On a Semantics for Active Logic

Johan Hovold

Examensarbete för 20 p, Institutionen för Datavetenskap,
Naturvetenskapliga fakulteten, Lunds universitet

Thesis for a diploma in computer science, 20 credit points,
Department of Computer Science, Faculty of Science, Lund University

Om en semantik för aktiv logik

Sammanfattning

Detta arbete behandlar modellering av realistiska agenters resonerande. Speciellt analyseras, kritiseras och utökas en semantik för aktiv logik som är en logisk formalism avsedd att kunna hantera tidens gång såväl som inkonsistens. Baserat på vad som kallas perceptionsfunktioner definieras uppfattade temporala strukturer som möjliggör för inkonsistenta kunskapsbaser att ha modeller. Genom dessa strukturer konstrueras en konsekvensrelation kallad aktiv konsekvens. Aktiv konsekvens troddes tidigare sammanfalla med klassisk logisk konsekvens när den begränsades till en bestämd delmängd av språket och konsistenta premisser. Vi visar att denna identitet inte håller på grund av problemet med Σ-oavgörbarhet – att det finns satser för vilka man inte kan avgöra huruvida de följer aktivt från en given mängd Σ – och föreslår en förfinad definition av aktiv konsekvens som lösning. Vårt viktigaste resultat är emellertid att vi visar att aktiv konsekvens är explosiv, det vill säga att vad som helst följer aktivt från en direkt motsägelse. Därför, och i motsats till vad som tidigare har hävdats, är en logik baserad på denna konsekvensrelation inte parakonsistent.

On a Semantics for Active Logic

Abstract

The object of study for this thesis is the problem of modelling the reasoning of real-world agents. In particular, a semantics for active logic, which is a logical formalism conceived to be able to cope with the passing of time as well as inconsistency, is analysed, criticised and refined. Based on what is called perception functions, a notion of perceived temporal structure is defined, which allows inconsistent knowledge bases to have models. Using such structures, a consequence relation called active consequence is constructed. Active consequence was previously believed to coincide with classical logical consequence when restricted to a certain subset of the language and consistent premises. We show that this identity does not hold due to the problem of Σ-undeterminism – that there are sentences for which it cannot be determined whether they follow actively from a given set Σ – and suggest a refined definition of active consequence as a solution. Our main result, however, is that we show that active consequence is explosive, that is, that anything follows actively from a direct contradiction. Consequently, and contrary to what has been previously claimed, a logic based on this consequence relation is not paraconsistent.

Contents

Acknowledgements

I would like to thank my supervisor Jacek Malec for introducing me to the subject of resource-bounded reasoning and for encouraging me during my research. I am also very grateful for his support in my other projects.

Thanks also to Michael L. Anderson, Walid Gomaa, John Grant, and Don Perlis for sending me a manuscript of their forthcoming article on the semantics in response to my comments on previously published work.

Finally, I would like to thank Mikael Asker and Pontus Melke for reading and commenting drafts of this thesis.

1 Introduction

The object of study for this thesis is the problem of modelling the reasoning of real-world agents. Real agents have some important characteristics that need to be taken into account when trying to formalise their reasoning. First, their reasoning takes place in time, and time passes as their reasoning proceeds. Secondly, real-world agents are fallible and must be able to reconsider previous beliefs and handle contradictions, which will inevitably arise from time to another. Classical logical formalisms are thus not, at least at first sight, particularly well-suited for modelling the ever-changing set of beliefs of such an agent. In the classical paradigm time is not the primary concern, and classical concepts such as logical consequence that lets us conclude that, for instance, ψ follows from φ gives us no information about how long it will take (for a specific agent) to deduce ψ given φ. For our purposes, knowing that something follows given *enough time* to deliberate is not sufficient. Furthermore, real-world agents are necessarily resource bounded, and at every point in time, only a finite subset of the consequences of an agent's beliefs are known to the agent. This set of *explicit beliefs* is contrasted with the the agent's *implicit beliefs*, which is the set of conclusions that follow logically from the explicit beliefs (and which may eventually become explicit given enough time to deliberate) [Lev84]. Classical logical formalisms do not make this distinction and only deal with what we have called implicit beliefs. These formalisms can thus only be used to model idealised agents that always believe every consequence of their current beliefs (i.e. their implicit and explicit beliefs coincide).

Inconsistency is also a problem with classical logic since anything follows from contradictory premises, *ex contradictione quodlibet*. There are, however, formalisms that challenge this logical principle. Let $\models$ be a consequence relation that is either proof theoretically or semantically defined. Then $\models$ is said to be *explosive* if and only if for every formula φ and ψ, $\{\varphi, \neg\varphi\} \models \psi$. Classical logic, intuitionistic logic, and most other standard logics are explosive. A logic with a consequence relation that is not explosive is said to be *paraconsistent* [PT04]. An agent reasoning in the real world would obviously benefit from using a logic that is paraconsistent – the agent should try to resolve inconsistencies rather than happily infer anything from them.

Active logic is a formalism conceived to be able to cope with the passing of time as well as inconsistency [EDP90, GKP00, AGGP05a]. Its first incarnation, step logic, was developed as a tool that could be used to reason about approaching deadlines. Active logic is a logical formalism but also a reasoning mechanism for an agent[1]. In particular, it is (primarily) an object-level formalism rather than a metalevel formalism *about* reasoning

[1]For a discussion of this widened notion of "logic" see [GW01].

agents (although it includes some metalevel elements from epistemic logic such as the ability for the agent to reason about its own beliefs).

Reasoning in active logic proceeds in a step-wise manner using rules like

$$\frac{t :}{t+1 : now(t+1)}$$

and active versions of classical rules such as *modus ponens*:

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

The rules are applied to the agent's knowledge base, which is a set of formulae representing the agent's current beliefs. A belief φ at time t is not necessarily believed at the next time step. Rather, beliefs are "inherited" using rules such as

$$\frac{t : \varphi}{t+1 : \varphi}$$

[condition: $\neg\varphi$ is not believed at time t and $\varphi \neq now(t)$],

which allow for a kind of default reasoning. Note, that active logic is by its nature non-monotonic (e.g. $now(t)$ is retracted at time $t+1$ for every t).

The step-wise control of the deductive process allows inference rules to refer to previous steps in the reasoning history. In particular, this allows the agent to reason about its own past, thereby making it possible, for example, to reason about the causes of arisen contradictions as well as possible remedial actions to take in response to them. Current implementations of active logic use rules like

$$\frac{t : \varphi, \neg\varphi}{t+1 : contra(\ulcorner\varphi\urcorner, \ulcorner\neg\varphi\urcorner, t)},$$

whereby direct contradictions (φ and $\neg\varphi$) can be spotted and some reasoning process to handle them initiated.

Now, what we have is a proof-theoretical (syntactical) characterisation of active logic, and we would like to have a model-theoretical (semantical) characterisation as well. In particular, we are in this thesis concerned with a model-theoretical characterisation at the object level – of modelling the evolving set of explicit beliefs of a reasoning agent (expressed in the agent's own language).

"Semantics" is probably best translated with "study of meaning", and the term itself is a derivation from the Greek word *sema*, which means "sign" [vW65]. The (philosophical) semantics concerns concepts such as meaning, truth, and soundness. In particular, semantics involve connecting formal languages and mathematical structures through the notion of truth. Following Tarski we shall say that a sentence φ is true in a structure $\mathfrak{A}$, or

that $\mathfrak{A}$ is a model of φ, if it is actually the case that φ holds in $\mathfrak{A}$. (The sentence "Snow is white" is true if snow is actually white.) [Man99]

We will later define a formal language of active logic and mathematical structures that (hopefully) will be appropriate for modelling reasoning that uses this formal language. We will also define formally when a sentence of this language holds in such a structure. In particular, we will – in contrast to classical logic – define a notion of structure that allows us to model inconsistencies. This way, we shall try to define a semantical concept of consequence that does not suffer from the explosiveness of classical logical consequence. (The explosiveness of the classical consequence relation follows from the fact that inconsistent sets of sentences have no classical models.)

There are many related theories on resource-bounded reasoning that give semantical accounts of reasoning agents at the metalevel (e.g. [GKP00, Ågo04]). That is, they model a theory (expressed in a metalanguage) that describes an agent that reasons using some (object-level) formalism such as active logic. Metalevel modelling can be done using classical logical formalisms since even an incomplete and inconsistent object-level theory has a corresponding complete and consistent theory at the metalevel. For example, even if an agent believes in a contradiction, say φ and $\neg\varphi$, the metalevel propositions *"the agent believes* $\ulcorner\varphi\urcorner$*"* and *"the agent believes* $\ulcorner\neg\varphi\urcorner$*"* do not constitute a contradiction (see also Section 2.4). When modelling at the object level on the other hand, one cannot resort to classical logic because inconsistent theories have no models in classical formalisms. We are not aware of any theory besides the one recently published in [AGGP05a] that tries to model the reasoning of real-world agents at the object level.

The step-model semantics that accompanied the original presentation of step logic (e.g. in [EDP90]) does not model at the object level either in that it does not model individual belief sets. The only theory (again, that we are aware of) that models individual belief sets of real-world agents and has a (semantical) consequence relation between such sets is the one in [AGGP05a]. Step-model semantics models the complete agent theory, that is, the infinite sequence of belief sets of a reasoning agent (expressed in the agent language, though) and lacks, and cannot have, a consequence relation between single belief sets. In this sense, it is also a metalevel semantics. Furthermore, the step model semantics is restricted to infinite sequences of *consistent* knowledge bases, and thus cannot be used to model real-world agents which may have inconsistent belief sets at times. In particular, step-model semantics does not allow the agent, at a particular time, to have a model of its potentially inconsistent beliefs.

Since active logic allows the agents to reason about their own beliefs (i.e. to do some metareasoning), the agents may have metabeliefs, and these require metalevel modelling. This part of the semantics thus may be influenced by, or use elements from, theories on metalevel modelling of reasoning agents. Where step models depend on the whole sequence of belief sets of

an agent (i.e. also on future beliefs), the perceived temporal structures of the new semantics "only" depend on what has been (and what is) believed. That is, also the new semantics refers to relations outside the current belief set, namely to the actual reasoning history of the agent. In this sense, also the new theory is a metalevel theory (if only in part).

The focus of this thesis is on the semantics for active logic proposed in [AGGP05a]. This first serious attempt of modelling an agent's beliefs at the object level is analysed and refined, and some useful metatheorems are proved. The semantics is found to have several flaws – the gravest being the failure to deliver a consequence relation which would make the logic paraconsistent.

The rest of this thesis is organised as follows. Section 2 presents (a reformulation of) the semantics for active logic from [AGGP05a]. Section 3 focuses on problems with the semantics and suggests some refinements of the theory. Several metatheorems are also proved here. In Section 4 we show that the proposed consequence relation is explosive. Our results are summarised in Section 5, where some points for future work are also outlined.

2 Perception Functions and Active Consequence

In [AGGP05a] a semantics for active logic based on what the authors call perception functions is proposed. In this first sketch of the semantics, a heavily weakened active logic is considered. Most notably the language is based on propositional logic rather than full first-order logic. Furthermore, a useful distinction between world and agent language is introduced, allowing further restrictions on the logic. The semantics include semantic-like concepts such as *active consequence* and *active soundness*, as well as some theorems regarding their relations to their classical counterparts.

The following presentation of the logic and its semantics follows the one given in [AGGP05a]. With the intention to increase stringency and understandability, the original theory has been reformulated and some remarks and examples have been added. Several errors in the original paper have also been corrected.[2]

While this thesis was being finished, an extended article about the semantics has been submitted for publishing [AGGP05b]. We had already pointed out some of the errors and ambiguities found in the original paper to the authors, and some of our suggestions have been acknowledged in the new article.[3] Otherwise, the theory is essentially the same (although perception functions are called *apperception functions*), and several problems still remain.

Our analysis of the semantics is postponed to Section 3.

2.1 Starting Assumptions

The logic is presented under following assumptions

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

[2] In particular, the definitions of the language $\mathcal{L}_a$ and its semantics have been heavily modified, the language $\mathcal{L}$ has been more precisely defined and a definition of $\mathcal{L}$-satisfaction has been added. The definition of perception functions has also been heavily modified. The definitions of active consequence and of the timing rule have been changed slightly. The proof of Theorem 2.1 has been rewritten and corrected, and the proof of Theorem 2.2 has been corrected and extended so that the theorem is now (fully) proved. The proof of Theorem 2.4 has been rewritten. The "theorem" concerning the active unsoundness of the explosive rule has been disproved. Examples 2.1, 2.2 and 2.3 are new, and examples 2.4 and 2.7 have been rewritten.

[3] The definitions of the language $\mathcal{L}$, $\mathcal{L}_a$-structure, perception function and active consequence (via G_{per_t}) have been modified. The proof of Theorem 2.1 has also been corrected.

However, we will in our presentation in a sense relax the assumption of one sole agent. In particular, we shall interpret it as meaning that we only model one agent at a time and that agents can only reason about their own beliefs (and not about the beliefs of other agents). We will in Section 3.4 argue that this interpretation is in conformity with the original theory.

2.2 The Language $\mathcal{L}$

The language used is a sorted first-order language $\mathcal{L}$ defined in two parts: the propositional language $\mathcal{L}_w$ used to express facts about the world and the first-order language $\mathcal{L}_a$ used to express facts about the agent and the agent's beliefs. Let Sn_L denote the set of all sentences of a language L, that is, the set of all closed formulae of L. (Note that for a language L without variables, Sn_L coincides with the set of formulae of L.) Let $\mathbb{N}$ denote the set of natural numbers (non-negative integers).

Definition 2.1 ($\mathcal{L}_w$). Let $\mathcal{L}_w$ be the propositional language consisting of the following symbols:

- a set $S = \{S_i \mid i \in \mathbb{N}\}$ of sentence symbols (propositional or sentential variables),
- the propositional connectives $\neg$ and $\rightarrow$, and
- parentheses (and).

The formulae of $\mathcal{L}_w$ are defined in the standard way: Every sentence symbol is an atomic formula, and if φ and ψ are formulae, then so are $\neg\varphi$ and $(\varphi \rightarrow \psi)$.

Note that we will allow ourselves to drop the parentheses from formulae when there is no risk for ambiguity.

Definition 2.2 ($\mathcal{L}_a$). Consider the sequence $\langle \mathcal{L}_n \rangle_{n\in\mathbb{N}}$ of restricted, sorted first-order languages, where each $\mathcal{L}_n$ has three sorts, $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$, but no variables or quantifiers. The sequence is defined inductively as follows.[4]

1. The language $\mathcal{L}_0$ consists of the following symbols:
   - a set $C = \{c_i \mid i \in \mathbb{N}\}$ of constant symbols of sort $\mathcal{S}_2$,
   - a set $D = \{d_\varphi \mid \varphi \in Sn_{\mathcal{L}_w}\}$ of constant symbols of sort $\mathcal{S}_1$,
   - a set $E_0 = \{e_\varphi \mid \varphi \in Sn_{\mathcal{L}_w}\}$ of constant symbols of sort $\mathcal{S}_3$,
   - the unary predicate symbol *now* of sort $\mathcal{S}_2$,
   - the ternary predicate symbol *contra* of sort $(\mathcal{S}_1 \times \mathcal{S}_1 \times \mathcal{S}_2)$,

[4]This definition is inspired by [AGGP05b]. The original definition of $\mathcal{L}_a$ suffered from problems related to self-reference.

- the binary predicate symbol *bel* of sort $(\mathcal{S}_3 \times \mathcal{S}_2)$, and
- the propositional connective $\neg$.

The formulae of $\mathcal{L}_0$ are defined in the standard way: Every predicate symbol applied to constant symbols of the appropriate sorts is an atomic formula, and if φ is a formula, then $\neg\varphi$ is a compound formula.

2. The language $\mathcal{L}_{n+1}$, $n \geq 0$, consists of the same symbols as $\mathcal{L}_n$ and an additional set $E_{n+1} = \{e_\varphi \mid \varphi \in Sn_{\mathcal{L}_n}\}$ of constant symbols of sort $\mathcal{S}_3$.

   The formulae of $\mathcal{L}_{n+1}$ are defined analogously to those of $\mathcal{L}_0$.

We define $\mathcal{L}_a$ to be the language consisting of the symbols of every language in $\langle\mathcal{L}_n\rangle_{n\in\mathbb{N}}$. In particular, $\mathcal{L}_a$ consists of the same symbols as $\mathcal{L}_0$ and a set $E = \bigcup_{n\in\mathbb{N}} E_n$ of constant symbols. The formulae of $\mathcal{L}_a$ are defined analogously to those of $\mathcal{L}_0$. In particular, $Sn_{\mathcal{L}_a} = \bigcup_{n\in\mathbb{N}} Sn_{\mathcal{L}_n}$.

