Honesty in Partial Logic

Wiebe van der Hoek*
Dept. of Computer Science
Utrecht University
P.O. Box 80089
3508 TB Utrecht
the Netherlands
wiebe@cs.ruu.nl

Jan Jaspars
Dept. of Language & Literature
Tilburg University
P.O. Box 90153
5000 LE Tilburg
the Netherlands
jaspars@kub.nl

Elias Thijsse
Dept. of Language & Literature
Tilburg University
P.O. Box 90153
5000 LE Tilburg
the Netherlands
thysse@kub.nl

Contents

Abstract

We propose an epistemic logic in which knowledge is fully introspective and implies truth, although truth need not imply epistemic possibility. The logic is presented in sequential format and is interpreted in a natural class of partial models, called balloon models. We examine the notions of honesty and circumscription in this logic: What is the state of an agent that 'only knows φ' and which *honest* φ enable such circumscription? Redefining *stable sets* enables us to provide suitable syntactic and semantic criteria for honesty. The rough syntactic definition of honesty is the existence of a minimal stable expansion, so the problem resides in the ordering relation underlying minimality. We discuss three different proposals for this ordering, together with their semantic counterparts, and show their effects on the induced notions of honesty.

keywords: *circumscription, honesty, modal logic, partial models, stable sets*

*This author was partially supported by ESPRIT Basic Research Action No. 6156 (DRUMS).

1 Introduction

In this paper we argue that honesty in knowledge representation calls for a *partial* approach, for reasons of adequacy and efficiency. Let us first (re)introduce the central concepts, honesty and partiality.

Honesty is the quality of a proposition which can be said to be *only* known, i.e. knowing that fact and its consequences, but not knowing more than that. For example, you may only know that Pat will come tomorrow, without knowing anything at all about, say, Sue. Also, you may only know that either Pat or Sue will come tomorrow, which implies you do not know which one of the two will come. These are examples of honest knowledge. By contrast, you cannot honestly claim to only know that you know *whether* Pat will come, for then you would either know that Pat will come or know that she won't come, both options being logically stronger than what is supposed to be known.

Partiality is the idea of not giving a truth value to every proposition. In a given situation the truth of a formula may be undefined, for example due to lack of information. Such undefinedness may even occur for classical tautologies, such as the well-known 'law of excluded middle' (*tertium non datur*) $\varphi \vee \neg\varphi$, which is therefore not valid in the partial semantics we advocate.

Partiality and honesty may seem totally unrelated themes, but we will argue that in fact they are very closely related. Let us reinspect the case in which you only know that Pat will come tomorrow. This, we claim, does not involve any knowledge about Sue or some other part of the universe. For example, it does not even imply that you know the possibility that Sue will come too: you may not be acquainted to her, or just not consider her possible arrival. In a straightforward total semantics ignorance leads to wide knowledge of possibilities, which, however, contradicts the initial idea of *only* knowing some honest formula. The proliferation problem simply does not occur in our partial semantics, since facts unrelated to some honest formula can be left undefined.

This, in a nutshell, is our prime motivation for 'going partial': it provides a more adequate and natural account of circumscription (i.e. describing what you only know). But there is more to it. One of the other advantages of partial semantics is its efficiency, which is reflected in the much smaller size of the characterizing models. Classical possible world semantics leads to a combinatorial explosion: the less one knows, the bigger the model. For example, honest knowledge of p and complete ignorance of n other propositional variables leads to 2^n worlds in a model that represents what one only knows. This may be contrasted to a partial model to the same effect that uses but one or two worlds. Moreover, addition of information may lead to growth of the partial model, unlike the elimination usual in possible worlds models — again the partial approach seems more natural and intuitive.

Finally partial semantics allows for a greater flexibility with respect to the epistemic background logic. For the case study presented in this paper this is revealed in adopting the veridicality principle of knowledge $\Box\varphi \Rightarrow \varphi$ ('if you know something, it must be true'), without being forced to accept its contrapositive $\varphi \Rightarrow \Diamond\varphi$ ('if something is true, you must consider it possible'). Moreover, knowledge will be fully introspective: both positive introspection ('if you know something, then you know that you know it') and negative introspection ('if you consider something possible, then you know you do'), as well as their contrapositives are properties our logic embraces. In all, the logic resembles a weak variant of the classical system **S5**. Apart from fitting our intuitions about (strong) knowledge, this logic enables us to simplify the kind of models needed, essentially omitting most of the relational structure.

The rest of the paper is organized as follows. In the next section we introduce the epistemic logic, presenting its language, semantics, inference system, and proving its completeness. Then, in section 3, we study circumscription for this logic, discussing different notions of honesty, both from a deductive (minimal stable sets) and from a modeltheoretic perspective (minimal models). Moreover, we provide a useful inferential test for honesty (disjunction properties). We round off by summarizing our results in the conclusion.

2 The Logic

In this section we introduce a partial modal logic **L** of which we will investigate the notions of stability, honesty and several disjunction properties in subsequent sections. We present our logic following a common pattern: we first give its language and a partial semantics (section 2.1), then we provide a deductive system for **L** (2.2) and round off with a completeness result (2.3) connecting them.

2.1 Language and Semantics

Definition 2.1 Let $\mathcal{P}$ be a non-empty countable set of propositional variables. The *language* $\mathcal{L}$ is the smallest superset of $\mathcal{P}$ such that

$$\varphi, \psi \in \mathcal{L} \Rightarrow \neg\varphi, (\varphi \wedge \psi), \bot, \Box\varphi \in \mathcal{L}.$$

$\mathcal{L}_0$ is the subset of $\mathcal{L}$ of all formulas which do not contain $\Box$-operators. For any $\Gamma \subseteq \mathcal{L}$, we write Γ_0 for $\Gamma \cap \mathcal{L}_0$ and $\overline{\Gamma}$ for $\{\varphi \in \mathcal{L} \mid \varphi \notin \Gamma\}$. Moreover, for any $\Gamma \subseteq \mathcal{L}$ and any $\odot \in \{\neg, \Box, \Diamond\}$, we define $\odot\Gamma = \{\odot\gamma \mid \gamma \in \Gamma\}$ and $\odot^-\Gamma = \{\gamma \mid \odot\gamma \in \Gamma\}$.

Here, the intended meaning of $\Box\varphi$ is that 'φ is known'. We write $\top$ for $\neg\bot$, $\varphi \vee \psi$ for $\neg(\neg\varphi \wedge \neg\psi)$ and $\Diamond\varphi$ means $\neg\Box\neg\varphi$. It is important to note that in our set-up, $\Diamond\varphi$ does not just mean that $\neg\varphi$ is not known, but that the agent considers some epistemic alternative to be possible, in which φ has a meaning: it is true!

Given a set of formulas Γ, we may consider its *objective kernel* (Γ_0), the *knowledge* it encodes ($\Box^-\Gamma$) and its *possibilities* ($\Diamond^-\Gamma$). These sets induce the following orderings.

Definition 2.2 Let Γ and Γ' be sets of formulas of $\mathcal{L}$. Then:

- $\Gamma \subseteq_0 \Gamma' \Leftrightarrow \Gamma_0 \subseteq \Gamma'_0$
- $\Gamma \subseteq_\Box \Gamma' \Leftrightarrow \Box^-\Gamma \subseteq \Box^-\Gamma'$
- $\Gamma \subseteq_\Diamond \Gamma' \Leftrightarrow \Diamond^-\Gamma \subseteq \Diamond^-\Gamma'$

Each of these orders can be linked to an equivalence relation, e.g. $\Gamma =_0 \Gamma' \Leftrightarrow \Gamma_0 = \Gamma'_0$. Let $\Re$ be some subset of $\wp(\mathcal{L})$, the power set of $\mathcal{L}$. For any $\star \in \{0, \Box, \Diamond\}$, we say that $\Gamma \in \Re$ is $\subseteq_\star$-*minimal in* $\Re$, if for all $\Gamma' \in \Re$, $\Gamma \subseteq_\star \Gamma'$, and similarly for $\subseteq$-minimality in $\Re$. Note that these minimal sets are the first (smallest) elements with respect to the corresponding ordering, rather than the elements without a predecessor.

We now give a formal interpretation of the language $\mathcal{L}$. The mathematical structure for such an interpretation is a Kripke model with partial worlds. Since we are only interested in models for our epistemic logic here, we do not have to consider arbitrary partial Kripke models.[1] Instead, we restrict attention to what we call 'balloon models', which somewhat remind of the well-known **KD45**-Kripke models. The basic entities in our balloon models are partial worlds, which are defined in terms of partial valuations.

