proceedings of the National Conference on Artificial Intelligence*, pages 198–202, Austin, TX, 1984. American Association for Artificial Intelligence.

[Martins and Shapiro, 1988] J. P. Martins and S. C. Shapiro. A model for belief revision. *Artificial Intelligence*, 35(1):25–79, 1988.

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

[Miller and Perlis, 1993] M. Miller and D. Perlis. Presentations and this and that: logic in action. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, Boulder, Colorado, 1993.

[Miller, 1993] M. Miller. *A View of One's Past and Other Aspects of Reasoned Change in Belief*. PhD thesis, Department of Computer Science, University of Maryland, College Park, Maryland, 1993.

[Nirkhe *et al.*, 1997] M. Nirkhe, S. Kraus, M. Miller, and D. Perlis. How to (plan to) meet a deadline between *now* and *then*. *Journal of logic computation*, 7(1):109–156, 1997.

[Priest and Tanaka, Summer 2004] Graham Priest and Koji Tanaka. Paraconsistent logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2004.

[Priest *et al.*, 1989] G. Priest, R. Routley, and J. Norman. *Paraconsistent Logic: Essays on the Inconsistent*. Philosophia Verlag, Mnchen, 1989.

[Priest, 2002] G Priest. Paraconsistent logic. In D. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic, 2ed*, pages 287–393. Kluwer Academic Publishers, 2002.

[Purang, 2001] K. Purang. *Systems that detect and repair their own mistakes*. PhD thesis, Department of Computer Science, University of Maryland, College Park, Maryland, 2001.