The intended meanings of *now*, *contra*, and *bel* are to indicate the current time, the existence of a direct contradiction at some time, and that an agent had a belief at some time, respectively. The semantics of the symbols is defined formally in Definition 2.9 below. Note that $\mathcal{L}_a$ only includes one connective $\neg$, thus heavily limiting the expressiveness of the language. The sets of constants are introduced to name time points and sentences of the languages.

Example 2.1. The following formulae are all sentences of $\mathcal{L}_a$: $now(c_5)$, $contra(d_{S_1}, d_{\neg S_1}, c_7)$, and $\neg bel(e_{now(c_5)}, c_4)$.

Definition 2.3 ($\mathcal{L}$). The language $\mathcal{L}$ is the sorted first-order language consisting of the symbols of the languages $\mathcal{L}_w$ and $\mathcal{L}_a$. The symbols are sorted, and the connectives only apply to formulae of either language using standard syntax. In particular, the set of sentences of $\mathcal{L}$, $Sn_{\mathcal{L}}$, is the union of the disjoint sets $Sn_{\mathcal{L}_w}$ and $Sn_{\mathcal{L}_a}$.

Note that the language $\mathcal{L}$ is defined in such a way that its set of sentences is the union of the sentences of $\mathcal{L}_w$ and $\mathcal{L}_a$, respectively. In particular, formulae such as $now(c_5) \rightarrow S_1$ are are not well formed because the implication connective applies only to $\mathcal{L}_w$-formulae (i.e. the connective, which could be written $\rightarrow_w$ as it stems from $\mathcal{L}_w$, is sorted). The definition of the language $\mathcal{L}$ and its consequences for the logic are discussed further in Section 3.2.[5]

The agent's knowledge is expressed within $Sn_{\mathcal{L}}$. In fact, at any time t, the agent's knowledge base KB^a_t is identified with a finite[6] subset of $Sn_{\mathcal{L}}$.[7]

[5]The definition presented here is an interpretation of the original definition found in [AGGP05a]. The possibility of other interpretations is also discussed later.

[6]Note that, although not explicitly stated in [AGGP05a], it is implicit that the agent's knowledge base is always finite. For example, perception functions (to be defined later), which are applied to knowledge bases, are functions with finite subsets of $Sn_{\mathcal{L}}$ as domain. In [AGGP05b] it has been made explicit.

[7]Cf. Ågotnes' term *finite syntactic epistemic states* [Ågo04].

The knowledge base will thus always be incomplete, and it may at any time also be incorrect (with respect to the actual world) or contradictory.

Definition 2.4 (KB_t^a). The *knowledge base* of an agent a at time t, denoted KB_t^a, is a finite subset of $Sn_{\mathcal{L}}$.

2.3 Semantics for $\mathcal{L}_w$

In the following definitions, a semantics for $\mathcal{L}_w$ is defined in the standard way.

Definition 2.5 ($\mathcal{L}_w$-Truth Assignment). An *$\mathcal{L}_w$-truth assignment* is a function $h : S \to \{\top, \bot\}$, which assigns truth values to all sentence symbols in $\mathcal{L}_w$.

Definition 2.6 ($\mathcal{L}_w$-Interpretation). An *$\mathcal{L}_w$-interpretation* h (keeping the same notation) is a function $h : Sn_{\mathcal{L}_w} \to \{\top, \bot\}$ that extends an $\mathcal{L}_w$-truth assignment h as follows:

$$\begin{aligned} h(\neg\varphi) = \top \quad &\text{iff} \quad h(\varphi) = \bot \\ h(\varphi \to \psi) = \bot \quad &\text{iff} \quad h(\varphi) = \top \text{ and } h(\psi) = \bot \end{aligned}$$

("Iff" means "if and only if".) If $h(\varphi) = \top$, we say that h is a model of φ and that h satisfies φ. We will also use the notation $h \models \varphi$ for $h(\varphi) = \top$.

Definition 2.7 ($\mathcal{L}_w$-Consistency). A set of sentences $\Sigma \subseteq Sn_{\mathcal{L}_w}$ is *consistent* if there exists an $\mathcal{L}_w$-interpretation h in which all the sentences are true. Denote with $h \models \Sigma$ the fact that every sentence of Σ is assigned $\top$ by h. If $h \models \Sigma$, we say that h is a model of Σ and that h satisfies Σ.

2.4 The $\mathcal{L}_a$-Model

Even an incomplete and inconsistent agent theory has a corresponding complete and consistent theory at the metalevel. For instance, if the agent believes both S_1 and $\neg S_1$, then its theory is inconsistent. As mentioned above, an agent theory is always incomplete since the knowledge base is finite. But at the metalevel, a theory containing the sentences *"the agent believes* $\ulcorner S_1 \urcorner$*"* and *"the agent believes* $\ulcorner \neg S_1 \urcorner$*"* at the same time, need not be inconsistent. Furthermore, if it includes a countable number of sentences of the form '*"the agent does not believe* $\ulcorner \varphi \urcorner$*"* it may also be complete.

Such a metatheory may be expressed using the language $\mathcal{L}_a$. Below, a semantical account is given by introducing a structure that models a theory at a given time t.

Definition 2.8 ($\mathcal{L}_a$-Structure). The *$\mathcal{L}_a$-structure* H_t^a at time t, given the agent's knowledge bases KB_k^a for $0 \leq k \leq t$, is the structure

$$H_t^a = \left\langle Sn_{\mathcal{L}_w}, \mathbb{N}, Sn_{\mathcal{L}}, \langle \mathbf{c}_k \rangle_{k\in\mathbb{N}}, \langle \mathbf{d}_\varphi \rangle_{\varphi\in Sn_{\mathcal{L}_w}}, \langle \mathbf{e}_\varphi \rangle_{\varphi\in Sn_{\mathcal{L}}}, \mathbf{now}, \mathbf{bel}, \mathbf{contra} \right\rangle,$$

where

1. for every $k \in \mathbb{N}$, $\mathbf{c}_k$ names the time index k,
2. for every $\varphi \in Sn_{\mathcal{L}_w}$, $\mathbf{d}_\varphi$ names the sentence φ,
3. for every $\varphi \in Sn_{\mathcal{L}}$, $\mathbf{e}_\varphi$ names the sentence φ,
4. the relation $\mathbf{now} \subseteq \mathbb{N}$ has only one element t,
5. the relation $\mathbf{bel} \subseteq (Sn_{\mathcal{L}} \times \mathbb{N})$ is the set $\{\langle \varphi, k\rangle \mid k \leq t \text{ and } \varphi \in KB_k^a\}$, and
6. the relation $\mathbf{contra} \subseteq (Sn_{\mathcal{L}_w} \times Sn_{\mathcal{L}_w} \times \mathbb{N})$ is the set

$$\{\langle \varphi, \psi, k\rangle \mid k \leq t,\ \{\varphi, \psi\} \subseteq KB_k^a, \text{ and either } \varphi = \neg\psi \text{ or } \psi = \neg\varphi\}.$$

Note that the $\mathcal{L}_a$-structure depends heavily on the agent's actual reasoning process, that is, on the sequence $\langle KB_k^a\rangle_{k=0}^t$: the **bel**-relation contains the agent's complete history, and the **contra**-relation stores a complete record of all direct contradictions (involving world knowledge) that have ever occurred.

The sentences of $\mathcal{L}_a$ are interpreted in an $\mathcal{L}_a$-structure in the standard way:

Definition 2.9 (Satisfaction in H_t^a). Let H_t^a be an $\mathcal{L}_a$-structure at time t. Then *satisfaction in* H_t^a, written $H_t^a \models \varphi$, is defined inductively by

1. $H_t^a \models now(c_k)$ iff $\mathbf{c}_k \in \mathbf{now}$,
2. $H_t^a \models contra(d_\varphi, d_\psi, c_k)$ iff $\langle \mathbf{d}_\varphi, \mathbf{d}_\psi, \mathbf{c}_k\rangle \in \mathbf{contra}$,
3. $H_t^a \models bel(e_\varphi, c_k)$ iff $\langle \mathbf{e}_\varphi, \mathbf{c}_k\rangle \in \mathbf{bel}$, and
4. $H_t^a \models \neg\varphi$ iff $H_t \not\models \varphi$.

Let $H_t^a \models \Sigma$ denote the fact that H_t^a satisfies every sentence in the set $\Sigma \subseteq Sn_{\mathcal{L}_a}$. If $H_t^a \models \varphi$, $\varphi \in Sn_{\mathcal{L}_a}$, we say that φ is true in H_t^a, otherwise φ is false in H_t^a.

Example 2.2. Let H_t^a be an $\mathcal{L}_a$-structure at time t. Then, a sentence $now(c_k)$ is true (in H_t^a) if and only if $k = t$, that is, if the time is actually k. Furthermore, a sentence $bel(e_\varphi, c_k)$ is true if and only if the agent actually believed φ at time $k \leq t$. Finally, a sentence $contra(d_\varphi, d_\psi, c_k)$ is true if and only the agent actually believed both φ and $\neg\varphi$ (or, ψ and $\neg\psi$) at time $k \leq t$ *and* φ (and ψ) are about the world (i.e. they are sentences of $\mathcal{L}_w$). Note that the meaning of "believe" in this paragraph (and thesis) is simply that the sentence believed was in the agent's knowledge base.

2.5 $\mathcal{L}$-Structures

As mentioned above, the agent expresses its knowledge in $Sn_{\mathcal{L}}$. To be able to model the agent's full reasoning, we need to define a structure appropriate for $\mathcal{L}$-theories.

Definition 2.10 (Active Structure). An *active structure*, or *a-structure*, at time t is an $\mathcal{L}$-structure M_t^a defined as

$$M_t^a = \langle h_t, H_t^a \rangle,$$

where h_t is an $\mathcal{L}_w$-interpretation and H_t^a is an $\mathcal{L}_a$-structure at time t.

Note that active structures depend implicitly on the agent's history of reasoning (the sequence $\langle KB_k^a \rangle_{k=0}^t$) via the $\mathcal{L}_a$-structure H_t^a.

Since the set of sentences of $Sn_{\mathcal{L}}$ is the union of the two disjoint sets $Sn_{\mathcal{L}_w}$ and $Sn_{\mathcal{L}_a}$, satisfaction in $\mathcal{L}$-structures is straightforward to define.[8]

Definition 2.11 (Satisfaction in M_t^a). Let $M_t^a = \langle h_t, H_t^a \rangle$ be an active structure at time t. Then *satisfaction in* M_t^a, written $M_t^a \models \varphi$, is defined by

1. if $\varphi \in Sn_{\mathcal{L}_w}$, then ($M_t^a \models \varphi$ iff $h_t \models \varphi$), and
2. if $\varphi \in Sn_{\mathcal{L}_a}$, then ($M_t^a \models \varphi$ iff $H_t^a \models \varphi$).

We say that M_t^a satisfies $\Sigma \subseteq Sn_{\mathcal{L}}$, denoted $M_t^a \models \Sigma$, if M_t^a satisfies every sentence in Σ. If $M_t^a \models \Sigma$ we also say that M_t^a is a model of Σ.

2.6 Perception Functions

In this section an attempt to model the agent's beliefs at the object level is made. First, two notions of *temporal consistency* relative to the language $\mathcal{L}$ are defined.

Definition 2.12 (Temporal Strong Consistency). A set of sentences $\Sigma \subseteq Sn_{\mathcal{L}}$ is said to be *temporally strongly consistent at time t*, t-strongly consistent for short, if there exists an active structure M_t^a such that $M_t^a \models \Sigma$.

Definition 2.13 (Temporal Weak Consistency). A set of sentences $\Sigma \subseteq Sn_{\mathcal{L}}$ is said to be *temporally weakly consistent at time t*, t-weakly consistent for short, if there exists an active structure M_t^a such that $M_t^a \models (\Sigma \cap Sn_{\mathcal{L}_a})$.

That is, a set of $\mathcal{L}$-sentences is t-weakly consistent if at least the $\mathcal{L}_a$-sentences of the set are consistent. For t-strong consistency, also the $\mathcal{L}_w$-sentences of the set must be consistent. Note that we will often drop the time specifier "at time t" as it is usually clear from context.

From now on, it is assumed that the agent's knowledge base is t-weakly consistent.

[8] This definition is not found in [AGGP05a] and of course depends on how we define $\mathcal{L}$, which also is an object of interpretation (see Section 3.2).

Definition 2.14 (Σ_t^ω). Let Σ_t^ω denote the set of finite, t-weakly consistent subsets of $Sn_\mathcal{L}$ at time t.

In order to express the agent's awareness of its knowledge about the world, a new language $\mathcal{L}'_w$ is defined.

Definition 2.15 ($\mathcal{L}'_w$). The propositional language $\mathcal{L}'_w$ derived from $\mathcal{L}_w$ consists of the following symbols:

- a set $S' = \{S_i^j \mid S_i \in \mathcal{L}_w \text{ and } j \in \mathbb{N}\}$ of sentence symbols,
- the propositional connectives $\neg$ and $\rightarrow$, and
- parentheses (and).

The formulae of $\mathcal{L}'_w$ are defined as in the definition of $\mathcal{L}_w$.

Thus, for every sentence symbol in $\mathcal{L}_w$ there is a corresponding infinite pool of sentence symbols in $\mathcal{L}'_w$. We define $\mathcal{L}'_w$-truth assignments $h : S' \rightarrow \{\top, \bot\}$ and $\mathcal{L}'_w$-interpretations $h : Sn_{\mathcal{L}'_w} \rightarrow \{\top, \bot\}$, and satisfaction and consistency for $\mathcal{L}'_w$ analogously to their $\mathcal{L}_w$-counterparts.

Definition 2.16 ($\mathcal{L}'$). Let $\mathcal{L}'$ be the sorted first-order language consisting of the (sorted) symbols of the languages $\mathcal{L}'_w$ and $\mathcal{L}_a$. In particular, its sentences is the set $Sn_{\mathcal{L}'} = Sn_{\mathcal{L}'_w} \cup Sn_{\mathcal{L}_a}$.

The language $\mathcal{L}'$ is used to express the agent's awareness of its agent-and-world knowledge. Note that the definitions of temporal consistency can be extended to $\mathcal{L}'$ (we will define formally structures corresponding to active structures at the end of this section). When clear from context, the language will not be mentioned.

The notion of *perception (awareness) function*, to be formally defined below, is intended to help capture, at least roughly, how the world might seem to the agent. The idea is that the agent's limited resources apply also to its ability to inspect its own knowledge base. Even if both S_1 and $\neg S_1$ are present in the knowledge base, the agent may be unaware of the contradiction by maintaining that S_1^1 and $\neg S_1^2$ are both true. Only later might it discover that S_1^1 and S_1^2 are in fact the same symbol (S_1) and start a process to handle the contradiction. This allows the agent to have inconsistent beliefs while still having a (classical) world model (see Section 3.3 for a critique of this approach). Furthermore, this mechanism can also be used to prevent the agent from deriving everything from an inconsistent knowledge base.

Let $\mathcal{P}(\Sigma)$ denote the power set of a set Σ.

Definition 2.17 (Perception Function). A *perception (awareness) function* at time t is a map $per_t : \Sigma_t^\omega \rightarrow \mathcal{P}(Sn_{\mathcal{L}'})$ that is defined by an infinite sequence of non-negative integers $\langle i_1, i_2, \ldots \rangle$ and

1. Let $\Sigma \in \Sigma_t^\omega$ and denote with Γ the set $\Sigma \cap Sn_{\mathcal{L}_w}$. Order the sentences of Γ alphabetically in a string, and let $\langle S_{j_1}, S_{j_2}, \ldots, S_{j_n} \rangle$ be the finite sequence of all sentence-symbol tokens occurring in this string. Let Γ' be the set of sentences obtained by replacing S_{j_k} in Γ with $S_{j_k}^{i_k}$ for $1 \leq k \leq n$.

2. Let $p : \Gamma \to \Gamma'$ be the bijection mapping every sentence of Γ to its corresponding Γ'-sentence.

3. Let the set of (perceived) direct contradictions, denoted DC, be the set
$$\{\varphi \in \Gamma \mid p(\varphi) \in \Gamma' \text{ and } \neg p(\varphi) \in \Gamma'\}.$$

4. Finally, let the image of Σ under per_t be the set
$$\begin{aligned}&(\Sigma - \Gamma) \cup \Gamma' - \\ &\{p(\varphi) \mid \varphi \in DC\} - \{\neg p(\varphi) \mid \varphi \in DC\} \cup \\ &\{contra(d_\varphi, d_{\neg\varphi}, c_t\} \mid \varphi \in DC\}.\end{aligned}$$

We denote with PER_t the set of all perception functions at time t.