Definition 2.3 A *partial valuation* V is a partial function which assigns truth-values to a given set of propositional variables $\mathcal{P}$. The collection of all partial valuations is denoted by VAL. The *domain of* V is defined as $Dom(V) = \{\, p \in \mathcal{P} \mid V(p) \in \{0,1\}\,\}$. $V' \in \mathsf{VAL}$ is said to be an *extension* of $V \in \mathsf{VAL}$ if $V(p) = V'(p)$ for all $p \in Dom(V)$. We abbreviate this relation by $V \sqsubseteq V'$.

[1] For a general approach, see [Thi92] or [JT93].

Definition 2.4 A *balloon model* is a triple $M = \langle W, g, V\rangle$ with W a non-empty finite set of worlds, called the *balloon*, g the *root* or *generator* of the model, and V a global valuation function $V : W \cup \{g\} \to \mathsf{VAL}$, such that $V(w) \sqsubseteq V(g)$ for certain $w \in W$. We also write M_g for such a model: note that any $w \in W$ and $V : W \to \mathsf{VAL}$ give rise to a model $M_w = \langle W, w, V\rangle$.

The *truth* and *falsity* of a formula $\varphi \in \mathcal{L}$ in a balloon model $M = \langle W, g, V\rangle$, written as $M \models \varphi$ and $M \mathrel{=\!\!|} \varphi$, respectively, are defined by induction:

$$
\begin{array}{llll}
M \not\models \bot & & M \mathrel{=\!\!|} \bot & \\
M \models p & \Leftrightarrow V(g)(p) = 1 \ (p \in \mathcal{P}) & M \mathrel{=\!\!|} p & \Leftrightarrow V(g)(p) = 0 \ (p \in \mathcal{P}) \\
M \models \neg\varphi & \Leftrightarrow M \mathrel{=\!\!|} \varphi & M \mathrel{=\!\!|} \neg\varphi & \Leftrightarrow M \models \varphi \\
M \models \varphi \wedge \psi & \Leftrightarrow M \models \varphi \text{ and } M \models \varphi & M \mathrel{=\!\!|} \varphi \wedge \psi & \Leftrightarrow M \mathrel{=\!\!|} \varphi \text{ or } M \mathrel{=\!\!|} \psi \\
M \models \Box\varphi & \Leftrightarrow M_w \models \varphi \text{ for all } w \in W & M \mathrel{=\!\!|} \Box\varphi & \Leftrightarrow M_w \mathrel{=\!\!|} \varphi \text{ for some } w \in W
\end{array}
$$

Note the special role played in the truth definition by the root of the model. Although we usually display the models with the root outside of the balloon, the recursive $\Box$-clauses show that this need not be the case. (Alternatively, we can duplicate the root, the new root being outside of the balloon.)

Also note that the truth-definitions yield the intended effect for $\Diamond$-formulas: we have that $M \models \Diamond\varphi \Leftrightarrow M_w \models \varphi$ for some $w \in W$. In particular, our partial semantics makes $\Box\varphi \vee \neg\Box\varphi$, and hence $\Box\neg\varphi \vee \Diamond\varphi$ invalid. This reflects the idea that, in our opinion, $\Diamond\varphi$-formulas should express some positive evidence about φ, not just lack of knowledge of $\neg\varphi$.

Example 2.5 Figure 1 denotes two typical balloon models M and M'. We call a world in which no atom is true or false an empty world; note that M' has such a world. Moreover note that $M \models \Box p \wedge \Diamond q$; $M \not\models \Diamond\neg q$ and $M \not\models \Box q$. For M', we have $M' \models \Box\top$, but at the same time $M' \not\models \Box(\neg p \vee p)$, $M' \mathrel{\not=\!\!|} \Box(\neg p \vee p)$.

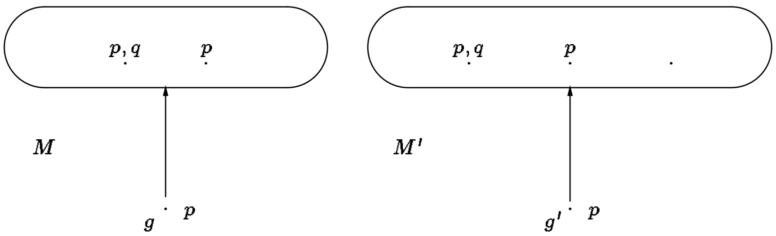

Figure 1: two balloon models M and M'

For any model $M = \langle W, g, V\rangle$ we define the *theory* $Th(M)$ of M by $Th(M) = \{\varphi \in \mathcal{L} \mid M \models \varphi\}$, the *knowledge* $\kappa(M)$ in M by $\kappa(M) = \{\varphi \in \mathcal{L} \mid M \models \Box\varphi\}$ and the *possibilities* $\pi(M)$ by $\pi(M) = \{\varphi \in \mathcal{L} \mid M \models \Diamond\varphi\}$. Note that $\kappa M = \Box^{-}Th(M)$ and $\pi(M) = \Diamond^{-}Th(M)$. Let Γ and Δ be sets of formulas. We write $\Gamma \models \Delta$, if all balloons which verify all members of Γ also verify at least one of the elements of Δ, i.e. $\forall M : \Gamma \subseteq Th(M) \Rightarrow \Delta \cap Th(M) \neq \emptyset$. Finally, we write $M \models \Gamma$ for $\Gamma \subseteq Th(M)$.

Lemma 2.6 (Propositional Persistence)
Let $M = \langle W, g, V\rangle$ and $M' = \langle W', g', V'\rangle$ be two balloon models. Then, for all $w \in W \cup \{g\}$ and all $w' \in W' \cup \{g'\}$:

$$V(w) \sqsubseteq V'(w') \Leftrightarrow \forall \pi \in \mathcal{L}_0 : (M_w \models \pi \Rightarrow M'_{w'} \models \pi)$$

Lemma 2.7 (Internal Persistence)
For every balloon model $M = \langle W, g, V\rangle$ and $w, w' \in W \cup \{g\}$:

$$V(w) \sqsubseteq V(w') \Leftrightarrow \forall \varphi \in \mathcal{L} : (M_w \models \varphi \Rightarrow M_{w'} \models \varphi)$$

Proof: The '$\Leftarrow$'-part follows from Lemma 2.6. The '$\Rightarrow$'-part is proven using induction on φ; in fact, to take care of negations (and hence of $\Diamond$-formulas), it is more convenient to prove that for all $\varphi \in \mathcal{L}$:

$$V(w) \sqsubseteq V(w') \Rightarrow [(M_w \models \varphi \Rightarrow M_{w'} \models \varphi) \text{ and } (M_w \dashv \varphi \Rightarrow M_{w'} \dashv \varphi)]$$

For $\varphi \in \mathcal{P}$, this follows from Lemma 2.6; for conjunctions and negations it is straightforward. So let us consider $\varphi = \Box\psi$; it turns that no inductive assumption about ψ is needed: $M \models \Box\psi \Leftrightarrow \forall v \in W, M_v \models \psi \Leftrightarrow M_{w'} \models \Box\psi$ ∎

There is also a more global way to achieve persistence. The notion that we need is a special case of a definition given in [JT93].

Definition 2.8 For two balloon models $M = \langle W, g, V\rangle$ and $M' = \langle W', g', V'\rangle$ we say that M' *bisimulates* M, if

- $V(g) \sqsubseteq V'(g)$
- $\forall w \in W\ \exists w' \in W'$ such that $V(w) \sqsubseteq V'(w')$
- $\forall w' \in W'\ \exists w \in W$ such that $V(w) \sqsubseteq V'(w')$

Theorem 2.9 (General Persistence)

$$M' \text{ bisimulates } M \Leftrightarrow \forall \varphi \in \mathcal{L} : M \models \varphi \Rightarrow M' \models \varphi$$

Proof: ¿From left to right, the proof is similar to that of lemma 2.7. ¿From right to left, suppose that $M' = \langle W', g', V'\rangle$ does not bisimulate $M = \langle W, g, V\rangle$. By definition 2.8, we have one of the following cases:

- $V(g) \not\sqsubseteq V'(g')$. Applying lemma 2.6 we find a formula $\pi \in \mathcal{L}_0$ such that $M \models \pi$ and $M' \not\models \pi$.
- $\exists w \in W\ \forall w' \in W'\ V(w) \not\sqsubseteq V'(w')$. For such w lemma 2.6 gives us, for each $w' \in W'$, a formula $\pi_{w'} \in \mathcal{L}_0$ for which $M_w \models \pi_{w'}$, but $M'_{w'} \not\models \pi_{w'}$. But then $M \models \Diamond \bigwedge_{w' \in W'} \pi_{w'}$, whereas $M' \not\models \Diamond \bigwedge_{w' \in W'} \pi_{w'}$.
- $\exists w' \in W'\ \forall w \in W\ V(w) \not\sqsubseteq V'(w')$. For such w' lemma 2.6 gives us, for each $w \in W$, formulas $\pi_w \in \mathcal{L}_0$ such that $M_w \models \pi_w$, $M'_{w'} \not\models \pi_w$. But then $M \models \Box \bigvee_{w \in W} \pi_w$, whereas $M' \not\models \Box \bigvee_{w \in W} \pi_w$.