Informally, what the above definition says is the following. A perception function only applies to world knowledge (Γ), by mapping every occurrence of a sentence-symbol token (S_{j_k}) in the set to a corresponding superscripted symbol (resulting in Γ'). Note that this mapping is completely determined by the sequence $\langle i_1, i_2, \ldots \rangle$. Once the sentence-symbol tokens have been superscripted, any direct contradiction in the set (Γ') is removed and a record, in form of a *contra*-sentence, is added.

Note that the set PER_t is infinite for every t since there are infinitely many ways of assigning superscripts to sentences (i.e. infinitely many sequences $\langle i_1, i_2, \ldots \rangle$).

Example 2.3. Let Σ be the set $\{now(c_5), S_1, \neg S_2, \neg S_1\}$, and let per_5 be the perception function (at time 5) determined by the infinite superscript sequence $\langle 1, 2, 1, \ldots \rangle$ (only the first three elements are shown). We now wish to apply per_5 to Σ. With the terminology of Definition 2.17, we get $\Gamma = \{S_1, \neg S_1, \neg S_2\}$ (presented in alphabetically ordered form). Hence the finite sequence of sentence-symbol tokens is $\langle S_1, S_1, S_2 \rangle$, and thus we get $\Gamma' = \{S_1^1, \neg S_1^2, \neg S_2^1\}$. Since there are no direct contradictions in Γ', we have $per_5(\Sigma) = \{now(c_5), S_1^1, \neg S_1^2, \neg S_2^1\}$.

Should instead both occurrences of S_1 have been mapped to the same symbol, say S_1^1, we would have had a direct contradiction and the resulting image would then have been the set

$$\{now(c_5), \neg S_2^1, contra(d_{S_1}, d_{\neg S_1}, c_5)\},$$

instead.

Theorem 2.1. *If KB_t^a is t-weakly consistent at time t, then there exists a perception function $per_t \in PER_t$ such that $per_t(KB_t^a)$ is t-strongly consistent (in $\mathcal{L}'$) at time t.*

Proof. Let $\Gamma = KB_t^a \cap Sn_{\mathcal{L}_w}$ and consider the perception function per_t^u determined by the sequence $\langle 1, 2, 3, \ldots \rangle$. Applying per_t^u according to Definition 2.17 renders a set Γ' in which every sentence-symbol token is unique. Thus, there exists an $\mathcal{L}'_w$-interpretation which satisfies Γ' since we can choose the truth value of every atomic formula independently.[9] (A concrete truth assignment can easily be constructed through a recursive procedure.) Since the remaining sentences of KB_t^a are assumed consistent, $per_t(KB_t^a)$ is t-strongly consistent. □

Definition 2.18 ($KB^a_{per_t}$ and $W^a_{per_t}$). Let $per_t \in PER_t$ be a perception function at time t. Let $KB^a_{per_t}$ denote the agent's perception (under per_t) of its knowledge base at time t, that is, $KB^a_{per_t} = per_t(KB_t^a)$. Let $W^a_{per_t} = KB^a_{per_t} \cap Sn_{\mathcal{L}'_w}$ be the agent's perception of the part of its knowledge base which concerns the external world.

Definition 2.19 ($G^a_{per_t}$). Let $per_t \in PER_t$ be a perception function at time t. Define $G^a_{per_t}$ to be the set of $\mathcal{L}'_w$-interpretations determined by $W^a_{per_t}$, that is, $G^a_{per_t} = \{h_{per_t} \mid h_{per_t} \models W^a_{per_t}\}$.

Remark. Note that, without further restrictions on per_t, $W^a_{per_t}$ need not be consistent and thus the set $G^a_{per_t}$ might be empty (contrary to what is claimed in [AGGP05a]).[10] The consequences of this remark will be discussed in Section 3.3.

Finally, an $\mathcal{L}'$-structure used to model the agent's knowledge base after a perception function has been applied to it is defined. This is meant to capture the way the world might seem to an agent at a given time.

Definition 2.20 (Perceived Temporal Structure). Let $per_t \in PER_t$ be a perception function at time t. Then a *perceived temporal structure* at time t, pt-structure for short, is an $\mathcal{L}'$-structure $M^a_{per_t}$ defined as follows: $M^a_{per_t} = \langle h_{per_t}, H_t^a \rangle$, where h_{per_t} is an $\mathcal{L}'_w$-interpretation satisfying $W^a_{per_t}$ (i.e. $h_{per_t} \in G^a_{per_t}$) and H_t^a is an $\mathcal{L}_a$-structure at time t. Let $\mathbb{M}^a_{per_t}$ denote the set of all pt-structures $M^a_{per_t}$.

[9] Note that the $\mathcal{L}'_w$-interpretation used in the proof of this theorem in [AGGP05a] does not model every Γ'. Consider, for instance, the set $\Gamma = \{\neg(S_1 \rightarrow S_2)\}$ which is mapped to $\Gamma' = \{\neg(S_1^1 \rightarrow S_2^2)\}$. Now, both sentence symbols occur positively and are thus assigned the value true. Consequently, the whole sentence is false, contrary to the claim of the proof.

[10] Consider, for example, $KB_t^a = \{S_1, \neg S_1\}$ and a perception function per_t mapping every token S_{j_k} to $S^1_{j_k}$. Then $W^a_{per_t}$ is the inconsistent set $\{S_1^1, \neg S_1^1\}$, which has no model.

Note once again that since $G^a_{per_t}$ might be empty, the existence of a pt-structure at time t is not guaranteed (thus, $\mathbb{M}^a_{per_t}$ might be empty as well).

Satisfaction in pt-structures is defined analogously to satisfaction in active structures.

2.7 Active Consequence

In this section we introduce the concept of *active consequence* – a concept purported to be the active-logic equivalent of logical consequence.

Definition 2.21 (1-Step Active Consequence). Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$, and let a be an agent with $KB^a_t = \Sigma$. We say that Θ is a *1-step active consequence* of Σ at time t, written $\Sigma \models_1 \Theta$, if and only if

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

Roughly, *if* for the $\mathcal{L}_w$-sentences, the set of conclusions as perceived by the agent at time $t+1$ are yielded by the antecedent as perceived by the agent at time t, *and if* for the $\mathcal{L}_a$-sentences, the $\mathcal{L}_a$-structure H^a_{t+1} models the agent's perception of the conclusions at time $t+1$, *then* it can be said that the conclusions are 1-step active consequences of the antecedent.

Note, that in the above definition, the apparently unbound variable H^a_{t+1} (it is not a constituent of $M^a_{per_t}$) must be interpreted as a constant in order for the formula to be well-formed. We will return to this in Section 3.[11]

Let us demonstrate the concept with a few examples.

Example 2.4. Let $\Sigma = \{\varphi, \neg\varphi\}$ and $\Theta = \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$. We want to show that $\Sigma \models_1 \Theta$ at time t. To do this we first fix an agent a with $KB^a_t = \Sigma$. Then we pick two perception functions per_t and per_{t+1} at time t and $t+1$ respectively: in this case, we can choose an arbitrary per_t, and we choose a per_{t+1} such that $per_{t+1}(\Theta) = \{contra(d_\varphi, d_{\neg\varphi}, c_t)\}$. Note that such functions do exist: PER_t is non-empty (as noted above, it is infinite) for all t, and any $per_{t+1} \in PER_{t+1}$ has the desired image (since perception functions only "modify" $\mathcal{L}_w$-sentences). Now, since the $\mathcal{L}_a$-structure H^a_{t+1} contains a complete record of any direct contradiction present in the agent's knowledge base at any time $k \leq t+1$, we have in particular that $H^a_{t+1} \models contra(d_\varphi, d_{\neg\varphi}, c_t)$ since $\{\varphi, \neg\varphi\} \subseteq KB^a_t$ (see Definition 2.9). Since $per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w}$ is empty, the second conjunct in Definition 2.21 is satisfied vacuously (this is why we could choose an arbitrary per_t). We have shown that $\Sigma \models_1 \Theta$.

[11] Note also that there is a slight difference between the definition of 1-step active consequence given above and the definition found in [AGGP05a]. The original definition reads: "Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$ such that $\Sigma = KB^a_t$. ...". This difference is discussed and motivated in Section 3.4.

Example 2.5. Let $\Sigma = \{now(c_t), S_1, S_1 \rightarrow S_4, S_{12}\}$ and let Θ be the set $\{now(c_{t+1}), S_4, S_{12}\}$. Again, we want to show that $\Sigma \models_1 \Theta$ at time t.

We first fix an agent a with $KB_t^a = \Sigma$. Let $per_t \in PER_t$ be a perception function with $per_t(\Sigma) = \{now(c_t), S_1^i, S_1^i \rightarrow S_4^j, S_{12}^k\}$ for some non-negative integers i, j and k. Then for every $h_{per_t} \in G^a_{per_t}$ the following must hold: $h_{per_t}(S_1^i) = h_{per_t}(S_4^j) = h_{per_t}(S_{12}^k) = \top$. Choose $per_{t+1} \in PER_{t+1}$ such that $per_{t+1}(\Theta) = \{now(c_{t+1}), S_4^j, S_{12}^k\}$. Clearly, $H^a_{t+1} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a})$ and $h_{per_t} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w})$ for every $h_{per_t} \in G^a_{per_t}$. We have shown that $\Sigma \models_1 \Theta$.

Note that once the perception functions have been applied, determining whether the $\mathcal{L}_w$-sentences of Θ follow actively from Σ is similar to determining this classically, whereas for the $\mathcal{L}_a$-sentences it is just a matter of determining whether they are modelled by H^a_{t+1} according to Definition 2.9. However, as we shall see in Section 3.4, this may not always be as straightforward as it sounds, and we will in fact eventually be forced to refine the concept of 1-step active consequence.

We will settle with one last example.

Example 2.6. Let Σ and Θ be as in the previous example, but with $bel(e_{S_5}, c_t)$ added to Θ. Now, since $S_5 \notin \Sigma$ (and hence, $S_5 \notin KB_t^a$) we have that $H^a_{t+1} \not\models bel(e_{S_5}, c_t)$, and consequently that $\Sigma \not\models_1 \Theta$ at time t.

A generalisation of 1-step active consequence is defined recursively[12]:

Definition 2.22 (n-Step Active Consequence). Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$, and let n be a positive integer. We say that Θ is an *n-step active consequence* of Σ at time t, written $\Sigma \models_n \Theta$, if there exists a set $\Gamma \subseteq Sn_{\mathcal{L}}$ such that $\Sigma \models_{n-1} \Gamma$ at time t and $\Gamma \models_1 \Theta$ at time $t + n - 1$.

Finally, active consequence is defined as follows.

Definition 2.23 (Active Consequence). Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$. We say that Θ is an *active consequence* of Σ at time t, written $\Sigma \models_a \Theta$, if Θ is an n-step active consequence of Σ at time t for some positive integer n.

Remark. When clear from context, the time specification "at time t" will sometimes be left out when discussing active consequence.

Example 2.7. We show that $\Sigma \models_a \Theta$ at time t with $\Sigma = \{S_1, S_2, S_2 \rightarrow \neg S_1\}$ and $\Theta = \{contra(d_{S_1}, d_{\neg S_1}, c_{t+1})\}$.

Let $\Gamma = \{S_1, \neg S_1\}$. We first show that $\Sigma \models_1 \Gamma$ at time t. Fix an agent a with $KB_t^a = \Sigma$. Let $per_t \in PER_t$ be a perception function with

[12]The original restriction that $\Sigma = KB_t^a$ has been removed from the definition (and from Definition 2.23) since it is implicit from Definition 2.21. Furthermore, the time specifications, which were mediated through examples in [AGGP05a], have been made explicit.

$per_t(\Sigma) = \{S_1^i, S_2^j, S_2^j \to \neg S_1^k\}$ for some $i, j, k \in \mathbb{N}$ with $i \neq k$. Then, for every $h_{per_t} \in G^a_{per_t}$ the following must hold: $h_{per_t}(S_1^i) = h_{per_t}(S_2^j) = \top$ and $h_{per_t}(S_1^k) = \bot$. Let $per_{t+1} \in PER_{t+1}$ be such that $per_{t+1}(\Gamma) = \{S_1^i, \neg S_1^k\}$. Clearly, $h_{per_t} \models per_{t+1}(\Gamma)$ for all $h_{per_t} \in G^a_{per_t}$. Since there are no $\mathcal{L}_a$-sentences in $per_{t+1}(\Gamma)$ to consider, we have shown that $\Sigma \models_1 \Gamma$.

To show that $\Gamma \models_1 \Theta$ at time $t+1$, we first fix an agent b with $KB^b_{t+1} = \Gamma$. (Note that we may have $b = a$, but this need not be the case.[13]) Choose an arbitrary $per_{t+1} \in PER_{t+1}$ and an arbitrary $per_{t+2} \in PER_{t+2}$. By definition of perception function, we have that $per_{t+2}(\Theta) = \{contra(d_{S_1}, d_{\neg S_1}, c_{t+1})\}$. Since $\{S_1, \neg S_1\} \subseteq KB^b_{t+1}$, we have, again by definition, that $H^b_{t+2} \models per_{t+2}(\Theta)$. As there are no $\mathcal{L}_w$-sentences in Θ to consider, we have shown that $\Gamma \models_1 \Theta$.

We have found a Γ such that $\Sigma \models_1 \Gamma$ and $\Gamma \models_1 \Theta$ so we conclude that $\Sigma \models_2 \Theta$, and hence $\Sigma \models_a \Theta$.

This last example demonstrated one case where n-step active consequence with $n > 1$ is applicable. Unfortunately, as we shall see in Section 3 where active consequence is discussed further, it does not get much more exciting than this.

The following theorem concerns the relation between active consequence (restricted to the sentences of $\mathcal{L}_w$) and the classical notion of logical consequence. It says that for a consistent knowledge base $KB = \Sigma$, a set Θ is an active consequence of Σ if and only if it is a logical consequence of Σ. This should be intuitively clear. A given set Σ of sentences has a fixed set of conclusions that may be drawn from it – the logical closure of Σ. We should expect the closure under active consequence to contain the closure under logical consequence since there exists a perception function which assigns the same superscript to every sentence-symbol token and thus essentially leaves the set of sentences, and its logical closure, unchanged. On the other hand, there is no perception function which extends the closure: a perception function either leaves the closure unchanged or reduces the number of sentences that can be inferred.

Remark. The last claim is crucial for the theorem to hold, but with the current definition of active consequence it is not clear whether it is actually true. The proof below depends (via Lemma 3.5) on a refined definition of active consequence which is given in Section 3.4, where this matter is discussed further.

Theorem 2.2. *Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$. If Σ is consistent, then*

$$\Sigma \models \Theta \text{ iff } \Sigma \models_1 \Theta.$$

[13] In fact, this interpretation of n-step active consequence is slightly different from the one found in [AGGP05a]. In the corresponding example it says that "Γ is potentially part of KB^a_{t+1}". But since this allows for different (potential) paths of reasoning for agent a, the result is essentially the same. See Section 3.4 for further discussion of this matter.

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}_w}$ be (classically) consistent, and let $\Theta \subseteq Sn_{\mathcal{L}_w}$ be arbitrary.

$\Rightarrow$) Assume $\Sigma \models \Theta$. This means that every $\mathcal{L}_w$-interpretation h that is a model of Σ is also a model of Θ. In particular, Θ is consistent since Σ is consistent. Consider the perception functions $per_t \in PER_t$ and $per_{t+1} \in PER_{t+1}$ which map every sentence-symbol token S_{i_k} to $S^1_{i_k}$. Note that $per_t(\Sigma)$ is consistent since Σ is consistent. Since $\Theta \subseteq Sn_{\mathcal{L}_w}$ and Θ is consistent, $per_t(\Theta) \subseteq Sn_{\mathcal{L}'_w}$ and $per_t(\Theta) = per_{t+1}(\Theta)$. Thus, we only need to show that for every model h' of $per_t(\Sigma)$, $h' \models per_t(\Theta)$.

Let h' be a model of $per_t(\Sigma)$. Consider the $\mathcal{L}_w$-interpretation $h = h' \circ per_t$, with $h(S_k) = h'(S^1_k)$ for all $k \in \mathbb{N}$. Obviously, for any consistent set $\Gamma \subseteq Sn_{\mathcal{L}_w}$, $h \models \Gamma$ if and only if $h' \models per_t(\Gamma)$. Consequently, we have that $h \models \Sigma$, and thus, by assumption, $h \models \Theta$. Hence, $h' \models per_t(\Theta)$, and we have shown that $\Sigma \models_1 \Theta$.

$\Leftarrow$) Assume $\Sigma \not\models \Theta$. This means that there exists an $\mathcal{L}_w$-interpretation h such that $h \models \Sigma$, but $h \not\models \Theta$. Let $per_t \in PER_t$ and $per_{t+1} \in PER_{t+1}$ be arbitrary.