Summarizing, we see that if M' does not bisimulate M, we always find a formula φ for which $M \models \varphi$, $M' \not\models \varphi$. ∎

2.2 Deductions in L

We now formally define the deductive machinery of our logic. The sequent $\Gamma \vdash \Delta$ should be understood as: 'the disjunction of the members of Δ follows from the conjunction of the formulas in Γ'. Instead of $\Gamma \cup \{\varphi\}$ and $\Gamma \cup \Delta$ we write Γ, φ and Γ, Δ respectively.

Definition 2.10 To start with, we distinguish the following *structural rules*:

- $\Gamma \cap \Delta \neq \emptyset \Rightarrow \Gamma \vdash \Delta$ START
- $\dfrac{\Gamma \vdash \Delta \quad \Delta \subseteq \Delta' \quad \Gamma \subseteq \Gamma'}{\Gamma' \vdash \Delta'}$ MON
- $\dfrac{\Gamma \vdash \varphi, \Delta \quad \Gamma', \varphi \vdash \Delta'}{\Gamma, \Gamma' \vdash \Delta, \Delta'}$ CUT

If we add to those structural rules the following *propositional rules*, we obtain the partial propositional logic $\mathbf{rL}^+$ (from [Thi90]). Those rules explain how the logical constants (c.q. connectives) are introduced on the left (L-TRUE) and right hand side (R-TRUE) of the '$\vdash$', respectively; possibly accompanied with a negation sign (L-FALSE or R-FALSE).

- $\dfrac{\Gamma \vdash \varphi, \Delta}{\Gamma, \neg\varphi \vdash \Delta}$ L-TRUE$\neg$
- $\dfrac{\Gamma, \varphi, \psi \vdash \Delta}{\Gamma, \varphi \wedge \psi \vdash \Delta}$ L-TRUE $\wedge$
- $\dfrac{\Gamma \vdash \varphi, \Delta \quad \Gamma' \vdash \psi, \Delta'}{\Gamma, \Gamma' \vdash \varphi \wedge \psi, \Delta, \Delta'}$ R-TRUE $\wedge$
- $\Gamma \vdash \neg\bot, \Delta$ R-FALSE $\bot$
- $\dfrac{\Gamma, \varphi \vdash \Delta}{\Gamma, \neg\neg\varphi \vdash \Delta}$ L-FALSE $\neg$
- $\dfrac{\Gamma \vdash \varphi, \Delta}{\Gamma \vdash \neg\neg\varphi, \Delta}$ R-FALSE $\neg$
- $\dfrac{\Gamma, \neg\varphi \vdash \Delta \quad \Gamma', \neg\psi \vdash \Delta'}{\Gamma, \Gamma', \neg(\varphi \wedge \psi) \vdash \Delta, \Delta'}$ L-FALSE $\wedge$
- $\dfrac{\Gamma \vdash \neg\varphi, \neg\psi, \Delta}{\Gamma \vdash \neg(\varphi \wedge \psi), \Delta}$ R-FALSE $\wedge$

Finally, we add to $\mathbf{rL}^+$ the following *epistemic rules*:

- $\dfrac{\Gamma, \varphi \vdash \Delta}{\Gamma, \Box\varphi \vdash \Delta}$ L-TRUE $\Box$
- $\dfrac{\Gamma \vdash \varphi, \neg\Delta}{\Box\Gamma \vdash \Box\varphi, \neg\Box\Delta}$ R-TRUE $\Box$
- $\dfrac{\Gamma, \neg\varphi \vdash \neg\Delta}{\Box\Gamma, \neg\Box\varphi \vdash \neg\Box\Delta}$ L-FALSE $\Box$
- $\dfrac{\Gamma \vdash \Box\varphi, \Delta}{\Gamma \vdash \Box\Box\varphi, \Delta}$ $4_\Box$
- $\dfrac{\Gamma, \Box\varphi \vdash \Delta}{\Gamma, \neg\Box\neg\Box\varphi \vdash \Delta}$ $5_\Diamond$
- $\dfrac{\Gamma \vdash \neg\Box\varphi, \Delta}{\Gamma \vdash \Box\neg\Box\varphi, \Delta}$ $5_\Box$

The rule L-TRUE $\bot$ ($\Gamma, \bot \vdash \Delta$) is derivable in $\mathbf{L}$. On the other hand, the rule R-TRUE $\neg$ is not derivable: adding such a rule to $\mathbf{rL}^+$ would yield a sequent system for classical propositional logic! (Cf. [Thi92]).

Definition 2.11 The rules above constitute the system $\mathbf{L}$ and are thus called $\mathbf{L}$-*rules*. A sequence $\Delta \subseteq \mathcal{L}$ is said to be $\mathbf{L}$-*derivable* from another sequence $\Gamma \subseteq \mathcal{L}$, $\Gamma \vdash_{\mathbf{L}} \Delta$, if $\Gamma \vdash \Delta$ can be derived after a finite number of applications of $\mathbf{L}$-rules. We usually drop the subscript '$\mathbf{L}$' in the sequel. Then, two formulas $\varphi, \psi \in \mathcal{L}$ are said to be equivalent, $\varphi \dashv\vdash \psi$, if $\varphi \vdash \psi$ and $\psi \vdash \varphi$.

Derivable sequents are at least valid on balloon models:

Lemma 2.12 (Soundness) For all $\Gamma, \Delta \subseteq \mathcal{L}$: $\Gamma \vdash \Delta \Rightarrow \Gamma \models \Delta$.

Proof: We prove soundness of L-TRUE $\Box$ and R-TRUE $\Box$. To start with L-TRUE $\Box$, suppose that $\Gamma, \varphi \models \Delta$. This means that for arbitrary balloon models M we have $M \models \Gamma \cup \{\varphi\} \Rightarrow \exists\delta \in \Delta, M \models \delta$ (*).

To prove $\Gamma, \Box\varphi \models \Delta$, suppose that $N = \langle W, g, V\rangle$ is an arbitrary balloon model for which $N \models \Gamma \cup \{\Box\varphi\}$. This means that both $N_g \models \Gamma$ and $N_g \models \Box\varphi$. By definition of balloon model, there is some $w \in W$ with $V(w) \sqsubseteq V(g)$. Since $N_g \models \Box\varphi$, we have for this w that $N_w \models \varphi$. Now by internal persistence we conclude that $N_g \models \varphi$. Thus we have $N_g \models \Gamma \cup \{\varphi\}$, and applying (*), we get $N_g \models \delta$, for some $\delta \in \Delta$.

To prove soundness of R-TRUE $\Box$, suppose that $\Gamma \models \varphi, \neg\Delta$. Let M be a model $\langle W, g, V\rangle$ such that $M \models \Box\Gamma$. So $M_w \models \Gamma$ for all $w \in W$. Now suppose $M \not\models \Box\varphi$, then, for some $u \in W$ we have $M_u \not\models \varphi$. Since $\Gamma \models \varphi, \neg\Delta$, we have $M_u \models \neg\delta$ for certain $\delta \in \Delta$, and hence $M \models \neg\Box\delta$. Therefore $M \models \Box\varphi, \neg\Box\Delta$ for an arbitrary balloon model M verifying $\Box\Gamma$, hence $\Box\Gamma \models \Box\varphi, \neg\Box\Delta$. ∎

Let us pause for a moment and reflect on our basic logic. Claims below that some sequents are not derivable, are now easily verified semantically, as is justified by lemma 2.12.

- The first thing to note about the logic is that it is indeed partial, which is mirrored by the fact that we do not have the *law of excluded middle*:

$$\not\vdash \varphi, \neg\varphi$$

  In fact, as is shown in [Thi92], there is not any theorem of **L** in the $\{\bot, \top\}$-free language.