Now, if it was the case that $\Sigma \models_1 \Theta$, then according to Lemma 3.5, for any per_{t+1} that satisfies the condition in the definition of 1-step active consequence, $per_{t+1}(\Theta) \subseteq Sn_{\mathcal{L}'_w}$. We can thus assume that $per_{t+1}(\Theta) \subseteq Sn_{\mathcal{L}'_w}$, that is, no direct contradiction in Θ is mapped to a *contra*-sentence.

Let h' be the $\mathcal{L}'_w$-interpretation satisfying $h'(S^j_i) = h(S_i)$ for all $i, j \in \mathbb{N}$. Then clearly $h' \models per_t(\Sigma)$, but $h' \not\models per_{t+1}(\Theta)$, and we have shown that $\Sigma \not\models_1 \Theta$. □

The theorem may be extended to general active consequence[14] since, as shall be proven later, when the consistent premises and the conclusions are $\mathcal{L}_w$-sentences, active consequence may be identified with 1-step active consequence (see Theorem 3.10).

Corollary. *Let* $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$. *If* Σ *is consistent, then*

$$\Sigma \models \Theta \text{ } \textit{iff} \text{ } \Sigma \models_a \Theta.$$

[14]This is claimed to be proved already in [AGGP05a] (and [AGGP05b]) but the proof does not hold for at least three reasons. First, the claim that $(\forall h \models \Sigma)\forall h'[h' \models per_t(\Sigma) \iff (\forall S_k \in A)(h(S^1_k) = h(S_k))]$ is not true: the sets $\Sigma = \Theta = \{S_1 \rightarrow S_2\}$ and interpretations h and h', with $h(S_1) = h'(S^1_1) = h'(S^1_2) = \top$ and $h(S_2) = \bot$ constitute a counterexample. (This has been corrected in [AGGP05b].)

Secondly, the if-part does not cover the case where Θ contains a direct contradiction and $per_{t+1}(\Theta) \not\subseteq Sn_{\mathcal{L}_w}$.

Thirdly, the proof can only be used to make claims about 1-step active consequence since it does not consider the recursive definition of n-step active consequence in the if-part of the proof. In particular, the proof of Theorem 2.2 above is based on the ideas found in this proof.

Finally, we note that a corresponding theorem for $\mathcal{L}_a$-sentences does not exist. For instance, $\{bel(e_{S_1}, c_t)\}$ follows actively, but not logically, from $\{S_1\}$ at time t.

2.8 Sound and Unsound Inference Rules

Using active consequence, a notion of soundness may be defined.

Definition 2.24 (Active-Sound Inference). An *active-sound* (*a-sound*) *inference* is one in which the conclusion is an active consequence of the premises.

It follows immediately from the definitions of active consequence, perception function and $\mathcal{L}_a$-structure that the following inference rules are active sound.

Definition 2.25 (Timing Rule). The *timing rule*[15] is defined by

$$\frac{t :}{t+1 : now(c_{t+1})}.$$

Definition 2.26 (Direct Contradiction Rule). The *direct contradiction rule* is defined by

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

where $\varphi \in Sn_{\mathcal{L}_w}$.

Definition 2.27 (Introspection Rule). The *introspection rule* is defined by

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

where $\varphi \in Sn_{\mathcal{L}}$.

Definition 2.28 (Negative Introspection Rule). The *negative introspection rule* is defined by

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

where $\varphi \in Sn_{\mathcal{L}}$ and $\varphi \notin KB_t^a$.

Theorem 2.3. *The timing, direct contradiction, introspection and negative introspection rules are all active sound.*

[15] The corresponding rule in [AGGP05a] has $now(c_t)$ as a premise (i.e. it requires $now(c_t)$ to be in the knowledge base at time t). This implies an unnecessary requirement on every agent that reasons using the timing rule to have $now(c_0)$ as an innate belief at time 0.

Proof. We show that the introspection rule is a-sound. The active soundness of the other rules follow immediately from definition in a similar way.

We show that $\{\varphi\} \models_1 \{bel(e_\varphi, c_t)\}$ at time t. This holds since by Definition 2.17 of perception function, $per_{t+1}(\{bel(e_\varphi, c_t)\}) = \{bel(e_\varphi, c_t)\}$ for every $per_{t+1} \in PER_{t+1}$, and by Definition 2.9 of $\mathcal{L}_a$-structure, $H^a_{t+1} \models \{bel(e_\varphi, c_t)\}$ for every agent a with $KB^a_t = \{\varphi\}$. □

Active versions of classical inference rules, such as modus ponens, may also be defined.

Definition 2.29 (Active Modus Ponens). The *active modus ponens rule* is defined by

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

where $\varphi, \psi \in Sn_{\mathcal{L}_w}$.

Theorem 2.4. *The active modus ponens rule is active sound.*

Proof. Let $\varphi, \psi \in Sn_{\mathcal{L}_w}$. Assume without restriction that the string representation of φ has i sentence-symbol tokens. We show that $\Theta = \{\psi\}$ follows actively from $\Sigma = \{\varphi, \varphi \to \psi\}$.

Let per_t be the perception function in PER_t determined by the sequence $\langle 1, 2, \ldots, i, 1, 2, \ldots \rangle$. Then $per_t(\Sigma) = \{\varphi^u, \varphi^u \to \psi^u\}$, where every sentence-symbol token in $\varphi^u \to \psi^u$ is unique. Since we can choose the truth value of every atomic formula (in $\varphi^u \to \psi^u$) independently, there exists an $\mathcal{L}'_w$-interpretation h such that $h \models per_t(\Sigma)$ (i.e. $per_t(\Sigma)$ is consistent). In particular, every model h_{per_t} of $per_t(\Sigma)$ must have $h_{per_t}(\varphi^u) = h(\psi^u) = \top$. Let $per_{t+1} \in PER_{t+1}$ be such that $per_{t+1}(\Theta) = \{\psi^u\}$. Then for every model h_{per_t} of $per_t(\Sigma)$, $h_{per_t} \models per_{t+1}(\Theta)$ as well, and we have shown that $\Sigma \models_1 \Theta$. □

In [AGGP05a] (and [AGGP05b]) there is an example of a classically sound inference rule that is claimed to be active unsound – a rule called the explosive rule.

Definition 2.30 (Explosive Rule). The *explosive rule* is defined by

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

where $\varphi, \psi \in Sn_{\mathcal{L}_w}$.

Unfortunately, the proof of the unsoundness of the explosive rule in the original paper does not hold.[16] As we shall see later, no such proof even

[16]The proof in [AGGP05a] (and [AGGP05b]) is not valid for at least two reasons. First, it is obvious that $\{\varphi, \neg\varphi\} \models_1 \{\varphi\}$ for every $\varphi \in Sn_{\mathcal{L}_w}$. Secondly, and more importantly, the proof only considers 1-step active consequence and thus cannot be directly used to make claims about general active consequence. In particular, proving that something does not follow in one step does not prove that it does not follow in $n > 1$ steps.

exists – the explosive rule *is* active sound. We will return to this matter and its dramatic implications for the theory in Section 4.

3 Analysis and Extensions

In this section the semantics found in [AGGP05a] (and [AGGP05b]) is analysed further, problems are pointed out, and extensions are suggested. In particular, we show that there are sentences for which it cannot be determined whether they follow actively from a given theory Σ, and we suggest a refined definition of active consequence which does not suffer from this problem that we call Σ-undeterminism. With this refined definition we are able to prove the fundamental equivalence between active consequence and classical logical consequence (see Theorem 2.2). Several useful metatheorems concerning n-step active consequence are also presented.

3.1 The $\mathcal{L}_a$-Semantics

When defining the $\mathcal{L}_a$-structure

$$H_t^a = \big\langle Sn_{\mathcal{L}_w}, \mathbb{N}, Sn_{\mathcal{L}}, \langle \mathbf{c}_k \rangle_{k\in\mathbb{N}}, \langle \mathbf{d}_\varphi \rangle_{\varphi\in Sn_{\mathcal{L}_w}}, \langle \mathbf{e}_\varphi \rangle_{\varphi\in Sn_{\mathcal{L}}}, \mathbf{now}, \mathbf{bel}, \mathbf{contra} \big\rangle,$$

at time t, we noted that the structure implicitly contains the sequence $\langle KB_k^a \rangle_{k=0}^t$ via the relations **bel** and **contra**. These relations store the agent's complete history of reasoning and a record of every direct contradiction that has ever occurred, respectively. In a sense, the $\mathcal{L}_a$-semantics can be summarised as follows: A sentence $\varphi \in Sn_{\mathcal{L}_a}$ is true if and only if what φ expresses *is* or *was* (if $\varphi \neq now(c_k)$) the case.

The following definitions will be useful in later discussions.

Definition 3.1 (t-Sentence). We say that a sentence $\varphi \in Sn_{\mathcal{L}_a}$ *is about time* t, or a that it is a *t-sentence*, if either of the following holds

1. $\varphi = now(c_t)$,
2. $\varphi = contra(d_\psi, d_\theta, c_t)$, where $\psi, \theta \in Sn_{\mathcal{L}_w}$,
3. $\varphi = bel(e_\psi, c_t)$, where $\psi \in Sn_{\mathcal{L}}$, or
4. $\varphi = \neg\psi$ and $\psi \in Sn_{\mathcal{L}_a}$ is about time t.

Definition 3.2 (Affirmative Sentence). A sentence $\varphi \in Sn_{\mathcal{L}}$ is *affirmative* if

1. φ is an atom, or
2. $\varphi = \neg\neg\psi$ and ψ is an affirmative sentence.

A *negative sentence* is an $\mathcal{L}$-sentence that is not affirmative.

We also note that the semantics will not allow for metareasoning about future beliefs since every affirmative $\mathcal{L}_a$-sentence about time $k > t$ is, by definition, false in H_t^a as none of the **now**-, **contra**-, and **bel**-relations in the $\mathcal{L}_a$-structure has any element with time component $k > t$ (see Definition 2.8 and Definition 2.9). Consider, for example, the sentence $\varphi = now(c_k) \rightarrow S_1$, which says that S_1 will be the case at time $k > t$ (e.g. the sun will set at 9 p.m.). In fact, due to our heavily restricted language $\mathcal{L}$, φ is not even an $\mathcal{L}$-sentence (see below), but imagine it will be as part of a future extension. Now, consider an agent a that believes (has concluded) φ at time t. Then, if it reasons using modus ponens and a clock rule (and if φ is not retracted), it will conclude S_1 at time k.

$$\begin{aligned} t: &\quad now(c_k) \rightarrow S_1 \\ &\quad \cdots \\ k: &\quad now(c_k) \rightarrow S_1,\ now(c_k) \\ k+1: &\quad S_1 \end{aligned}$$

Hence, we could argue that it would be rational for the agent to believe the sentence $bel(e_{S_1}, c_{k+1})$ at time t, and thus it could seem appropriate to assert the truth of this sentence (in H_t^a) as well. But to the contrary, $H_t^a \models \neg bel(e_{S_1}, c_{k+1})$, that is, it is true (in H_t^a) that the agent will *not* believe S_1 at time $k+1$, although we have good reason to believe it will (and no apparent reason not to do so). Of course, one cannot (usually) be absolutely certain about the future, and assuming that the knowledge base is empty for every future time point may thus be the best default even though it is (usually) highly unlikely. We again stress, however, that this is not a problem for the current version of the logic since (this kind of) metareasoning is not even possible, but it might need to be taken under consideration in future work. We formulate the underlying observation as a theorem for future use.

Theorem 3.1. *Let φ be an affirmative sentence of $\mathcal{L}_a$, and let a be an agent with $\mathcal{L}_a$-structure H_t^a at time t. If φ is about time $k > t$, then $H_t^a \not\models \varphi$, that is, φ is false in H_t^a.*

3.2 The Expressiveness of $\mathcal{L}$

The definition of the language $\mathcal{L}$ in Section 2.2 is our interpretation of how $\mathcal{L}$ was defined in [AGGP05a]. In the original paper, it says that $\mathcal{L}$ is a first-order language that is defined in two parts – $\mathcal{L}_w$ and $\mathcal{L}_a$ – and in the definition of $\mathcal{L}_a$ we read that

"... $\mathcal{S}_3$ is the sort of sentences in the language $\mathcal{L} = \mathcal{L}_a \cup \mathcal{L}_w$".

Now, how is union between languages supposed to be interpreted? If we were discussing formal languages in general and identified a language with

a set of strings over an alphabet, then everything would be fine. But we are not.[17] Instead, we identify a language with its symbols and grammar (i.e. rules for how formulae of the language are formed). (In fact, nothing is said about the grammars of $\mathcal{L}_w$ and $\mathcal{L}_a$ in [AGGP05a] so we assume that standard propositional and first-order syntax, respectively, were intended.) It is apparent that $\mathcal{L}$ has to be interpreted as being a language consisting of the symbols of the languages $\mathcal{L}_w$ and $\mathcal{L}_a$. But what about the grammar of $\mathcal{L}$? Nothing is explicitly said about this in [AGGP05a], but we mean that the only reasonable interpretation is that the set of atomic formulae of $\mathcal{L}$ is the union of the atomic formulae of $\mathcal{L}_w$ and $\mathcal{L}_a$, respectively. For the compound formulae, there are two plausible interpretations:

1. The formulae are combined using the connectives of $\mathcal{L}$ and standard first-order syntax (just as we did with $\mathcal{L}_w$ and $\mathcal{L}_a$).

2. The formulae and connectives are considered to be sorted and we may only combine formulae and connectives of the same sort (using standard first-order syntax).

We decided on the second interpretation for our definition of $\mathcal{L}$ (and $\mathcal{L}'$). The reason was that this implies that the set of sentences of $\mathcal{L}$, $Sn_{\mathcal{L}}$, is the union of the two disjoint sets $Sn_{\mathcal{L}_w}$ and $Sn_{\mathcal{L}_a}$ (and similarly, $Sn'_{\mathcal{L}} = Sn_{\mathcal{L}'_w} \cup Sn_{\mathcal{L}_a}$). This is in conformity with the second sentence in the definition of $\mathcal{L}'$ in [AGGP05a] which reads:

"Let $\mathcal{L}' = \mathcal{L}'_w \cup \mathcal{L}_a$. Let $Sn_{\mathcal{L}'} = Sn_{\mathcal{L}'_w} \cup Sn_{\mathcal{L}_a}$."

Had we instead defined $\mathcal{L}$ according to the first interpretation then the union $Sn_{\mathcal{L}_w} \cup Sn_{\mathcal{L}_a}$ would have been a proper subset of $Sn_{\mathcal{L}}$. That is, the set $\Delta = Sn_{\mathcal{L}} - Sn_{\mathcal{L}_w} - Sn_{\mathcal{L}_a}$ would have been non-empty. We would then (in $\mathcal{L}$) have been able to express sentences such as $now(c_{21}) \rightarrow \neg S_1$ and $(now(c_3) \wedge S_2) \rightarrow bel(e_{S_2}, c_3)$.[18] This may seem fine, since the ability to reason in terms of *"If it is 9 pm, then the sun is set."* (default reasoning) and *"If it is 3 am and it is dark, then I believe that its is dark at 3 am."* may be desirable. Indeed, what is the purpose of doing metareasoning if you cannot relate it to the external world?[19]

Unfortunately, this is not an acceptable alternative since anything in Δ follows actively from anything. To see why, consider the sets $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq \Delta$. Let per_t and per_{t+1} be arbitrary perception functions at time t and $t+1$, respectively. Now, for every $\mathcal{L}_a$-structure H^a_{t+1} and pt-structure $M^a_{per_t}$ of Σ, H^a_{t+1} and $M^a_{per_t}$ satisfies vacuously $per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a}$

[17] Even if we were, the (lack of) expressiveness of $\mathcal{L}$ discussed below is still a problem.

[18] We allow ourselves to use conjunction as syntactic sugar.

[19] In earlier work on active logic (e.g. [GKP00]), the world language has been a part of the agent's metalanguage.

and $per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w}$, respectively, since, by Definition 2.17, $per_{t+1}(\Theta) = \Theta$. Thus, $\Sigma \models_1 \Theta$. To avoid this, either $Sn_{\mathcal{L}}$ needs to be restricted in such a way that $\Delta = \emptyset$, or the concept of active consequence (and other related concepts) needs to be redefined. We assume that the former solution is the intended one.[20]

But as noted above, restricting the syntax in such a way will render $\mathcal{L}$, more or less, uninteresting as metareasoning cannot be related to the external word. Furthermore, with the additional restrictions on $\mathcal{L}_a$ – only one connective and no quantifiers or variables – we seem to be left with a metareasoning incapable of much more than a mere record keeping of beliefs and contradictions.

3.3 Existence of a Model

As noted at the end of Section 2.6, the existence of a pt-structure for a given perception function is not guaranteed. A pt-structure $M^a_{per_t} = \langle h_{per_t}, H^a_t \rangle$ contains an $\mathcal{L}'_w$-interpretation h_{per_t}, which, by definition, models the agents perception of its $\mathcal{L}_w$-sentences ($W^a_{per_t}$). But there exist perception functions which may render inconsistent perceived knowledge bases. Take, for example, the perception function per^i_t, which assigns superscript 1 to every sentence-symbol token and thus, in a sense, leaves the set unchanged. When per^i_t is applied to a set Σ that contains an indirect inconsistency, the perceived set will be inconsistent as well (i.e it has no model). Then, the set $\mathbb{M}^a_{per^i_t}$ of pt-structures is empty and the condition

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

in Definition 2.21 holds vacuously for every conclusion $\Theta \subseteq Sn_{\mathcal{L}}$. (Note that Θ follows even if $H^a_{t+1} \not\models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a})$.) That is, with the current definition of active consequence, we may conclude that everything follows actively from an indirectly inconsistent knowledge base, contrary to our intention to limit the logical closure of inconsistent sets.

Of course, there always exists *a* perception function per_t such that the perceived knowledge base $KB^a_{per_t}$ is t-strongly consistent, for instance the one used in Theorem 2.1. Thus we could postulate that an agent always perceives its knowledge base in such a way that it appears consistent. Active consequence could then be redefined as follows.

Definition 3.3 (1-Step Active Consequence). Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$, and let a be an agent with $KB^a_t = \Sigma$. We say that Θ is a *1-step active conse-*

[20] Our interpretation has been confirmed by the authors in [AGGP05b].

quence of Σ at time t, written $\Sigma \models_1 \Theta$, if and only if

$$(\exists per_t \in PER_t)(\exists per_{t+1} \in PER_{t+1})$$
$$\big[\mathbb{M}^a_{per_t} \neq \emptyset \ \wedge \ H^a_{t+1} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a}) \wedge$$
$$(\forall M^a_{per_t} \in \mathbb{M}^a_{per_t})(M^a_{per_t} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w})\big].$$

By adding a condition on $\mathbb{M}^a_{per_t}$ (and thus indirectly on per_t), the problem pointed out above is avoided.[21] Note also that our previous results that involve active consequence (i.e. the theorems and examples in Section 2.7 and Section 2.8) are still valid since we have argued using perception functions that have rendered consistent perceived knowledge bases.

Of course, one may raise philosophical objections to this solution: It is indeed possible to actually believe that, for example, $\{S_1, S_1 \rightarrow S_2, \neg S_2\}$ is the case while still maintaining that the two instances of S_2 name the same proposition. In fact, human beings do this all the time – we have inconsistent belief sets without being aware of the inconsistencies simply because we have not yet drawn the appropriate conclusions.

3.4 Agent a and 1-Step Active Consequence

Consider two sets $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$ such that $\Sigma \models_1 \Theta$ at time t. By Definition 3.3, we know that $\Sigma = KB^a_t$ for some agent a, that there exist perception functions $per_t \in PER_t$ and $per_{t+1} \in PER_{t+1}$ such that $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ is consistent, and that

1. $H^a_{t+1} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a})$, and
2. $(\forall h_{per_t} \in G^a_{per_t})(h_{per_t} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w}))$.

Assume we want to verify that Θ actually follows actively in one step from Σ given the two perception functions. We then need to verify the conditions on the perception of Θ. The second condition, that every model of $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ is also a model of $per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w}$, is straightforward and easily verified. But what about condition one? In order to verify if an arbitrary $\mathcal{L}_a$-sentence is satisfied by H^a_{t+1}, we need to know the complete history of the agent a up until time $t+1$, that is, the sequence $\langle KB^a_k \rangle^{t+1}_{k=0}$, or the derived **bel**- and **contra**-relations, that is, the structure H^a_{t+1} itself.[22] But all we have is a single snapshot of the knowledge base at time t, namely

[21] This restriction on per_t is most likely what the authors intended since in a remark after the definition of $W^a_{per_t}$, they claim that this set is consistent. But, as mentioned before, as their definition stands in [AGGP05a], this is not true. For clarity, we have chosen to restrict per_t (via $\mathbb{M}^a_{per_t}$) in the definition of active consequence rather than in the definition of $W^a_{per_t}$. The problem has been acknowledged in [AGGP05b], where instead G_{per_t} has been redefined.

[22] Notice the apparent paradox in referring to KB^a_{t+1} when trying to determine whether something follows from KB^a_t since if it does it may end up in KB^a_{t+1}.

$\Sigma = KB_t^a$, and we are thus only *certain* to be able to determine the truth values of sentences about time t besides those of sentences about time $k > t+1$ (which follow from definition, see Theorem 3.1). Indeed, Σ can also contain sentences about time $k < t$, and if it does, we can, since Σ is assumed t-weakly consistent, conclude that these sentences are modelled by H_{t+1}^a. If we have been able to assert the truth value of an $\mathcal{L}_a$-sentence φ (in H_{t+1}^a), it is also possible to derive the truth values of some related sentences. If the true, affirmative *bel*-sentence φ is about another *bel*- or *contra*-sentence, we know, again due to the assumption of t-weak consistency, that this sentence is also true, for example,

$$H_{t+1}^a \models bel(e_{bel(e_\psi,c_l)}, c_k) \Rightarrow H_{t+1}^a \models bel(e_\psi, c_l),$$

for every $\psi \in Sn_{\mathcal{L}}$ and time points k and l. Furthermore, since $\mathcal{L}_a$ has negation as its only connective, we can of course also derive the truth value of every subformula of φ and of every sentence in which φ is a subformula. Except for some special cases (e.g. *now*-sentences), this is about all we can do – the truth value of every other $\mathcal{L}_a$-sentence is undeterminable. We formulate this discussion as a theorem after first defining the concept of Σ-determinism.

Definition 3.4 (Σ-Determinism). Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-weakly consistent, and let a be an agent with $KB_t^a = \Sigma$. A sentence $\varphi \in Sn_{\mathcal{L}_a}$ is *determinable* from Σ (or Σ-*determinable*) at time $t+1$ if we can determine whether $H_{t+1}^a \models \varphi$ without additional information about a.

If φ is not determinable from Σ at time $t+1$, then it is Σ-*undeterminable* at time $t+1$. Denote the set of Σ-undeterminable sentences at time $t+1$ with $UND_{t+1}(\Sigma)$.

Remark. We will sometimes allow ourselves to drop the "at time $t+1$" specifier when it is clear from context.

Theorem 3.2. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$, and let a be an agent with $KB_t^a = \Sigma$. For any sentence $\varphi \in Sn_{\mathcal{L}_a}$, φ is Σ-determinable at time $t+1$ if and only if one of the following conditions holds:*

1. $\varphi = now(c_k)$,
2. $\varphi = contra(d_\psi, d_\theta, c_k)$ *but* $\psi \neq \neg\theta$ *and* $\theta \neq \neg\psi$,
3. φ *is about time* t *or time* $k > t+1$,
4. $\varphi \in \Sigma$ *and* φ *is about time* $k < t$,
5. $\varphi = \neg\psi$ *and* ψ *is* Σ-*determinable*,
6. $\psi = \neg\varphi$ *and* ψ *is* Σ-*determinable*,

7. $\varphi = contra(d_\psi, d_\theta, c_k)$ *and* $contra(d_\theta, d_\psi, c_k)$ *is* Σ*-determinable,*

8. $\varphi = contra(d_\psi, d_\theta, c_k)$ *and both* $bel(e_\psi, c_k)$ *and* $bel(e_\theta, c_k)$ *are* Σ*-determinable,*

9. $\varphi = contra(d_\psi, d_\theta, c_k)$, $bel(e_\psi, c_k)$ *or* $bel(e_\theta, c_k)$ *is* Σ*-determinable, and* $H^a_{t+1} \not\models bel(e_\psi, c_k)$ *or* $H^a_{t+1} \not\models bel(e_\theta, c_k)$*, respectively,*

10. $\varphi = bel(e_\psi, c_k)$ *for some* $\psi \in Sn_{\mathcal{L}_w}$*, and* $contra(d_\psi, d_\theta, c_k)$ *is* Σ*-determinable and* $H^a_{t+1} \models contra(d_\psi, d_\theta, c_k)$ *for some* $\theta \in Sn_{\mathcal{L}_w}$*,*

11. $\varphi = bel(e_\psi, c_k)$ *and* $bel(e_\varphi, c_l)$ *is* Σ*-determinable and* $H^a_{t+1} \models bel(e_\varphi, c_l)$ *for some time* l*, or*

12. $\varphi = contra(d_\psi, d_\theta, c_k)$ *and* $bel(e_\varphi, c_l)$ *is* Σ*-determinable and* $H^a_{t+1} \models bel(e_\varphi, c_l)$ *for some time* l*.*

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$, let a be an agent with $KB^a_t = \Sigma$, and let $\varphi \in Sn_{\mathcal{L}_a}$.

1. From Definition 2.8 and Definition 2.9 it follows that we can determine the truth value of every *now*-sentence given the time $(t + 1)$.

2. From the same definitions, it is also apparent that *contra*-sentences are never satisfied unless the involved sentences are direct contradictions of each other.

3. We have already handled *now*-sentences so assume φ is not a *now*-sentence. If φ is about time t then its truth value in H^a_{t+1} is determinable from $\Sigma = KB^a_t$, and if φ is about time $k > t+1$ then its truth value in H^a_{t+1} follows from Theorem 3.1 and Definition 2.9.

4. Let $\varphi \in \Sigma$ be about time $k < t$. Then $H^a_{t+1} \models \varphi$ since Σ is assumed t-weakly consistent and what φ expresses (e.g. the presence of a contradiction) is or was indeed the case according to Definition 2.8.

5. According to Definition 2.9, $H^a_{t+1} \models \neg\psi$ if and only if $H^a_{t+1} \not\models \psi$.

6. Similarly, $H^a_{t+1} \not\models \varphi$ if and only if $H^a_{t+1} \models \neg\varphi$.

7. Follows by symmetry from Definition 2.8.

8. Assume that $\psi = \neg\theta$ ($\theta = \neg\psi$ follows by symmetry, and the other possibilities were handled in case 2 above). Now, again by Definition 2.8 and Definition 2.9: If we know for each of the sentences ψ and θ whether it was in KB^a_k or not, then we know whether $H^a_{t+1} \models contra(d_\psi, d_\theta, c_k)$.

9. If $H^a_{t+1} \not\models bel(e_\psi, c_k)$ or $H^a_{t+1} \not\models bel(e_\theta, c_k)$, then ψ or θ was not part of KB^a_k, and hence $H^a_{t+1} \not\models contra(d_\psi, d_\theta, c_k)$.

10. Similarly, if $H^a_{t+1} \models contra(d_\psi, d_\theta, c_k)$, then both ψ and θ were in KB^a_k, and thus $H^a_{t+1} \models bel(e_\psi, c_k)$.

11. If $H^a_{t+1} \models bel(e_\varphi, c_l)$, then $\varphi \in KB^a_l$. Since KB^a_l is assumed t-weakly consistent and $\varphi \in Sn_{\mathcal{L}_a}$, $H^a_l \models bel(e_\psi, c_k)$. Consequently, we know that $\psi \in KB^a_k$, and thus also $H^a_{t+1} \models bel(e_\psi, c_k)$.

12. If $H^a_{t+1} \models bel(e_\varphi, c_l)$, then $\varphi \in KB^a_l$. Since KB^a_l is assumed t-weakly consistent and $\varphi \in Sn_{\mathcal{L}_a}$, $H^a_l \models contra(d_\psi, d_\theta, c_k)$. Consequently, we know that $\psi, \theta \in KB^a_k$ and that $\psi = \neg\theta$ or $\theta = \neg\psi$, and thus also $H^a_{t+1} \models contra(d_\psi, d_\theta, c_k)$.

We now argue that any other sentence is not determinable from Σ. For any sentence $\varphi \in Sn_{\mathcal{L}_a}$ not already handled above, we have that it is about time $k < t$ or time $k = t + 1$, it is not a *now*-sentence, and it is not the object of any Σ-determinable, affirmative, true *bel*-sentence. We also know that φ is not in Σ if $k < t$. Assume without restriction that φ is atomic.[23] Then φ is of one of the following forms:

1. $\varphi = contra(d_\psi, d_\theta, c_k)$, $\psi = \neg\theta$ or $\theta = \neg\psi$, and either
   (a) neither $bel(e_\psi, c_k)$ nor $bel(e_\theta, c_k)$ is Σ-determinable, or
   (b) either $bel(e_\psi, c_k)$ or $bel(e_\theta, c_k)$ is Σ-determinable, and H^a_{t+1} satisfies $bel(e_\psi, c_k)$ or $bel(e_\theta, c_k)$, respectively.

   In the first case we cannot tell whether $\{\psi, \theta\} \subseteq KB^a_k$, and thus φ is not Σ-determinable. In the second case we have no more information than that ψ or θ was in KB^a_k, and thus φ is not Σ-determinable.

2. $\varphi = bel(e_\psi, c_k)$, and either
   (a) $\psi \in Sn_{\mathcal{L}_a}$,
   (b) $contra(d_\psi, d_\theta, c_k)$ is not Σ-determinable for any $\theta \in Sn_{\mathcal{L}_w}$, or
   (c) $H^a_{t+1} \not\models contra(d_\psi, d_\theta, c_k)$ for every $\theta \in Sn_{\mathcal{L}_w}$ for which the sentence $contra(d_\psi, d_\theta, c_k)$ is Σ-determinable.

   Since the only (remaining) possible ways of knowing that $\psi \in KB^a_k$ is through *contra*-sentences, and since these are about $\mathcal{L}_w$-sentences, we can not determine whether φ is satisfied in the first case. Case two follows since we can not determine the needed *contra*-sentences, and the third case follows since the fact that there was no direct contradiction involving ψ in KB^a_k does not rule out that ψ was in the knowledge base at that time. □

[23] Note that knowing that φ is in Σ when φ is about time $t+1$ is not enough to be able to determine whether $H^a_{t+1} \models \varphi$: Since Σ is assumed t-weakly consistent, φ must be negative (and thus non-atomic), but its truth value in $\mathcal{L}_a$-structures at time t is determined solely by its syntactical structure (see Theorem 3.1).

The following lemma sheds some more light on the concept of Σ-undeterminism. Note that in the Lemma, and in the rest of this thesis, we identify an agent with its representation. In particular, we identify an agent a with an infinite sequence of knowledge bases (i.e. subsets of $Sn_{\mathcal{L}}$), $\langle KB_k^a \rangle_{k=0}^{\infty}$.

Lemma 3.3. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-weakly consistent, and let $\varphi \in Sn_{\mathcal{L}_a}$. If φ is Σ-undeterminable at time $t+1$, then there exists an agent a with $KB_t^a = \Sigma$ but $H_{t+1}^a \not\models \varphi$.*

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$, and let $\varphi \in Sn_{\mathcal{L}_a}$ be Σ-undeterminable at time $t+1$. We assume without loss of generality that φ is atomic. Then, since *now*-sentences are obviously Σ-determinable, φ is either a *contra*-sentence, $contra(d_\psi, d_\theta, c_k)$ for some $\psi, \theta \in Sn_{\mathcal{L}_w}$ and time k, or a *bel*-sentence, $bel(e_\psi, c_k)$ for some $\psi \in Sn_{\mathcal{L}}$ and time k.

Given an arbitrary agent a with $KB_t^a = \Sigma$, we can, by definition, not tell if $H_{t+1}^a \models \varphi$ or not without additional information about a. In particular, this implies that there exists an agent a with $KB_t^a = \Sigma$ such that $H_{t+1}^a \not\models \varphi$:

We consider first the *contra*-case. Assume the contrary, that for every agent a with $KB_t^a = \Sigma$, $H_{t+1}^a \models contra(d_\psi, d_\theta, c_k)$, that is, for every such agent, $\psi, \theta \in KB_k^a$. This is obviously not true, since there are no sentences ψ and θ that have to be in KB_k^a for every such agent – their presence in KB_k^a is not determined by $KB_t^a = \Sigma$ and this is the only restriction we have on the agents.

Similarly, if $\varphi = bel(e_\psi, c_k)$, not every agent with $KB_t^a = \Sigma$ has to have $\psi \in KB_k^a$. □

Who is agent a? According to the initial assumptions in [AGGP05a], "There is only one agent a", and their definition of active consequence begins with:

> "Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$ such that $\Sigma = KB_t^a$. Then Θ is said to be a 1-step active consequence of Σ at time t, ..., if and only if ...".

The way we have interpreted this formulation is that we can choose an agent a with $KB_t^a = \Sigma$ (see Definition 3.3 above). Another, possibly far-fetched, interpretation is the literal one: We have one sole agent a, and the concept of active consequence only applies to its specific knowledge base at each given time, that is, we can only speak about active consequences of Σ at time t if $\Sigma = KB_t^a$. With an additional assumption of perfect knowledge about this unique agent, we are able to determine whether $\Sigma \models_1 \Theta$. This seem to have strange implications as, for example, if $\{S_1\} \models_1 \{S_1\}$ at time t, then $\{S_1\} \not\models_1 \{S_1\}$ at any other time k with $KB_k^a \neq \{S_1\}$ (since nothing follows actively at time k from a set $\Sigma \neq KB_k^a$). We do not pursue this interpretation further.