- Moreover, we do not have *contraposition*:

$$\Gamma \vdash \Delta \not\Rightarrow \neg\Delta \vdash \neg\Gamma$$

- Although **L** lacks contraposition and does not have any $\{\bot, \top\}$-free theorems, there are the following *propositional equivalences:*

  De Morgan: $\neg(\varphi \wedge \psi) \dashv\vdash \neg\varphi \vee \neg\psi \qquad \neg(\varphi \vee \psi) \dashv\vdash \neg\varphi \wedge \neg\psi$

  Double negation: $\neg\neg\varphi \dashv\vdash \varphi$

  Distribution: $\varphi \wedge (\psi \vee \chi) \dashv\vdash (\varphi \wedge \psi) \vee (\varphi \wedge \chi)$
  $\varphi \vee (\psi \wedge \chi) \dashv\vdash (\varphi \vee \psi) \wedge (\varphi \vee \chi)$

  Associativity: $\varphi \wedge (\psi \wedge \chi) \dashv\vdash (\varphi \wedge \psi) \wedge \chi \qquad \varphi \vee (\psi \vee \chi) \dashv\vdash (\varphi \vee \psi) \vee \chi$

  Idempotence: $\varphi \wedge \varphi \dashv\vdash \varphi \qquad \varphi \vee \varphi \dashv\vdash \varphi$

  Commutativity: $\varphi \wedge \psi \dashv\vdash \psi \wedge \varphi \qquad \varphi \vee \psi \dashv\vdash \psi \vee \varphi$

  Absorption: $\varphi \vee (\varphi \wedge \psi) \dashv\vdash \varphi \qquad \varphi \wedge (\varphi \vee \psi) \dashv\vdash \varphi$

- For the defined symbols one easily proves the following *derived rules:*

$$\Gamma, \neg\top \vdash \Delta \quad \text{L-FALSE } \top \qquad\qquad \Gamma \vdash \top, \Delta \quad \text{R-TRUE } \top$$

$$\frac{\Gamma, \varphi \vdash \Delta \quad \Gamma', \psi \vdash \Delta'}{\Gamma, \Gamma', \varphi \vee \psi \vdash \Delta, \Delta'} \text{ L-TRUE } \vee \qquad \frac{\Gamma \vdash \varphi, \psi, \Delta}{\Gamma \vdash \varphi \vee \psi, \Delta} \text{ R-TRUE } \vee$$

$$\frac{\Gamma, \neg\varphi, \neg\psi \vdash \Delta}{\Gamma, \neg(\varphi \vee \psi) \vdash \Delta} \text{ L-FALSE } \vee \qquad \frac{\Gamma \vdash \neg\varphi, \Delta \quad \Gamma' \vdash \neg\psi, \Delta'}{\Gamma, \Gamma' \vdash \neg(\varphi \vee \psi), \Delta, \Delta'} \text{ R-FALSE } \vee$$

$$\frac{\Gamma, \varphi \vdash \Delta}{\Box\Gamma, \Diamond\varphi \vdash \Diamond\Delta} \text{ L-TRUE } \Diamond \qquad \frac{\Gamma \vdash \neg\varphi, \Delta}{\Box\Gamma \vdash \neg\Diamond\varphi, \Diamond\Delta} \text{ R-FALSE } \Diamond$$

- **L** has the following distribution property:

$$\Box\varphi \wedge \Box\psi \dashv\vdash \Box(\varphi \wedge \psi)$$

- For the *epistemic part*, we have the following:

  Positive introspection: $\Box\varphi \vdash \Box\Box\varphi \quad \Diamond\Diamond\varphi \vdash \Diamond\varphi$

  Negative introspection: $\neg\Box\varphi \vdash \Box\neg\Box\varphi \quad \Diamond\Box\varphi \vdash \Box\varphi$

  Veridicality: $\Box\varphi \vdash \varphi \quad \varphi \not\vdash \Diamond\varphi$!

Note that, although we *do* have veridicality of knowledge ('known facts are true') we got rid of its contrapositive ('true facts are considered to be possible'). Negative introspection is now better motivated than in classical **S5**: if some fact is considered possible by the agent, it is explicitly present in his set of alternatives, so he knows that particular possibility. This should be contrasted to the classical case where merely not knowing the opposite is supposed to involve knowledge of the possibility. In the sequel, we will denote a property like positive introspection by '$\Box \Rightarrow \Box\Box$' or '$\Diamond\Diamond \Rightarrow \Diamond$'.

To see the system **L** at work, we will provide a proof of the property that nestings of modal operators are in fact superfluous (theorem 2.16). Later on, this property will be used in our completeness proof. To start some preliminary work, let us first define the *modal depth* $md(\varphi)$ of a formula φ by: $md(p) = md(\bot) = 0$ $(p \in \mathcal{P})$; $md(\neg\varphi) = md(\varphi)$; $md(\varphi \wedge \psi) = \max(md(\varphi), md(\psi))$; $md(\Box\varphi) = 1 + md(\varphi)$. Secondly, by using the propositional equivalences as stated above and treating formulas like $\Box\alpha$ and $\Diamond\beta$ as literals, we obtain the following *normal forms* in **L** :

Proposition 2.13 Every $\varphi \in \mathcal{L}$ is equivalent with a formula of the form

$$\bigvee_{i=1}^{n} \bigwedge_{j=1}^{m} \varphi_{i,j}$$

where each $\varphi_{i,j}$ is of the form $\Box\alpha$ with $md(\alpha) < md(\varphi)$, $\Diamond\beta$ with $md(\beta) < md(\varphi)$, $\neg p$ or p. If $n = 0$, we interpret φ as $\bot$ (the 'empty' disjunction) and if $n > 0$ but $m = 0$, φ is to be understood as $\top$. We call the format displayed above a *semi-disjunctive normal form* of φ. There also exists a *semi-conjunctive normal form:*

$$\bigwedge_{i=1}^{n} \bigvee_{j=1}^{m} \varphi_{i,j}$$

where the formulas $\varphi_{i,j}$ are of the same form as above.

The following proposition is the heart of theorem 2.16: it explains how nesting of modal operators are removed.

Proposition 2.14 We have:

1. $\Box(\Box\alpha \vee \psi) \dashv\vdash \Box\alpha \vee \Box\psi$
2. $\Diamond(\Box\alpha \wedge \psi) \dashv\vdash \Box\alpha \wedge \Diamond\psi$

Proof: We only prove the first equivalence, the second is similar.

1.	$\Box\alpha \vdash \Box\alpha$	START	1.	$\psi \vdash \Box\alpha, \psi$	START
2.	$\psi \vdash \psi$	START	2.	$\psi \vdash \Box\alpha \vee \psi$	R-TRUE $\vee$ (1)
3.	$\Box\alpha \vee \psi \vdash \Box\alpha, \psi$	L-TRUE $\vee$ (1,2)	3.	$\Box\psi \vdash \Box(\Box\alpha \vee \psi)$	R-TRUE $\Box$ (2)
4.	$\Box(\Box\alpha \vee \psi) \vdash \Diamond\Box\alpha, \Box\psi$	R-TRUE $\Box$ (3)	4.	$\Box\alpha \vdash \Box\alpha, \psi$	START
5.	$\Diamond\Box\alpha \vdash \Box\alpha$	$\Diamond\Box \Rightarrow \Box$	5.	$\Box\alpha \vdash \Box\alpha \vee \psi$	R-TRUE $\vee$ (4)
6.	$\Box(\Box\alpha \vee \psi) \vdash \Box\alpha, \Box\psi$	CUT (4,5)	6.	$\Box\Box\alpha \vdash \Box(\Box\alpha \vee \psi)$	R-TRUE $\Box$ (5)
7.	$\Box(\Box\alpha \vee \psi) \vdash \Box\alpha \vee \Box\psi$	R-TRUE $\vee$ (6)	7.	$\Box\alpha \vdash \Box\Box\alpha$	$\Box \Rightarrow \Box\Box$
			8.	$\Box\alpha \vdash \Box(\Box\alpha \vee \psi)$	CUT (6,7)
			9.	$\Box\alpha \vee \Box\psi \vdash \Box(\Box\alpha \vee \psi)$	L-TRUE $\vee$ (3,8)

■

Corollary 2.15

$$\Box(\bigvee_{i=1}^{n} \Box\alpha_i \vee \bigvee_{j=1}^{m} \Diamond\beta_j \vee \pi) \dashv\vdash \bigvee_{i=1}^{n} \Box\alpha_i \vee \bigvee_{j=1}^{m} \Diamond\beta_j \vee \Box\pi$$

$$\Diamond(\bigwedge_{i=1}^{n} \Box\alpha_i \wedge \bigwedge_{j=1}^{m} \Diamond\beta_j \wedge \pi) \dashv\vdash \bigwedge_{i=1}^{n} \Box\alpha_i \wedge \bigwedge_{j=1}^{m} \Diamond\beta_j \wedge \Diamond\pi$$

By means of these preliminaries we now easily establish:

Theorem 2.16 Every $\varphi \in \mathcal{L}$ is equivalent with a formula φ' with $md(\varphi') \leq 1$.