Note also that our definition is applied in the exact same way as is the definition in [AGGP05a]. In particular, even under the assumption of "only one agent a", the original examples starts with sentences such as "Let $\Sigma = \{\varphi, \neg\varphi\} = KB_t^a$". Instead of having a single agent whose knowledge base changes between examples, we found it more natural to change the agent.

Our interpretation is in compliance with the assumption of one agent in a weaker sense, namely that the assumption imposes a restriction on the agent's metareasoning capabilities: An agent can only reason about its own beliefs and not about beliefs of other agents. Furthermore, we are modelling one agent at a time in the sense that we are modelling single belief sets and not families of belief sets. In particular, we are studying belief sets and a binary relation of consequence between such sets.[24]

Another argument for our definition follows: In Example 2.7 we concluded that $\Theta = \{contra(d_{S_1}, d_{\neg S_1}, c_{t+1})\}$ follows actively in two steps from $\Sigma = \{S_1, S_2, S_2 \rightarrow \neg S_1\}$ by showing that for $\Gamma = \{S_1, \neg S_1\}$, $\Sigma \models_1 \Gamma$ and $\Gamma \models_1 \Theta$. In particular, when applying our definition of 1-step active consequence, we referred to two agents a and b with $KB_t^a = \Sigma$ and $KB_{t+1}^b = \Gamma$, respectively. The corresponding example in [AGGP05a] only involves one agent a with $KB_t^a = \Sigma$, and instead it is said that "Γ is *potentially* part of KB_{t+1}^a" (our emphasis). This constitutes more evidence against the literal interpretation as the example is not about a fixed individual a about which perfect knowledge is assumed. Rather, it opens for more than one potential path of reasoning for agent a, and thus we are essentially dealing with more than one agent. We consider, however, our definitions to be more transparent on this point than are their counterparts in [AGGP05a].

Before we return to our discussion of agent a, note also the use of the phrase "part of" in "Γ is potentially part of KB_{t+1}^a" above. This formulation indicates an even more relaxed definition of 1-step active consequence, a definition that could start with:

> Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$, and let a be an agent with $\Sigma \subseteq KB_t^a$.
>
> ...

Although it would imply that the set of Σ-undeterminable sentences would also contain sentences about time t, this could be a reasonable extension. We leave this discussion for future work.

As we see it, there are three possible ways to refine the definition of active consequence in order to handle the problem of Σ-undeterminism (i.e. that there are sentences for which we cannot tell whether they follow actively from a given set):

[24] In the concluding remarks in [AGGP05a] it is said the starting assumption will be dropped in future work, which will include "multiple agents, reasoning both about the world and about one another's beliefs".

1. Existential quantification of the agent. That is, we require that there exists an agent a with $KB_t^a = \Sigma$ and $H_{t+1}^a \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a})$.

2. Universal quantification of the agent. That is, we require that for every agent a with $KB_t^a = \Sigma$, $H_{t+1}^a \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a})$, or equivalently, we postulate that if $(\Theta \cap UND_{t+1}(\Sigma)) \neq \emptyset$, then $\Sigma \not\models_1 \Theta$ at time t.

3. We accept that there are sets Σ and Θ for which it cannot be determined whether $\Sigma \models_1 \Theta$ (i.e. we clarify that this is the case and leave the definition essentially unchanged).

Consider the first option, which, for instance, would make every singleton subset of $UND_{t+1}(\Sigma)$ follow actively from Σ at time t. This is hardly an acceptable solution as then, for example,

$$\emptyset \models_1 \{bel(e_\varphi, c_{t-1})\},$$

at time t for any $\varphi \in Sn_{\mathcal{L}}$. In fact, we would have $\emptyset \models_1 \{\neg bel(e_\varphi, c_{t-1})\}$ at time t as well, but still $\emptyset \not\models_1 \{bel(e_\varphi, c_{t-1}), \neg bel(e_\varphi, c_{t-1})\}$ due to the principle of bivalence – no agent a both believes and not believes φ at time point $t-1$, that is, either $\varphi \in KB_{t-1}^a$ or $\varphi \notin KB_{t-1}^a$.

The second and third options are more appealing, both ruling out the preceding example. Consider first the third option: that we sometimes cannot determine if $\Sigma \models_1 \Theta$ or not. This is a reasonable solution although it would severely limit the possibility of general results about the logic and its semantics. A variant of this option would be to allow external information about the agent when determining whether $\Sigma \models_1 \Theta$: Given enough information about an agent a, we can, even if Θ contains Σ-undeterminable sentences, determine whether $\Sigma \models_1 \Theta$ at time t *for agent* a. We would then have a situation where, given $\varphi \in Sn_{\mathcal{L}}$, sometimes $\emptyset \models_1 \{bel(e_\varphi, c_{t-1})\}$, sometimes $\emptyset \not\models_1 \{bel(e_\varphi, c_{t-1})\}$, and sometimes we simply cannot tell, depending on which agent a we are discussing *and* if we have enough information about a. We conclude that this option, whether we allow external information or not, is not satisfactory.

In particular, the fundamental equivalence of Theorem 2.2 (which is also found in [AGGP05a]) depends on us being able to determine whether $\Sigma \models_1 \Theta$ in the *general case*, that is, without referring to a specific agent about which we have perfect knowledge. The theorem states that if something in $Sn_{\mathcal{L}_w}$ follows actively in one step, then it also follows classically. But consider the sets $\Sigma = \emptyset$ and $\Theta = \{\varphi, \neg\varphi\}$ for some $\varphi \in Sn_{\mathcal{L}_w}$. Let $per_t \in PER_t$ be arbitrary, and let $per_{t+1} \in PER_{t+1}$ be such that $per_{t+1}(\Theta) = \{contra(d_\varphi, d_{\neg\varphi}, c_{t+1})\}$. Then, according to the definition of 1-step active consequence, $\Sigma \models_1 \Theta$ if $H_{t+1}^a \models per_{t+1}(\Theta)$. But $\Sigma \not\models \Theta$, so in order for the theorem to hold, we would have to have $H_{t+1}^a \not\models per_{t+1}(\Theta)$ for every agent a (even if $\varphi, \neg\varphi \in KB_{t+1}^a$). Consequently, Theorem 2.2 does not hold with original definition of active consequence.

This leaves us with the second option: that nothing in $UND_{t+1}(\Sigma)$ should follow actively from Σ at time t. The following formal definition of 1-step active consequence incorporates the idea.

Definition 3.5 (1-Step Active Consequence). Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-weakly consistent, and let $\Theta \subseteq Sn_{\mathcal{L}}$ be arbitrary. We say that Θ is a *1-step active consequence* of Σ at time t, written $\Sigma \models_1 \Theta$, if and only if

$$(\exists per_t \in PER_t)(\exists per_{t+1} \in PER_{t+1})\forall a\Big[KB_t^a = \Sigma \rightarrow \big[\mathbb{M}^a_{per_t} \neq \emptyset \ \wedge \ H^a_{t+1} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a}) \wedge (\forall M^a_{per_t} \in \mathbb{M}^a_{per_t})(M^a_{per_t} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w})\big]\Big].$$

Remark. Note that we now need to explicitly add the condition that Σ must be t-weakly consistent. Without this condition, everything would follow from a non-t-weakly consistent set since no agent can have such a set as its knowledge base. This condition was implicit in our previous definitions as we required that there was an agent a with $KB_t^a = \Sigma$.

We formulate the consequence of Definition 3.5 for Σ-undeterminable sentences as a theorem.

Theorem 3.4. *Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$. If $(\Theta \cap UND_{t+1}(\Sigma)) \neq \emptyset$, then $\Sigma \not\models_1 \Theta$ at time t.*

Proof. Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$ be such that $(\Theta \cap UND_{t+1}(\Sigma)) \neq \emptyset$, and let $\varphi \in \Theta$ be a Σ-undeterminable sentence at time $t+1$. Let per_t and per_{t+1} be arbitrary perception functions at time t and $t+1$, respectively. We note that $\varphi \in Sn_{\mathcal{L}_a}$ and that therefore $\varphi \in per_{t+1}(\Theta)$.

By Definition 3.5, $\Sigma \not\models_1 \Theta$ if Σ is not t-weakly consistent so assume Σ is t-weakly consistent. Then, by Lemma 3.3, we know that there exists an agent a with $KB_t^a = \Sigma$ but $H^a_{t+1} \not\models \varphi$. Hence, $\Sigma \not\models_1 \Theta$, by Definition 3.5. □

Note that if $\varphi \in UND_{t+1}(\Sigma)$, then neither φ nor $\neg\varphi$ follows actively in one step from Σ at time t.

As mentioned above, Theorem 2.2 depends on a refined definition of 1-step active consequence. In particular, the proof makes use of the following lemma, which states that the perception function per_{t+1} in the definition of 1-step active consequence does not map any direct contradictions in Θ to *contra*-sentences when $\Theta \subseteq Sn_{\mathcal{L}_w}$.

Lemma 3.5. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$. If $\Sigma \models_1 \Theta$ at time t, then $(per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a}) = \emptyset$ for every perception function $per_{t+1} \in PER_{t+1}$ such that*

$$(\exists per_t \in PER_t)\forall a\Big[KB_t^a = \Sigma \rightarrow \big[\mathbb{M}^a_{per_t} \neq \emptyset \ \wedge \ H^a_{t+1} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a}) \wedge (\forall M^a_{per_t} \in \mathbb{M}^a_{per_t})(M^a_{per_t} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w})\big]\Big].$$

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$ be such that $\Sigma \models_1 \Theta$ at time t. Let per_t and per_{t+1} be perception functions at time t and $t+1$, respectively, that satisfy the condition in Definition 3.5. In particular, $H^a_{t+1} \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a})$ for every agent a with $KB^a_t = \Sigma$.

Note that since $\Theta \subseteq Sn_{\mathcal{L}_w}$, the only $\mathcal{L}_a$-sentences that may be in the set $per_{t+1}(\Theta)$ are *contra*-sentences about time $t+1$ originating from direct contradictions in Θ. Now, either these sentences are Σ-undeterminable at time $t+1$, or they are not. We show that both cases lead to contradictions and hence that $per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a} = \emptyset$.

Assume $\varphi = contra(d_\psi, d_{\neg\psi}, c_{t+1})$ is in $per_{t+1}(\Theta)$ for some $\psi \in Sn_{\mathcal{L}_w}$. Now, if φ is Σ-undeterminable, we know from Lemma 3.3 that there exists an agent a with $KB^a_t = \Sigma$, but $H^a_{t+1} \not\models \varphi$. Thus $H^a_{t+1} \not\models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a})$, which is a contradiction.

If φ is Σ-determinable, then it is true in H^a_{t+1} for every agent a with $KB^a_t = \Sigma$ because otherwise we would have $H^a_{t+1} \not\models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a})$, contrary to our assumption. Now, for every such agent a with $KB^a_t = \Sigma$ and $H^a_{t+1} \models \varphi$, there exists an agent a' with the same reasoning history $\langle KB^{a'}_k \rangle^t_{k=0}$, but with $KB^{a'}_{t+1} = \emptyset$. In particular we have that $KB^{a'}_t = \Sigma$, but $H^{a'}_{t+1} \not\models \varphi$ (since $\psi \notin KB^{a'}_{t+1}$), contrary to our assumption. □

We are now able to prove the following theorem, which states that only the $\mathcal{L}_w$-sentences of Σ are involved when determining whether $\Theta \subseteq Sn_{\mathcal{L}_w}$ follows actively in one step from Σ. This result will be generalised to n-step active consequence in the next section.

Theorem 3.6. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$. Then*

$$\Sigma \models_1 \Theta \text{ iff } (\Sigma \cap Sn_{\mathcal{L}_w}) \models_1 \Theta.$$

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$.

First assume $\Sigma \models_1 \Theta$. Since $\Theta \subseteq Sn_{\mathcal{L}_w}$, we have from the previous lemma that there exist perception functions per_t and per_{t+1} at time t and $t+1$, respectively, such that $per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a} = \emptyset$, $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ is consistent, and $(per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}) \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w})$.

Since $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w} = per_t(\Sigma \cap Sn_{\mathcal{L}_w}) \cap Sn_{\mathcal{L}'_w}$ (for every perception function per_t), we have that $per_t(\Sigma \cap Sn_{\mathcal{L}_w}) \cap Sn_{\mathcal{L}'_w}$ is consistent, and $(per_t(\Sigma \cap Sn_{\mathcal{L}_w}) \cap Sn_{\mathcal{L}'_w}) \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w})$ as well. That is, $(\Sigma \cap Sn_{\mathcal{L}_w}) \models_1 \Theta$.

Now, assume $(\Sigma \cap Sn_{\mathcal{L}_w}) \models_1 \Theta$. Again by Lemma 3.5, we have that there exist perception functions per_t and per_{t+1} at time t and $t+1$, respectively, such that $per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a} = \emptyset$, $per_t(\Sigma \cap Sn_{\mathcal{L}_w}) \cap Sn_{\mathcal{L}'_w}$ is consistent, and $(per_t(\Sigma \cap Sn_{\mathcal{L}_w}) \cap Sn_{\mathcal{L}'_w}) \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w})$. The identity $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w} = per_t(\Sigma \cap Sn_{\mathcal{L}_w}) \cap Sn_{\mathcal{L}'_w}$ holds also for per_t, and thus $\Sigma \models_1 \Theta$ as well. □

We conclude this section with the observation that all our previous results that were based on the original definition of active consequence are still valid with the refined definition since they do not involve Σ-undeterminable sentences. In particular, Lemma 3.5 makes the new definition of 1-step active consequence equivalent with Definition 3.3 when the consequent is in $Sn_{\mathcal{L}_w}$, and the old definition could then be used in practice.

3.5 n-Step Active Consequence

In this section we prove several useful results regarding n-step active consequence. In particular, we show that subsets of $Sn_{\mathcal{L}_w}$ follow actively from t-strongly consistent sets if and only if they follow actively in one step.

According to the following lemma, we can when $\Theta \subseteq Sn_{\mathcal{L}_w}$ without restriction assume that the set Γ in the definition of n-step active consequence is in $Sn_{\mathcal{L}_w}$ as well.

Lemma 3.7. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$. If $\Sigma \models_n \Theta$, $n > 1$, then*

$$(\exists \Gamma' \subseteq Sn_{\mathcal{L}_w})(\Sigma \models_{n-1} \Gamma' \wedge \Gamma' \models_1 \Theta).$$

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$ be such that $\Sigma \models_n \Theta$, $n > 1$. Then, by definition,

$$(\exists \Gamma \subseteq Sn_{\mathcal{L}})(\Sigma \models_{n-1} \Gamma \wedge \Gamma \models_1 \Theta).$$

Let $\Gamma' = \Gamma \cap Sn_{\mathcal{L}_w}$. Now, if $\Gamma \models_1 \Theta$, then because $\Theta \subseteq Sn_{\mathcal{L}_w}$, $\Gamma' \models_1 \Theta$ as well according to Theorem 3.6. Furthermore, if $\Sigma \models_{n-1} \Gamma$, then obviously $\Sigma \models_{n-1} \Gamma'$ since $\Gamma' \subseteq \Gamma$. □

The preceding lemma is used to show that only the $\mathcal{L}_w$-sentences of the premises are involved in determining whether $\mathcal{L}_w$-sentences follow actively, thereby generalising Theorem 3.6.

Theorem 3.8. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$. Then for every $n \geq 1$,*

$$\Sigma \models_n \Theta \textit{ iff } (\Sigma \cap Sn_{\mathcal{L}_w}) \models_n \Theta.$$

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$. The theorem holds for $n = 1$ according to Theorem 3.6, so let $n > 1$.

$\Rightarrow$) Assume $\Sigma \models_n \Theta$. From Lemma 3.7 it follows that there exist sets $\Gamma_1, \Gamma_2, \ldots, \Gamma_{n-1} \subseteq Sn_{\mathcal{L}_w}$ such that

$$\Sigma \models_1 \Gamma_1 \wedge \Gamma_1 \models_1 \Gamma_2 \wedge \ldots \wedge \Gamma_{n-1} \models_1 \Theta.$$

Now, since $\Sigma \models_1 \Gamma_1$ and $\Gamma_1 \subseteq Sn_{\mathcal{L}_w}$, we have from Theorem 3.6 that $(\Sigma \cap Sn_{\mathcal{L}_w}) \models_1 \Gamma_1$. Hence, $(\Sigma \cap Sn_{\mathcal{L}_w}) \models_n \Theta$ as well.