Proof: The proof runs by induction on the modal depth of formulas. Obviously the result for the basic step is 'for free'.

Let the modal depth of φ be larger than 1. By purely propositional reasoning φ is equivalent with the following semi-disjunctive normal form

$$\bigvee_{i=1}^{n}(\bigwedge_{j=1}^{m} \Box\alpha_{i,j} \wedge \bigwedge_{k=1}^{l} \Diamond\beta_{i,k} \wedge \pi_i)$$

with $\pi \in \mathcal{L}_0$, $md(\alpha_{i,j}) < md(\varphi)$ and $md(\beta_{i,k}) < md(\varphi)$.

The induction hypothesis applies to all the formulas $\alpha_{i,j}$ and $\beta_{i,k}$. This means that these formulas can be assumed to have a modal depth at most 1. If this result can also be obtained for $\Box\alpha_{i,j}$ and $\Diamond\beta_{i,k}$, then the result has been shown for the formula φ. This result can be obtained quite easily by using proposition 2.14, together with $\Box(\alpha \wedge \alpha') \dashv\vdash (\Box\alpha \wedge \Box\alpha')$ and $\Diamond(\beta \vee \beta') \dashv\vdash (\Diamond\beta \vee \Diamond\beta')$. If α has modal depth 1 it must be equivalent with a semi-conjunctive normal form

$$\alpha \dashv\vdash \bigwedge_{i=1}^{n}(\bigvee_{j=1}^{m} \Box\sigma_{i,j} \vee \bigvee_{k=1}^{l} \Diamond\eta_{i,k} \vee \pi_i) \text{ with } \pi_i, \sigma_{i,j}, \eta_{i,k} \in \mathcal{L}_0,$$

while each β is equivalent to the semi-disjunctive normal form:

$$\beta \dashv\vdash \bigvee_{i=1}^{n}(\bigwedge_{j=1}^{m} \Box\epsilon_{i,j} \wedge \bigwedge_{k=1}^{l} \Diamond\lambda_{i,k} \wedge \varrho_i) \text{ with } \varrho_i, \epsilon_{i,j}, \lambda_{i,k} \in \mathcal{L}_0.$$

Corollary 2.15 yields, after applying sc r-true $\Box$ and $\Box$-distribution over $\wedge$ for the α's and β's in the semi-disjunctive normal form of φ:

$\Box\alpha \dashv\vdash \bigwedge_{i=1}^{n}(\bigvee_{j=1}^{m} \Box\sigma_{i,j} \vee \bigvee_{k=1}^{l} \Diamond\eta_{i,k} \vee \Box\pi_i)$, and

$\Diamond\beta \dashv\vdash \bigvee_{i=1}^{n}(\bigwedge_{j=1}^{m} \Box\epsilon_{i,j} \wedge \bigwedge_{k=1}^{l} \Diamond\lambda_{i,k} \wedge \Diamond\pi_i)$.

Clearly the latter formulas have a modal depth not larger than 1. ■

Combined with proposition 2.13, this theorem also implies that every formula has a semi-disjunctive and a semi-conjunctive normal form of at most modal depth 1.

2.3 Completeness

The aim of this section is to prove that the logic **L** is complete for the class of balloon models. By definition 2.4 our models are *finite*; as a consequence, not each consistent set will be satisfiable (e.g. $\{\Diamond(p_1 \wedge \cdots \wedge p_{n-1} \wedge \neg p_n) \mid n \in I\!N\}$ has only infinite models). We *can* guarantee satisfiability of *finite* sets. However, this requirement can be eased a little: what we can prove is that $\Gamma \vdash \Delta \Rightarrow \Gamma \models \Delta$ for those Γ and Δ for which the set of atoms in $\Gamma \cup \Delta$ is finite. To avoid cumbersome notation, from now on we simply assume that $\mathcal{P}$ itself is finite. We first show that this assumption implies that **L** is *logically finite*.

Proposition 2.17
L is logically finite: there are only finitely many non-equivalent formulas.

Proof: From the proof of theorem 2.16 we learn that every formula in $\mathcal{L}$ is equivalent to a semi-disjunctive normal form of modal degree ≤ 1. Since $\mathcal{P}$ is assumed to be finite, modulo logical equivalence there are only finitely many distinct formulas in $\mathcal{L}_0$. Thus there are only finitely many logically different choices for the $\alpha_{i,j}$, $\beta_{i,k}$ and π_i in the semi-disjunctive normal form displayed on page 9. Therefore there are only finitely many non-equivalent formulas. ∎

Now we are ready to give a Henkin-type construction of a canonical balloon model, based on consistent sets of formulas. However, instead of working with *maximally consistent sets*, we build such a model out of *consistent, disjunction-saturated, deductively closed theories.*

Definition 2.18 Let $\Sigma \subseteq \mathcal{L}$ and $\varphi, \psi \in \mathcal{L}$. Then:

- Σ is *consistent* iff $\Sigma \not\vdash \varphi \wedge \neg\varphi$ for all φ.
- Σ is a (deductively closed) *theory* iff $\Sigma \vdash \varphi \Rightarrow \varphi \in \Sigma$ for all φ.
- Σ is *disjunction-saturated* iff $\Sigma \vdash \varphi \vee \psi \Rightarrow \Sigma \vdash \varphi$ or $\Sigma \vdash \psi$ for all φ and ψ.

Using our sequent calculus, we have an elegant characterization of consistent disjunction-saturated theories:

Definition 2.19 Let Σ, Δ and Ω be subsets of $\mathcal{L}$. Then:

- $\Sigma \subseteq \mathcal{L}$ is *saturated* iff for every Δ: $\Sigma \vdash \Delta \Rightarrow \Sigma \cap \Delta \neq \emptyset$. $\mathcal{SAT}$ is the collection of all saturated sets (in **L**).
- Ω is a *saturator* of Σ iff $\Omega \cap \Delta \neq \emptyset$ for all Δ such that $\Sigma \vdash \Delta$. In such a case we write $\Sigma \trianglelefteq \Omega$.

Lemma 2.20

1. A set Γ is saturated iff it is a consistent disjunction-saturated theory.
2. $\Sigma \trianglelefteq \Omega$ iff $\Sigma \not\vdash \overline{\Omega}$.

Proof:

1. We only argue that a saturated set is consistent: suppose that Γ is not consistent, i.e. we have that for some formula φ, $\Gamma \vdash \varphi \wedge \neg\varphi$. Then:

1	$\Gamma \vdash \varphi \wedge \neg\varphi$	assumption
2	$\varphi \vdash \varphi$	START
3	$\varphi, \neg\varphi \vdash \emptyset$	L-TRUE $\neg$, 2
4	$\varphi \wedge \neg\varphi \vdash \emptyset$	L-TRUE $\wedge$, 3
5	$\Gamma \vdash \emptyset$	CUT, 1, 4

   So Γ cannot be saturated, since this would imply that $\Gamma \cap \emptyset \neq \emptyset$, which is impossible.

2. $\Sigma \trianglelefteq \Omega$ iff Ω is a saturator of Σ iff for all $\Delta \subseteq \mathcal{L} : (\Omega \cap \Delta = \emptyset \Rightarrow \Sigma \not\vdash \Delta)$ iff for all $\Delta \subseteq \mathcal{L} : (\Delta \subseteq \overline{\Omega} \Rightarrow \Sigma \not\vdash \Delta)$ iff (by R-MON) $\Sigma \not\vdash \overline{\Omega}$.

∎

Lemma 2.21 (Saturation Lemma)
If $\Sigma \trianglelefteq \Omega$, then there exists a saturated set Γ such that $\Sigma \subseteq \Gamma \subseteq \Omega$.

Proof: See [JT93]. ∎

The following lemma, from [Thi92, p.108], is equivalent to the Saturation lemma. It guarantees that each consistent set Σ and each set of non-consequences Δ can be separated by a saturated set $\Gamma \supseteq \Sigma$.

Lemma 2.22 (Separation Lemma)
If $\Sigma \nvdash \Delta$ then there exists a saturated set Γ such that $\Sigma \subseteq \Gamma$ and $\Delta \cap \Gamma = \emptyset$.