$\Leftarrow$) Assume $(\Sigma \cap Sn_{\mathcal{L}_w}) \models_n \Theta$. From Lemma 3.7 it follows that there exist sets $\Gamma_1, \Gamma_2, \ldots, \Gamma_{n-1} \subseteq Sn_{\mathcal{L}_w}$ such that

$$(\Sigma \cap Sn_{\mathcal{L}_w}) \models_1 \Gamma_1 \wedge \Gamma_1 \models_1 \Gamma_2 \wedge \ldots \wedge \Gamma_{n-1} \models_1 \Theta.$$

Now, since $(\Sigma \cap Sn_{\mathcal{L}_w}) \models_1 \Gamma_1$ and $\Gamma_1 \subseteq Sn_{\mathcal{L}_w}$, we have from Theorem 3.6 that $\Sigma \models_1 \Gamma_1$ as well. Hence, $\Sigma \models_n \Theta$. □

We now show that if a set of $\mathcal{L}_w$-sentences follows actively in two steps from a consistent set $\Sigma \subseteq Sn_{\mathcal{L}_w}$, then it follows actively in one step from Σ as well. The lemma will serve as the base case in an inductive proof of the equivalence of n-step and 1-step active consequence when restricted to $\mathcal{L}_w$-sentences and consistent premises.

Lemma 3.9. *Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$. If Σ is consistent and $\Sigma \models_2 \Theta$, then $\Sigma \models_1 \Theta$.*

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}_w}$ be consistent, and let $\Theta \subseteq Sn_{\mathcal{L}_w}$ be arbitrary. We show the contra-positive proposition: If $\Sigma \not\models_1 \Theta$, then $\Sigma \not\models_2 \Theta$.

Assume $\Sigma \not\models_1 \Theta$. From Theorem 2.2 it follows that $\Sigma \not\models \Theta$ classically.

We show that assuming $\Sigma \models_2 \Theta$, which by definition means that

$$(\exists \Gamma \subseteq Sn_{\mathcal{L}})(\Sigma \models_1 \Gamma \wedge \Gamma \models_1 \Theta),$$

leads to a contradiction. Note that since $\Theta \subseteq Sn_{\mathcal{L}_w}$, we can according to Lemma 3.7 without restriction assume that $\Gamma \subseteq Sn_{\mathcal{L}_w}$.

Let $\Gamma \subseteq Sn_{\mathcal{L}_w}$ be such that $\Sigma \models_1 \Gamma$. Then by Theorem 2.2, $\Sigma \models \Gamma$. Assume that $\Gamma \models_1 \Theta$, which again implies that $\Gamma \models \Theta$, or equivalently, that every model of Γ is a model of Θ. Then, since every model of Σ is a model of Γ, $\Sigma \models \Theta$, contrary to our assumption. Hence, $\Sigma \not\models_2 \Theta$. □

The following theorem allows results about 1-step active consequence to be generalised to active consequence under certain circumstances without having to resort to cumbersome inductive proofs. In particular, this allows us to extend Theorem 2.2 to general active consequence – that classical logical consequence and active consequence are equivalent (with respect to $\mathcal{L}_w$) when the premises are consistent.

Theorem 3.10. *Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$. If Σ is consistent, then*

$$\Sigma \models_a \Theta \text{ iff } \Sigma \models_1 \Theta.$$

Proof. The if part follows from definition. We show the only-if part, that for any sets $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$ such that Σ is (classically) consistent, $\Sigma \models_n \Theta \Rightarrow \Sigma \models_1 \Theta$, $n \geq 1$, by induction on n. The proposition obviously holds for $n = 1$, and by Lemma 3.9 it holds for $n = 2$.

Assume it holds for some $n \geq 2$. Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$ be such that Σ is consistent, and assume $\Sigma \models_{n+1} \Theta$. Then since $\Theta \subseteq Sn_{\mathcal{L}_w}$, we have according to Lemma 3.7 that

$$(\exists \Gamma \subseteq Sn_{\mathcal{L}_w})(\Sigma \models_n \Gamma \wedge \Gamma \models_1 \Theta),$$

which by the induction hypothesis implies

$$(\exists \Gamma \subseteq Sn_{\mathcal{L}_w})(\Sigma \models_1 \Gamma \wedge \Gamma \models_1 \Theta),$$

or equivalently, $\Sigma \models_2 \Theta$. Lemma 3.9 thus gives us $\Sigma \models_1 \Theta$. □

Note that Theorem 3.10 can also be proved by noting that Lemma 3.9 is equivalent to the following proposition: If $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$ and $\Sigma \models_1 \Theta$, then the closure of Θ under active consequence is a subset of the closure of Σ with respect to $\mathcal{L}_w$-sentences.[25]

Using Theorem 3.8, the restriction that $\Sigma \subseteq Sn_{\mathcal{L}_w}$ may be relaxed.

Corollary. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$. If Σ is t-strongly consistent, then*

$$\Sigma \models_a \Theta \textit{ iff } \Sigma \models_1 \Theta.$$

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-strongly consistent, and let $\Theta \subseteq Sn_{\mathcal{L}_w}$. Then $\Sigma \cap Sn_{\mathcal{L}_w}$ is classically consistent with respect to $\mathcal{L}_w$, and Theorem 3.10 thus gives

$$(\Sigma \cap Sn_{\mathcal{L}_w}) \models_a \Theta \text{ iff } (\Sigma \cap Sn_{\mathcal{L}_w}) \models_1 \Theta,$$

which by Theorem 3.8 implies

$$\Sigma \models_a \Theta \text{ iff } \Sigma \models_1 \Theta.$$ □

In Section 4 we prove that no corresponding theorem for t-weakly consistent sets exists.

Theorem 3.8 may also be used to give another characterisation of the relation between active consequence and its classical counterpart that was expressed in Theorem 2.2.

Theorem 3.11. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$ and $\Theta \subseteq Sn_{\mathcal{L}_w}$. If Σ is t-strongly consistent, then*

$$(\Sigma \cap Sn_{\mathcal{L}_w}) \models \Theta \textit{ iff } \Sigma \models_a \Theta.$$

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-strongly consistent, and let $\Theta \subseteq Sn_{\mathcal{L}_w}$. Then $\Sigma \cap Sn_{\mathcal{L}_w}$ is classically consistent with respect to $\mathcal{L}_w$, and the corollary to Theorem 2.2 thus gives

$$(\Sigma \cap Sn_{\mathcal{L}_w}) \models \Theta \text{ iff } (\Sigma \cap Sn_{\mathcal{L}_w}) \models_a \Theta,$$

[25] Denote the closure under active consequence of a set Σ with respect to $\mathcal{L}_w$-sentences with Σ^*. Assume $\Sigma \models_n \Theta$ for some $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$. Then, by Lemma 3.7, we know that there exist sets $\Gamma_1, \Gamma_2, \ldots, \Gamma_{n-1} \subseteq Sn_{\mathcal{L}_w}$ such that

$$\Sigma \models_1 \Gamma_1 \wedge \Gamma_1 \models_1 \Gamma_2 \wedge \ldots \wedge \Gamma_{n-1} \models_1 \Theta,$$

and thus we have that

$$\Sigma^* \supseteq \Gamma_1^* \supseteq \Gamma_2^* \supseteq \ldots \supseteq \Gamma_{n-1}^* \supseteq \Theta^*. \tag{1}$$

Obviously $\Theta \subseteq \Theta^*$, and hence it follows from (1) that $\Theta \subseteq \Sigma^*$. That is, $\Sigma \models_1 \Theta$.

which by Theorem 3.8 implies

$$(\Sigma \cap Sn_{\mathcal{L}_w}) \models \Theta \text{ iff } \Sigma \models_a \Theta. \qquad \square$$

We now turn our attention to the metalanguage $\mathcal{L}_a$. The following results about n-step active consequence consider conclusions that are in $Sn_{\mathcal{L}_a}$ rather than $Sn_{\mathcal{L}_w}$. In particular, we will try to characterise the least (and sometimes exact) number of steps it takes for an $\mathcal{L}_a$-sentence to follow if it follows at all.

Lemma 3.12. *Let Σ be a t-weakly consistent subset of $Sn_{\mathcal{L}}$. Then $\Sigma \models_n \{now(c_{t+n})\}$ at time t for every $n \geq 1$.*

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-weakly consistent, and let $n \geq 1$ be arbitrary. We first note that since $\Sigma \models_1 \emptyset$ and $\emptyset \models_1 \emptyset$ at any time t, $\Sigma \models_k \emptyset$ at any time t and $k > 1$ as well. Note also that $per_l(\{now(c_{t+n})\}) = \{now(c_{t+n})\}$ for every perception function per_l at every time l. Now, if $n = 1$ then obviously $\Sigma \models_1 \{now(c_{t+n})\}$ at time t since for every agent a, $H^a_{t+1} \models now(c_{t+1})$.

If instead $n > 1$ then, as noted above, $\Sigma \models_{n-1} \emptyset$, and since for every agent a, $H^a_{t+n} \models now(c_{t+n})$, $\emptyset \models_1 \{now(c_{t+n})\}$ at time $t + n - 1$. That is, $\Sigma \models_n \{now(c_{t+n})\}$ at time t. $\square$

Lemma 3.13. *Let Σ be a t-weakly consistent subset of $Sn_{\mathcal{L}}$. Then $\Sigma \models_a \{\neg now(c_k)\}$ at time t for every $k \in \mathbb{N}$. In particular, the conclusion follows in two steps if $k = t + 1$, otherwise it follows in one step.*

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-weakly consistent, and let $k \in \mathbb{N}$ be arbitrary.

Obviously, $\Sigma \models_1 \{\neg now(c_k)\}$ if $k \leq t$ or if $k > t + 1$ since for every agent a, $H^a_{t+1} \not\models now(c_k)$.

Now, assume $k = t+1$. As in the previous lemma, we have that $\Sigma \models_1 \emptyset$. Furthermore, $\emptyset \models_1 \{\neg now(c_k)\}$ at time $t+1$ since for every agent a, $H^a_{t+2} \not\models now(c_{t+1})$. That is, $\Sigma \models_2 \{\neg now(c_k)\}$. $\square$

Note that from the previous two lemmas, it follows that both singleton sets $\{now(c_{t+n})\}$ and $\{\neg now(c_{t+n})\}$ follow actively at time t from anything. (Their union is, by definition, never satisfied in an $\mathcal{L}_a$-structure and thus does not follow, though.)

The following result concerns negative sentences about the future.

Lemma 3.14. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-weakly consistent, and let φ be a negative $\mathcal{L}_a$-sentence about time $k > t + 1$. Then $\Sigma \models_1 \{\varphi\}$ at time t.*

Proof. Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-weakly consistent, and let φ be a negative $\mathcal{L}_a$-sentence about time $k > t + 1$.

Since $\neg\varphi$ is an affirmative sentence about time $k > t + 1$, we have by Theorem 3.1 that for every agent a, $H^a_{t+1} \not\models \neg\varphi$, and thus $H^a_{t+1} \models \varphi$. That is, $\Sigma \models_1 \{\varphi\}$. $\square$

We hypothesise that similar results may be shown for every other $\mathcal{L}_a$-sentence, and that these may be used to prove the following conjectures that characterise n-step active consequence when the conclusions are in $Sn_{\mathcal{L}_a}$.

Conjecture 3.15. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$, and let $\varphi \in Sn_{\mathcal{L}_a}$ be about time k. If $\Sigma \models_a \{\varphi\}$ at time t, then the least number of steps in which $\{\varphi\}$ follows actively from Σ is*

1. *1 step if $k \leq t$,*
2. *2 steps if $k = t + 1$ and φ is a negative now-sentence,*
3. *1 step if $k > t + 1$ and φ is a negative now-sentence,*
4. *n steps if $k = t + n$, $n \geq 1$, and φ is an affirmative now-sentence,*
5. *1 step if $k = t + 1$ and φ is Σ-determinable at time $t + 1$,*
6. *2 steps if $k = t + 1$ and φ is Σ-undeterminable at time $t + 1$,*
7. *1 step if $k > t + 1$ and φ is negative, or*
8. *$n + 1$ steps if $k = t + n$, $n > 1$, and φ is an affirmative contra- or bel-sentence.*

Conjecture 3.16. *Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$ be such that $\Sigma \models_a \Theta$ at time t. The $\mathcal{L}_a$-sentence in Θ that require the greatest number of steps to follow determine the least number of steps in which Θ follows from Σ.*

Case 1 in Conjecture 3.15 follows because if $k \leq t$, then the necessary information is already in H^a_{t+1}. We have already proved cases 2–4 and 7 in the preceding lemmas. Case 5 follows because if φ is Σ-determinable at time $t + 1$, then the needed evidence is already present at time t. If φ is Σ-undeterminable at time $t+1$, then it does not follow in one step according to Lemma 3.4, and it does not take more than two steps for it to follow since φ is about time $t + 1$ (cf. case 1 and 8). The last case follows since affirmative sentences about time k are by definition false prior to time k and we need an additional step to get the sentences that φ concerns into the knowledge base of every agent (cf. Example 2.7).

We note that affirmative *now*-sentences about time $t+n$ follow in exactly n steps and that affirmative *contra*- and *bel*-sentences about time $k > t + 1$ are false prior time k (see Theorem 3.1). This could be used to prove that under certain circumstances, if something follows actively in n steps then it also follows in $n + 1$ steps.

Conjecture 3.17. *Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}}$ be such that $\Sigma \models_n \Theta$ at time t for some $n \geq 1$. If there are no affirmative now-sentences and no sentences about time $k > t + n$ in Θ, then*

$$\Sigma \models_l \Theta \text{ for every } l \geq n.$$

The formal proofs are left for future work.

We conclude with the observation that there is nothing "special" about n-step active consequence. For $\mathcal{L}_w$-sentences it is equivalent to 1-step active consequence when the premises are consistent (otherwise it is, as we shall see later, equivalent to 4-step active consequence), and for $\mathcal{L}_a$-sentences the number of steps n is completely deterministic as well. It can be used to capture the least number of steps necessary for an $\mathcal{L}_a$-conclusion to follow. In particular, it captures that the time must be t for $now(t+1)$ to follow in one step and that the time must be at least t for affirmative *contra*- or *bel*-sentences about time t to follow.

3.6 Active-Sound Inferences

By Definition 2.24, an active sound inference is an inference in which the conclusion is an active consequence of the premises. Consequently, the following inference is active sound:

$$\frac{t:}{t+1 : now(c_{t+2})},$$

since, by Lemma 3.12 from the previous section, $now(c_{t+2})$ follows actively in two steps, and thus actively, from anything at time t. This is obviously problematic since $now(c_{t+2})$ is, by definition, not satisfied in any $\mathcal{L}_a$-structure at time t, and thus an agent reasoning using this rule will end up with a knowledge base that is not t-weakly consistent.

By altering the definition of active sound inference in such a way that time is taken under consideration, the problem pointed out above can be avoided.[26]

Definition 3.6 (n-step Active-Sound Inference). An *n-step active-sound inference* is one in which the conclusion is an n-step active consequence of the premises.

Note that the inference rules that were proved active sound in Section 2.8 are also 1-step active sound. When clear from context, we shall allow ourselves to drop the "n-step" prefix.

As mentioned before, we will in the next section prove that the explosive rule is active sound. With the new definition of active soundness, this claim must be modified with an "n-step" prefix. In particular, we shall prove that the explosive rule is 1-step active unsound, whereas the 2-step variant

$$\frac{t : \varphi, \neg\varphi}{t+2 : \psi},$$

where $\varphi, \psi \in Sn_{\mathcal{L}_w}$, is 2-step active sound.

[26] A similar modification has also been done in [AGGP05b].

In Section 2.8 we saw that the active version of the classical rule modus ponens was active sound. The rule is defined by

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

where $\varphi, \psi \in Sn_{\mathcal{L}_w}$. Note that the proof of Theorem 2.4 is valid also when φ or ψ are contradictions. Thus, like in the classical case, ψ follows actively from $\{\varphi, \varphi \rightarrow \psi\}$ even if φ (or ψ) is a contradiction.

A more general result follows from Theorem 2.2 if we restrict ourselves to consistent premises, namely that an inference from consistent premises using a classically sound inference rule is active sound.

Theorem 3.18. *All inferences using active versions of classically sound inference rules that involve only $\mathcal{L}_w$-sentences and consistent premises are active sound.*

Proof. Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$ be such that Σ is consistent, and let the following inference be an inference that uses an active version of a classically sound inference rule

$$\frac{t : \Sigma}{t+1 : \Theta}.$$

Since the underlying classical rule is sound, $\Sigma \models \Theta$, and hence, by Theorem 2.2, $\Sigma \models_1 \Theta$, that is, the inference is active sound. □

Finally, we note that, without further restrictions on perception functions, active consequence cannot incorporate the observation functions of step logic (see [EDP90]). Assume $\varphi \in Sn_{\mathcal{L}_w}$ is an observation at time $t+1$. Then the observation rule says that

$$\frac{t :}{t+1 : \varphi}.$$

Obviously, for any pair of perception functions, not every model of the perceived premises ($\emptyset$) models the perception of φ. That is, $\emptyset \not\models_1 \{\varphi\}$ at time t, and hence, by Theorem 3.10, the observation rule is active unsound.

4 Active Consequence is Explosive

In this section the main result of this thesis is presented, namely that the task of constructing a consequence relation that does not suffer from the drawbacks of classical logical consequence has failed. In particular, we show that the logic presented in [AGGP05a] (and [AGGP05b]) is not paraconsistent as anything follows actively from an inconsistency.

Note that the proofs below do not depend on any of the extensions of the original logic that were made in the previous section.

Theorem 4.1. *Let $\varphi, \psi \in Sn_{\mathcal{L}_w}$ be arbitrary. Then $\{\varphi, \neg\varphi\} \models_2 \{\psi\}$. That is, anything in $Sn_{\mathcal{L}_w}$ follows actively from a direct contradiction.*

Proof. Let $\varphi, \psi \in Sn_{\mathcal{L}_w}$ be arbitrary, and let $\Sigma = \{\varphi, \neg\varphi\}$, $\Theta = \{\psi\}$, and $\Gamma = \{\varphi \to \psi, \neg\varphi \to \psi\}$. Assume without loss of generality that the string representations of φ and $\varphi \to \psi$ precede alphabetically the representations of $\neg\varphi$ and $\neg\varphi \to \psi$, respectively, and that φ and ψ have k and l sentence-symbol tokens, respectively. We show that $\Sigma \models_1 \Gamma$ and $\Gamma \models_1 \Theta$.

Let per^u_t be the perception function at time t determined by the sequence $\langle 1, 2, 3, \ldots\rangle$, with $per^u_t(\Sigma) = \{\varphi^1, \neg\varphi^2\}$. Let per'_{t+1} be the perception function at time $t+1$ determined by the sequence

$$\langle \underbrace{k+1, k+2, \ldots, 2k}_{k}, \underbrace{1, 1, \ldots, 1}_{l}, \underbrace{1, 2, \ldots, k}_{k}, 1, 1, 1, \ldots\rangle,$$

with $per'_{t+1}(\Gamma) = \{\varphi^2 \to \psi^i, \neg\varphi^1 \to \psi^i\}$. Note that φ^1 and φ^2 have no sentence symbols in common and that every sentence symbol in them is unique. In particular, this implies that $per^u_t(\Sigma)$ is consistent (see Theorem 2.1). Now, for every model h of $per^u_t(\Sigma)$, $h(\varphi^1) = \top$ and $h(\varphi^2) = \bot$, and thus, $h(\neg\varphi^1 \to \psi^i) = \top$ and $h(\varphi^2 \to \psi^i) = \top$. That is, $\Sigma \models_1 \Gamma$.

Let per''_{t+1} be the perception function at time $t+1$ determined by the sequence

$$\langle \underbrace{1, 1, \ldots, 1}_{k}, \underbrace{1, 2, \ldots, l}_{l}, \underbrace{1, 1, \ldots, 1}_{k}, 1, 2, 3, \ldots\rangle,$$

with $per''_{t+1}(\Gamma) = \{\varphi^i \to \psi^u, \neg\varphi^i \to \psi^u\}$. Let per^u_{t+2} be the perception function at time $t+2$ determined by the sequence $\langle 1, 2, 3, \ldots\rangle$, with $per^u_{t+2}(\Theta) = \{\psi^u\}$. Note that every sentence symbol in ψ^u is unique and that therefore ψ^u is satisfiable. In particular, this means that $per''_{t+1}(\Gamma)$ is consistent. Now, for every model h of $per''_{t+1}(\Gamma)$, either $h(\varphi^i) = \top$ or $h(\varphi^i) = \bot$, and thus $h(\psi^u) = \top$. That is, $\Gamma \models_1 \Theta$. □

Corollary. *Active consequence is explosive, and the logic is not paraconsistent.*

Corollary. *The explosive rule is active sound.*[27]

Note that Theorem 4.1 can be generalised to sets of $\mathcal{L}_w$-sentences since the perception function per'_{t+1} above easily can be extended in such a way that the sentences $\varphi \to \psi$ and $\neg\varphi \to \psi$ follow for every sentence ψ in the conclusion Θ.

Corollary. *Let $\varphi \in Sn_{\mathcal{L}_w}$ be arbitrary. If $\Theta \subseteq Sn_{\mathcal{L}_w}$, then $\{\varphi, \neg\varphi\} \models_2 \Theta$. That is, every subset of $Sn_{\mathcal{L}_w}$ follows actively from a direct contradiction.*

The following lemma states that a direct contradiction follows from every set that is inconsistent with respect to $\mathcal{L}_w$. Then the proposition that everything in $Sn_{\mathcal{L}_w}$ follows actively from a t-weakly consistent set that is not t-strongly consistent follows immediately from the preceding corollary.

Lemma 4.2. *Let $\Sigma \subseteq Sn_{\mathcal{L}}$ be t-weakly consistent. If $\Sigma \cap Sn_{\mathcal{L}_w}$ is inconsistent, then*

$$\Sigma \models_2 \{S_i, \neg S_i\},$$

for some $i \in \mathbb{N}$.

Proof. Consider the identity perception function determined by the sequence $\langle 1, 1, 1, \ldots\rangle$. The identity perception of a set of $\mathcal{L}_w$-sentences is consistent if and only if the set itself is consistent. In fact, this is the case for any perception function that maps every occurrence of a sentence symbol using the same superscript. Thus, if the perception of an inconsistent set is to be consistent, we must have that every model h of the perceived set has $h(S_i^j) \neq h(S_i^k)$ for some i and $j \neq k$.

Now, let per_t be an arbitrary perception function at time t such that $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ is consistent. (The existence of per_t is guaranteed by Theorem 2.1.) From the observation above, we know that for every model h of $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$, there exists an index i for which $h(S_i^j) \neq (S_i^k)$ for some superscripts j and k. Note that these index-superscript triples $\langle i, j, k\rangle$ need not be the same for every model. Let I be the set of all such triples (possibly originating from different models h).

Assume $\langle i, j, k\rangle \in I$ is a triple for which there exists a model h of $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ that assigns the same truth value to both S_i^j and S_i^k. (Note that if such a triple exists then the cardinality of I is strictly greater than one.) Consider the perception function per'_t that maps every sentence-symbol token to the same sentence symbol as per_t does except for the S_i-tokens previously mapped to S_i^k, which now get mapped to S_i^j. Now, the models of $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ that assigned the same truth value to both S_i^j and S_i^k are also models of $per'_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$. We can thus, without loss of

[27] As noted in the previous section, with a step-sensitive definition of active soundness, we have shown that the 2-step explosive rule is 2-step active sound. The 1-step version is, with this definition, proved 1-step active unsound later in this section.

generality, assume that every model h of $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ assigns different truth values to the two sentence symbols S_i^j and S_i^k for every triple $\langle i, j, k\rangle$ in I (i.e. the elements of I do not originate from different models).

We have a perception function per_t such that $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ is consistent. Furthermore, there exists a triple $\langle i, j, k\rangle \in I$ such that every model h of $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ has $h(S_i^j) \neq h(S_i^k)$ and thus also $h(S_i^j \wedge S_i^k) = \bot$ and $h(S_i^j \vee S_i^k) = \top$.[28] Denote with Γ the set $\{\neg(S_i \wedge S_i), S_i \vee S_i\}$ and let per'_{t+1} be a perception function at time $t+1$ with $per'_{t+1}(\Gamma) = \{\neg(S_i^j \wedge S_i^k), S_i^j \vee S_i^k\}$. Then every model h of $per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}$ is also a model of $per'_{t+1}(\Gamma)$, and thus we have shown that $\Sigma \models_1 \Gamma$.

Now, let per''_{t+1} and per_{t+2} be perception functions at time $t+1$ and $t+2$, respectively, with $per''_{t+1}(\Gamma) = \{\neg(S_i^1 \wedge S_i^1), S_i^2 \vee S_i^2\}$ and $per_{t+2}(\Theta) = \{S_i^2, \neg S_i^1\}$. Since $per''_{t+1}(\Gamma)$ is consistent and every model h of $per''_{t+1}(\Gamma)$ has $h(S_i^1) = \bot$ and $h(S_i^2) = \top$, h is a model of $per_{t+2}(\Theta)$ as well. That is, $\Gamma \models_1 \Theta$ at time $t+1$, and we have proved that $\Sigma \models_2 \Theta$ at time t. □

Theorem 4.3. *Everything in $Sn_{\mathcal{L}_w}$ follows actively in four steps from a t-weakly consistent set that is not t-strongly consistent. In particular, everything follows actively from an inconsistent knowledge base.*

Note that even though everything follows actively from an inconsistent set, not everything follows in one step. Consider, for instance, the sets $\Sigma = \{S_1, \neg S_1\}$ and $\Theta = \{S_2\}$. Let per_t and per_{t+1} be arbitrary perception functions at time t and $t+1$, respectively. There are two cases to consider: Either $per_t(\Sigma) = \{S_1^i, \neg S_1^j\}$ with $i \neq j$, or $per_t(\Sigma) = \{contra(d_{S_1}, d_{\neg S_1}, c_t)\}$.

1. $per_t(\Sigma) = \{S_1^i, \neg S_1^j\}$, with $i \neq j$. Then there exists an $\mathcal{L}'_w$-interpretation h which models the perceived premises but not the conclusion, namely one with $h(S_1^i) = \top$ and $h(S_1^j) = h(S_2^k) = \bot$ for every superscript k.

2. $per_t(\Sigma) = \{contra(d_{S_1}, d_{\neg S_1}, c_t)\}$. Obviously, there exists an $\mathcal{L}'_w$-interpretation h which models the perceived premises,

$$per_t(\Sigma) \cap Sn_{\mathcal{L}'_w} = \emptyset,$$

   but not the conclusion (for instance h above).

Hence, we have shown that $\Sigma \not\models_1 \Theta$. In particular, this means that there is no equivalence between active consequence and 1-step active consequence when the conclusion is a subset of $Sn_{\mathcal{L}_w}$, and thus Theorem 3.10 can not be generalised to t-weakly consistent sets.

[28] We allow ourselves to use $\varphi \wedge \psi$ and $\varphi \vee \psi$ as syntactic sugar for $\neg(\varphi \rightarrow \neg\psi)$ and $\neg\varphi \rightarrow \psi$, respectively.

We conclude this section by noting that Theorem 2.2 can be extended to include also inconsistent premises.[29] That is, active consequence is equivalent to classical logical consequence with respect to $\mathcal{L}_w$.

Theorem 4.4. *Let* $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$. *Then*

$$\Sigma \models \Theta \textit{ iff } \Sigma \models_a \Theta.$$

Proof. Let $\Sigma, \Theta \subseteq Sn_{\mathcal{L}_w}$. If Σ is consistent, then by Theorem 2.2, $\Sigma \models \Theta$ if and only if $\Sigma \models_a \Theta$, so assume Σ is inconsistent.

Assume $\Sigma \models \Theta$. Then, by Theorem 4.3, $\Sigma \models_a \Theta$.

Assume $\Sigma \models_a \Theta$. Then, since Σ is inconsistent, $\Sigma \models \Theta$. □

[29] As noted above, none of the results presented so far in this section depend on any extension of the original logic. Theorem 2.2 does, however, indirectly depend on the refined concept of active consequence presented in Section 3.4.

5 Conclusions

In this thesis we have analysed a proposal for a semantics for active logic based on the concept of perception functions. Using perception functions a notion of perceived temporal structure is defined, which allows inconsistent knowledge bases to have models. These structures are then used to construct a consequence relation called active consequence.

Active consequence was previously believed to coincide with classical logical consequence when restricted to the world language and consistent premises. We have shown that this identity does not hold due to the problem of Σ-undeterminism – that there are sentences for which it cannot be determined whether they follow actively from a given set Σ – and suggest a refined definition of active consequence as a solution.

Our main result, however, is that we have shown that active consequence is explosive, that is, that anything follows actively from a direct contradiction. Consequently, and contrary to what has been previously claimed, a logic based on this consequence relation is not paraconsistent.

5.1 Accomplishments

- We have shown that active consequence is explosive, and thus cannot be used for a paraconsistent logic.
- We have pointed out that there are sentences which are not Σ-determinable, by which we mean that it cannot be determined whether they follow actively from a given set Σ.
- As a result of the problem of Σ-undeterminism, active consequence does not coincide with classical logical consequence when restricted to the world language and consistent premises.
- We have suggested a refined definition of active consequence, which solves the problem of Σ-undeterminism and makes active consequence equivalent to classical logical consequence when restricted to the world language and consistent premises.
- Several lemmas and theorems regarding active consequence have been proved. These results shed more light on active consequence and its relation to classical logical consequence. The results and their proofs can probably also be reused or serve as inspiration in future work on a modified consequence relation.
- In particular, it has been proved that active consequence is equivalent to one-step active consequence when the premises are consistent and the conclusion is part of the world language. Furthermore, the restriction to consistent premises cannot be relaxed since every conclusion

follows in two steps – but not in one step – from directly inconsistent sets. Similarly, anything follows in four steps from inconsistent sets.

- As part of our analysis, we have reformulated the original theory in order to gain stringency and understandability. Several errors in the original paper [AGGP05a] (and [AGGP05b]) have also been corrected.

- We have stressed that the restrictions imposed on the logic by the definition of the language $\mathcal{L}$ render a logic that is too weak. In particular, metareasoning cannot involve world knowledge.

- We have also noted that even if the language restrictions are relaxed in order to allow non-trivial metareasoning, modelling reasoning about future beliefs is problematic with the current $\mathcal{L}_a$-semantics.

- Since there exist perception functions that render inconsistent perceived knowledge bases, we have concluded that with the original definition of active consequence from [AGGP05a], everything follows vacuously in one step from an indirectly inconsistent set. We have proposed a minor modification of the definition as a solution.

5.2 Future Work

In order to meet the requirements put on active logic, a new consequence relation that is not explosive needs to be defined. Perhaps it is possible to redefine active consequence or at least to use some part of the original theory, such as perception functions or perceived temporal structures.

Future work will also need to relax the restrictions on the language $\mathcal{L}$ so that metareasoning can involve world knowledge.

Non-trivial reasoning about future beliefs requires the $\mathcal{L}_a$-semantics to be refined in such a way that not every affirmative sentence about the future is false by definition. One way could be to extend the $\mathcal{L}_a$-structure to include also the agent's future reasoning. This way, also sentences about the future would become Σ-undeterminable at time $t+1$ and would thus, in a sense, be contingent rather than determined by their syntactical structure.

If active consequence can be made non-explosive, it would be interesting to analyse the consequences of modifying the relation between Σ and the agent knowledge base KB_t^a by allowing Σ to be a proper subset of KB_t^a in the definition of 1-step active consequence.

Conjecture 3.15, Conjecture 3.16 and Conjecture 3.17, which characterise n-step active consequence when the conclusion is in $Sn_{\mathcal{L}_a}$, should be formally proved using the lemmas and ideas found in Section 3.5.

References

[AGGP05a] Michael L. Anderson, Walid Gomaa, John Grant, and Don Perlis. On the reasoning of real-world agents: Toward a semantics for active logic. In *Proceedings of the 7th Annual Symposium on the Logical Formalization of Commonsense Reasoning*, Corfu, Greece, 2005.

[AGGP05b] Michael L. Anderson, Walid Gomaa, John Grant, and Don Perlis. Active logic semantics for a single agent in a static world, 2005. Forthcoming. Manuscript from October 14, 2005.

[Ågo04] Thomas Ågotness. *A Logic of Finite Syntactic Epistemic States*. PhD thesis, Department of Informatics, University of Bergen, Norway, 2004.

[EDP90] Jennifer J. Elgot-Drapkin and Donald Perlis. Reasoning situated in time I: Basic concepts. *Journal of Experimental & Theoretical Artificial Intelligence*, 2(1):75–98, 1990.

[GKP00] John Grant, Sarit Kraus, and Donald Perlis. A logic for characterizing multiple bounded agents. *Autonomous Agents and Multi-Agent Systems*, 3(4):351–387, 2000.

[GW01] Dov M. Gabbay and John Woods. The new logic. *Logic Journal of the IGPL*, 9(2):141–174, 2001.

[Lev84] Hector J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202, Austin, TX, USA, 1984. American Association for Artificial Intelligence.

[Man99] María Manzano. *Model theory*. Oxford University Press, Oxford, UK, 1999.

[PT04] Graham Priest and Koji Tanaka. Paraconsistent logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2004. `http://plato.stanford.edu/archives/win2004/entries/logic-paraconsistent/`.

[vW65] Georg Henrik von Wright. *Logik, filosofi och språk*. Nya Doxa, Nora, Sweden, 1965.