Proof: If $\Sigma \nvdash \Delta$ then, by lemma 2.20 (2) $\overline{\Delta}$ is a saturator of Σ. But then, the saturation lemma guarantees the existence of a saturated set Γ with $\Sigma \subseteq \Gamma \subseteq \overline{\Delta}$, i.e. a saturated set Γ with $\Sigma \subseteq \Gamma$ and $\Gamma \cap \Delta = \emptyset$. ∎

We will now build a canonical model for **L** . Whereas in classical modal logic the canonical worlds are maximally consistent sets, in partial logic this role is taken over by saturated sets.

Definition 2.23 (Canonical Model)
Let Γ be a saturated set. We define the canonical model for Γ as $\mathcal{M}_\Gamma = \langle \mathcal{W}_\Gamma, \Gamma, \mathcal{V} \rangle$, where

- $\mathcal{W}_\Gamma = \{\Sigma \mid \Sigma \text{ is saturated and } \Box^-\Gamma \subseteq \Sigma \subseteq \Diamond^-\Gamma\}$
- For all $\Sigma \in \mathcal{W}_\Gamma \cup \{\Gamma\}$ and $p \in \mathcal{P}$: $\mathcal{V}(\Sigma)(p) = \begin{cases} 1 & \text{if} \quad p \in \Sigma \\ 0 & \text{if} \quad \neg p \in \Sigma \end{cases}$

Lemma 2.24 The canonical model $\mathcal{M}_\Gamma$ is a balloon model.

Proof: (Cf. Definition 2.4)

1. $\mathcal{W}$ is finite by proposition 2.17[2]
2. $\mathcal{V}$ is well-defined since saturated sets are consistent (lemma 2.20).
3. The root Γ is an extension of some world in the balloon, i.e. for some $\Sigma \in \mathcal{W}_\Gamma$ it holds that $\mathcal{V}(\Sigma) \sqsubseteq \mathcal{V}(\Gamma)$. To see this, we claim that $\Box^-\Gamma \trianglelefteq \Omega = \Diamond^-\Gamma \cap \Gamma$; then we are done, since then (by the saturation lemma) there is a saturated Σ such that $\Box^-\Gamma \subseteq \Sigma \subseteq \Diamond^-\Gamma \cap \Gamma$. Therefore $\Sigma \in \mathcal{W}_\Gamma$ and $\Sigma \subseteq \Gamma$, and so $\mathcal{V}(\Sigma) \sqsubseteq \mathcal{V}(\Gamma)$. The proof of the claim about Ω is as follows.

   By induction on finite $\Lambda \subseteq \mathcal{L}$ we prove that

   $$\Box^-\Gamma \vdash \Lambda \Rightarrow \Lambda \cap \Gamma \cap \Diamond^-\Gamma \neq \emptyset.$$

   Because $\Box^-\Gamma \nvdash \emptyset$ [3], the implication above holds in a trivial way for $\Lambda = \emptyset$. So suppose $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$, $(n \geq 1)$, and $\Box^-\Gamma \vdash \Lambda$. Then, by $n-1$ applications of R-TRUE $\vee$ we have, for each $(i \leq n)$, $\Box^-\Gamma \vdash (\lambda_1 \vee \cdots \vee \lambda_{i-1} \vee \lambda_{i+1} \vee \cdots \vee \lambda_n), \lambda_i$ and hence, by using R-TRUE $\Box$, we obtain

   $$\forall i \leq n : \Gamma \vdash \Box(\lambda_1 \vee \cdots \vee \lambda_{i-1} \vee \lambda_{i+1} \vee \cdots \vee \lambda_n), \Diamond\lambda_i$$

   Since Γ is saturated, we have two possibilities:

[2] Notice this is virtually the only place where the specific (introspection) rules of **L** are used in the completeness proof. I.e. these rules license the special form of our balloon models.

[3] $\Box^-\Gamma \vdash \emptyset \Rightarrow \Box^-\Gamma \vdash \bot \Rightarrow \Gamma \vdash \Box\bot \Rightarrow \Gamma \vdash \bot$.

- For some $i \leq n$ $\Box(\lambda_1 \vee \cdots \vee \lambda_{i-1} \vee \lambda_{i+1} \vee \cdots \vee \lambda_n) \in \Gamma$. Then $\Box^-\Gamma \vdash \Lambda \setminus \{\lambda_i\}$ and, by the induction hypothesis, $\Lambda \setminus \{\lambda_i\} \cap \Gamma \cap \Diamond^-\Gamma \neq \emptyset$, and hence $\Lambda \cap \Gamma \cap \Diamond^-\Gamma \neq \emptyset$.
- For all $i \leq n, \Diamond\lambda_i \in \Gamma$. Then $\Lambda \subseteq \Diamond^-\Gamma$ (a), and, since $\Box^-\Gamma \vdash \Lambda$, by the L-TRUE $\Box$-rule, we have $\Gamma \vdash \Lambda$ and hence, by saturation of Γ, $\Gamma \cap \Lambda \neq \emptyset$ (b). Combining (a) and (b), we obtain $\Lambda \cap \Gamma \cap \Diamond^-\Gamma \neq \emptyset$.

■

Lemma 2.25 (Truth Lemma)
For all formulas $\varphi \in \mathcal{L}$, and all sets $\Gamma \in \mathcal{SAT}$ and each canonical model $\mathcal{M}_\Gamma$:

$$\mathcal{M}_\Gamma \models \varphi \Leftrightarrow \varphi \in \Gamma \qquad \mathcal{M}_\Gamma \dashv \varphi \Leftrightarrow \neg\varphi \in \Gamma$$

Proof: By induction on φ: we only give the $\Box$-step. So we assume that $\varphi = \Box\psi$, while the induction hypothesis (IH) says that the lemma holds for ψ.
First we show the equivalence for $\models$:

($\Rightarrow$) If $\mathcal{M}_\Gamma \models \Box\psi$, then, by the truth definition of $\Box$, for all $\Delta \in \mathcal{W}_\Gamma : \mathcal{M}_\Delta \models \psi$. By IH, we conclude that for all $\Delta \in \mathcal{W}_\Gamma, \psi \in \Delta$. Now consider

$$\Gamma \vdash \Box\psi, \Diamond\overline{\Diamond^-\Gamma} \qquad (*)$$

After observing that $\Diamond\overline{\Diamond^-\Gamma} = \{\Diamond\gamma \mid \Diamond\gamma \notin \Gamma\}$, we claim that $(*)$ holds: for, suppose not, then by R-TRUE $\Box$ and L-MON we also have $\Box^-\Gamma \not\vdash \psi, \overline{\Diamond^-\Gamma}$ and we use the separation lemma to find a Δ for which $\Box^-\Gamma \subseteq \Delta \subseteq \overline{\Diamond^-\Gamma}$, and $\psi \notin \Delta$, contradicting IH. Thus, since $(*)$ holds, we may use saturation of Γ to conclude that either $\Box\psi \in \Gamma$ or $\Gamma \cap \Diamond\overline{\Diamond^-\Gamma} \neq \emptyset$. Since the latter is impossible, we conclude that $\Box\psi \in \Gamma$.

($\Leftarrow$) Suppose $\Box\psi \in \Gamma$, and choose $\Delta \in \mathcal{W}_\Gamma$ which means that $\Delta \in \mathcal{SAT}$ and $\Box^-\Gamma \subseteq \Delta \subseteq \Diamond^-\Gamma$. We immediately find $\psi \in \Delta$ and, by IH, $\mathcal{M}_\Delta \models \psi$ so that $\mathcal{M}_\Gamma \models \Box\psi$.

Next the steps for $\dashv$ are:

($\Rightarrow$) If $\mathcal{M}_\Gamma \dashv \Box\psi$, then, by the falsity condition for $\Box$, for some $\Delta \in \mathcal{W}_\Gamma : \mathcal{M}_\Delta \dashv \psi$, and, using IH, $\neg\psi \in \Delta$. Since $\Delta \subseteq \Diamond^-\Gamma$, $\Diamond\neg\psi \in \Gamma$, and, since Γ is deductively closed, $\neg\Box\psi \in \Gamma$.

($\Leftarrow$) Suppose $\neg\Box\psi \in \Gamma$. We claim that $\Box^-\Gamma \cup \{\neg\psi\} \trianglelefteq \Diamond^-\Gamma$. To see this, suppose that Θ is such that $\Box^-\Gamma \cup \{\neg\psi\} \vdash \Theta$, then, by L-FALSE $\Box$, also $\Box\Box^-\Gamma, \neg\Box\psi \vdash \Diamond\Theta$ and, by monotonicity, $\Gamma, \neg\Box\psi \vdash \Diamond\Theta$. Since $\neg\Box\psi$ is already a member of Γ, this implies $\Gamma \vdash \Diamond\Theta$. Now we use saturation of Γ to find a formula $\Diamond\theta$ in $\Gamma \cap \Diamond\Theta$, so $\theta \in \Diamond^-\Gamma \cap \Theta$. Now we have proven the claim, we use the saturation lemma to obtain a saturated set Δ with $\Box^-\Gamma \cup \{\neg\psi\} \subseteq \Delta \subseteq \Diamond^-\Gamma$. Clearly, $\Delta \in W_\Gamma, \neg\psi \in \Delta$, so we apply IH to conclude $\mathcal{M}_\Delta \dashv \psi$. ■

Theorem 2.26 (Completeness) For all $\Sigma, \Delta \subseteq \mathcal{L}$, $\Sigma \models \Delta \Rightarrow \Sigma \vdash \Delta$.

Proof: Suppose $\Sigma \not\vdash \Delta$, then, using the separation lemma we obtain a saturated set Γ for which $\Sigma \subseteq \Gamma$ and $\Gamma \cap \Delta = \emptyset$. Clearly, by lemma 2.25, $\mathcal{M}_\Gamma \models \Sigma$ and $\mathcal{M}_\Gamma \not\models \delta$ for all $\delta \in \Delta$, hence $\Sigma \not\models \Delta$ ■

Corollary 2.27 For all Σ and Δ, $\Sigma \vdash \Delta \Leftrightarrow \Sigma \models \Delta$.

3 Honesty

This section concerns both the 'syntactic' and 'semantic' view on *circumscription* and *honesty*. Circumscribing the knowledge expressed by, say, φ, is to characterize what a rational agent knows when (s)he *only* knows φ (together with its logical consequences). If such circumscription is possible, φ is called 'honest' in [HM85]. Though it may seem, *prima facie*, that circumscribing φ is always possible (by taking, e.g., the deductive closure of φ), this need not be the case. For example, the formula $\varphi = \Box p \vee \Box \neg p$ cannot be circumscribed (and is, hence, *dishonest*): only knowing φ implies not knowing more than that, in particular, not knowing p and not knowing $\neg p$. However, the latter two conclusions, combined with φ, lead to an inconsistency.

The main issue we want to address in this section, is to decide which formulas φ can be rendered honest. We will in fact present several notions of honesty in section 3.1, and illustrate them using several examples. Most of the technical justification for these examples, (in particular, examples 3.6, 3.12 and 3.16) as well as for observation 3.14 is provided after we have given a semantic account of the various notions of honesty, and are therefore postponed until section 3.2. In section 3.3 we connect the semantic view on honesty with a syntactic one.

3.1 Stable Sets

We start out by investigating the deductive view on circumscription and honesty. Which criteria does the set $C_{\Box\varphi}$ consisting of consequences of $\Box\varphi$ have to meet to consider φ honest? The crucial notion here is that of a *stable set*[4]. Although stability can be defined in many ways, the notion itself is stable, since various definitions turn out to be equivalent.

Thinking of $C_{\Box\varphi}$ as the 'epistemic state' (in terms of [HM85]) of a rational agent knowing only φ, it is clear that a stable set at least has to be a *consistent theory* (Cf. definition 2.18). In addition to being a consistent theory we want a stable set to have the property that the ignorance of non-consequences is compatible with the knowledge of consequences. In [Moo85] and [Jas91b] this leads to the following requirements for a stable set with respect to a normal modal system:

- S is a theory
- $\Box S \cup \neg\Box\overline{S}$ is consistent

Though correct for normal systems, the latter requirement is too strong for the partial logic we advocate. Recall from section 2.2 that our logic does not have any $\{\top, \bot\}$-free theorems. We want to exploit this property by allowing the set $S = C_{\Box\top}$ to be stable, characterizing the epistemic state of an agent knowing nothing. However, S is unstable by the second requirement: since $\Box\top \not\vdash (p \vee \neg p)$, we have that $(p \vee \neg p) \in \overline{S}$, and therefore $\{\Box\top, \neg\Box(p \vee \neg p)\}$ would be consistent, which it is not. So we propose to replace the requirement above by the more general condition that knowledge of non-consequences does not follow from the initial knowledge.

Definition 3.1 (Stability)
A set S of formulas is *stable* iff S is a theory for which $\Box S \not\vdash \Box\overline{S}$

Notice that stable sets are consistent: suppose that S is an inconsistent theory. By the rule L-TRUE $\bot$ and the theoricity of S we have $S = \mathcal{L}$ and hence $\Box\overline{S} = \emptyset$. The inconsistency of S implies that of $\Box S$, so, by the proof of lemma 2.20, we have $\Box S \vdash \emptyset$, which means $\Box S \vdash \Box\overline{S}$, implying that S is not stable.

The insightful but somewhat esoteric definition 3.1 can be recast in a format which is closer to Stalnaker's original formulation:

[4] See [Sta], [Moo85], [HM85] for **S5** stability; [Jas91b] defines stability for arbitrary normal systems. Our text definition is from [Thi92].

Proposition 3.2
S is a stable set of formulas iff

1. S is a theory
2. if $\varphi \in S$ then $\Box\varphi \in S$ (positive introspection)
3. if $\Box\varphi \vee \Box\psi \in S$ then $\varphi \in S$ or $\psi \in S$ (modal saturation)
4. $\varphi \notin S$ for some φ

Lemma 3.3 (modal saturation)
For all consistent theories S modal saturation is equivalent to

$$S \vdash \Box\Gamma \Rightarrow S \cap \Gamma \neq \emptyset \text{ for all } \Gamma \subseteq \mathcal{L}$$

Proof: Modal saturation is obviously implied by the above requirement. For the other direction, suppose that $S \vdash \Box\Gamma$. First note that the consistency of S implies that $\Gamma \neq \emptyset$. By the finiteness of **L** we may assume that Γ is finite, say $\{\gamma_1, \ldots, \gamma_n\}$. So $S \vdash \Box\gamma_1 \vee \cdots \vee \Box\gamma_n$, therefore (by corollary 2.15) $S \vdash \Box(\Box\gamma_1 \vee \cdots \vee \Box\gamma_n)$, and thus (by proposition 2.14) $S \vdash \Box\gamma_1 \vee \Box(\Box\gamma_2 \vee \cdots \vee \Box\gamma_n)$. Therefore, by modal saturation, $\gamma_1 \in S$ or $\Box\gamma_2 \vee \cdots \vee \Box\gamma_n$. Repeating this argument it follows that for some i, $\gamma_i \in S$. ∎

Because of this equivalence, we will also refer to the elegant property displayed in lemma 3.3 by 'modal saturation'.

Proof of proposition 3.2:

($\Rightarrow$) Let S be a stable set. Then

1. by definition, S is a theory
2. suppose $\varphi \in S$ and $\Box\varphi \notin S$ then, since $\Box\varphi \vdash \Box\Box\varphi$, S violates the $\nvdash$ condition.
3. suppose for some Γ, we have $S \vdash \Box\Gamma$, and $S \cap \Gamma = \emptyset$, then $\Gamma \subseteq \overline{S}$, so $\Box\Gamma \subseteq \Box\overline{S}$, so that, by monotonicity, $S \vdash \Box\overline{S}$ and, by L-TRUE, $\Box S \vdash \Box\overline{S}$, contradicting the stability of S.
4. S is consistent, so $p \wedge \neg p \notin S$.

($\Leftarrow$) Next let S satisfy the conditions (1–4) then S is a theory, thus (by 4 and *ex falso*) consistent. Suppose $\Box S \vdash \Box\overline{S}$. By (2) $\Box S \subseteq S$, so by MON $S \vdash \Box\overline{S}$. Lemma 3.3 tells us that $S \cap \overline{S} \neq \emptyset$, a contradiction. ∎

Although the characterization of stability given by proposition 3.2 is useful, sometimes a more concise requirement is convenient.

Proposition 3.4 S is stable iff $S = \Box^{-}\Gamma$ for some $\Gamma \in \mathcal{SAT}$.

Proof: ($\Rightarrow$) Let S be stable, then $\Box S \nvdash \Box\overline{S}$. By our Separation Lemma there is a saturated set Γ such that (i) $\Box S \subseteq \Gamma$ and (ii) $\Gamma \cap \Box\overline{S} = \emptyset$. Then $\varphi \in S \Rightarrow$ (by i) $\Box\varphi \in \Gamma \Rightarrow \varphi \in \Box^{-}\Gamma$ and $\varphi \notin S \Rightarrow \Box\varphi \in \Box\overline{S} \Rightarrow$ (by ii) $\Box\varphi \notin \Gamma \Rightarrow \varphi \notin \Box^{-}\Gamma$. Hence $S = \Box^{-}\Gamma$. ($\Leftarrow$) Suppose $S = \Box^{-}\Gamma$ for some saturated set Γ, and also that $\Box S \vdash \Box\overline{S}$. Since $\Box S \subseteq \Gamma$, and using R-MON we have $\Gamma \vdash \Box\overline{S}$. Γ is saturated, and hence there is some $\psi \notin S$ with $\Gamma \vdash \Box\psi$. But then, since Γ is deductively closed (lemma 2.20), we have $\Box\psi \in \Gamma$ and hence $\psi \in S$, a contradiction. ∎

Having characterized stability in different ways, we are ready for a formal account of circumscription and honesty. If we write $\mathcal{ST}(\varphi)$ for $\{S \subseteq \mathcal{L} | \varphi \in S$ and S is stable $\}$, then circumscription of knowledge of φ involves finding a minimal element in $\mathcal{ST}(\varphi)$, the stable sets containing the initial information φ. If there is a stable set which is minimum, according to some order on sets of formulas, the knowledge is honest. What is this ordering relation? In the paradigm case of the (total) system **S5**, different stable sets are incomparable, so set inclusion does not work. This is not the case for the present (partial) system, basically because the notorious Stalnaker condition $\varphi \notin S \Rightarrow \neg\Box\varphi \in S$ does not hold for stable sets in partial logic. The invalidity of the latter condition implies that in **L** a stable set is not determined by its propositional content (the purely propositional formulas in it), although a stable set is determined by its formulas of degree 1 (i.e. with modal depth less or equal than 1), by theorem 2.16. This might suggest set inclusion as the ordering relation of the stable sets, and a definition of honesty induced by $\subseteq$: φ would then be 'stable-honest' iff there is a $\subseteq$-smallest stable expansion of $\{\varphi\}$.

Definition 3.5 (Naïve Honesty)
φ is called *naïvely honest* iff there is a $\subseteq$-minimal element in $\mathcal{ST}(\varphi)$.

Example 3.6 The formulas p, $p \wedge q$, $\Box p$, $\Box(p \wedge q)$, $\Diamond p$ and $\Diamond(p \wedge q)$ are all naïvely honest.

Can we give other sufficient and necessary conditions for naïve honesty? To this purpose reinspect $C_{\Box\varphi} = \{\psi \mid \Box\varphi \vdash \psi\}$. First observe that

- $C_{\Box\varphi}$ is a *theory*, since $\vdash$ is transitive;
- $C_{\Box\varphi}$ is contained in every stable set containing φ: by proposition 3.2(2) if $\varphi \in$ some stable S, then $\Box\varphi \in S$, so, by proposition 3.2(1) S contains all the consequences of $\Box\varphi$, i.e. $C_{\Box\varphi} \subseteq S$.

As an easy result, we now present a necessary and sufficient condition for a stable set being $\subseteq$-minimal.

Theorem 3.7 A set S is $\subseteq$-minimal in $\mathcal{ST}(\varphi)$ iff $S = C_{\Box\varphi}$ is stable.

Proof: ($\Rightarrow$) Suppose S is $\subseteq$-minimal for φ. By definition $\varphi \in S$, and, by the remark above, $C_{\Box\varphi} \subseteq S$. Now suppose that $S \not\subseteq C_{\Box\varphi}$, then we have a ψ with $\psi \in S$ and $\Box\varphi \not\vdash \psi$. The separation lemma then provides a saturated set Γ for which $\Box\varphi \in \Gamma, \psi \notin \Gamma$. Since $\Box\psi \vdash \psi$, by lemma 2.20 we also have $\Box\psi \notin \Gamma$. By proposition 3.4, $\Box^{-}\Gamma$ is a stable set containing φ, contradicting the $\subseteq$-minimality of S.
($\Leftarrow$) If $C_{\Box\varphi}$ is a stable set, by the remarks above it must be $\subseteq$-minimal. ∎

The theorem above immediately provides a necessary and sufficient condition for naïve honesty; more strict characterizations are given in section 3.3:

Corollary 3.8 φ is naïvely honest iff $C_{\Box\varphi}$ is stable.

Proof: Let $C_{\Box\varphi}$ be stable. By L-TRUE $\Box$, $\varphi \in C_{\Box\varphi}$. By theorem 3.7, $C_{\Box\varphi}$ is also $\subseteq$-minimal for φ, implying that φ is naïvely honest. The other direction is obvious. ∎

Intuition says that all *objective* (i.e. propositional) formulas should be rendered honest: it seems to be perfectly sensible to claim to only know some objective information. This is where the definition of naïve honesty is too strong (and also too naïve):

Observation 3.9 The objective formula $p \vee q$ is not naïvely honest.

Proof: Suppose that S would be $\subseteq$-minimal in $\mathcal{ST}(p \vee q)$, then $(p \vee q) \in S$, and, by proposition 3.2(2), also $\Box(p \vee q) \in S$. Since (by R-TRUE and $\Diamond \Rightarrow \Box\Diamond$)), we have $\Box(p \vee q) \vdash \Box p \vee \Box\Diamond q$, we use proposition 3.2(3) to conclude that either $p \in S$ or $\Diamond q \in S$ (*). Now, let $\Sigma_1 = \{\Box p\}$ and $\Sigma_2 = \{\Box q\}$. Using completeness, we immediately see that $\Sigma_1 \not\vdash \Box\Diamond q$ and $\Sigma_2 \not\vdash \Box p$. The separation lemma then guarantees the existence of saturated sets Γ_1, Γ_2 for which $\Sigma_i \subseteq \Gamma_i (i = 1, 2), \Box\Diamond q \notin \Gamma_1$ and $\Box p \notin \Gamma_2$. By proposition 3.4 we find two stable sets $S_i = \Box^- \Gamma_i, (i = 1, 2), S_i \in \mathcal{ST}(p \vee q)$, for which $\Diamond q \notin S_1$ and $p \notin S_2$. Since S is $\subseteq$-minimal in $\mathcal{ST}(p \vee q)$ we find $p \notin S, \Diamond q \notin S$, contradicting (*). ∎

Therefore, though the set inclusion ordering of stable sets is (non-trivially) possible, it produces wrong results as far as honesty is concerned. Now one alternative is to replace ordinary set inclusion by the relation of epistemic inclusion $\subseteq_\Box$. This, however, will not produce any new results, due to the following observation.

Observation 3.10 For all stable sets Γ, Δ: $\Gamma \subseteq_\Box \Delta \Leftrightarrow \Gamma \subseteq \Delta$.

Proof: straightforward from the definitions. ∎

Somewhat surprisingly, since its propositional content does not determine a stable set, propositional minimality of a stable expansion produces a more adequate notion of honesty.

Definition 3.11 (Weak Honesty)
φ is *weakly honest* iff there is a $\subseteq_0$-minimal element in $\mathcal{ST}(\varphi)$.

This is in fact the same definition of honesty that was proposed by Halpern Moses in [HM85]. However, in **L** one can generally derive less conclusions from the circumscription of a weakly honest formula than in the **S5** case ($\Diamond p$ for instance, is honest in [HM85] and also weakly honest in **L**, but 'knowing only $\Diamond p$' entails different conclusions in both set-ups).

Example 3.12 Naïvely honest formulas are weakly honest. The disjunction $p \vee q$ is also weakly honest: more generally, for each consistent objective formula π, π itself, $\Box\pi$ and $\Diamond\pi$ are weakly honest. Other examples are $\Box p \wedge \Diamond q$, and disjunctions like $\Box p \vee \neg\Box p$ and $p \vee \neg\Box p$. The formula $\Box p \vee \Box q$ is not weakly honest, neither is $\Box p \vee \neg p$. (This will be proved in section 3.3.)

Notice that a propositionally smallest stable expansion for some formula need not be unique: $S \cap \mathcal{L}_0$ does not determine S. For example, in the case of $p \vee q$, S may or may not contain $\Diamond(p \wedge q)$.

Theorem 3.13 A set S is $\subseteq_0$-minimal in $\mathcal{ST}(\varphi)$ iff $S \in \mathcal{ST}(\varphi)$ and $S_0 = (C_{\Box\varphi})_0 = \{\mu \in \mathcal{L}_0 \mid \Box\varphi \vdash \mu\}$.

Proof: Omitted; essentially the same as the proof of theorem 3.7. ∎

In the introduction to this section we explained why $\Box p \vee \Box q$ should not be rendered honest: although it makes perfect sense for an agent to claim that he knows that he either knows p or q, it is absurd for an agent to claim that *all* he knows is that either he knows p or he knows q, because it would imply that he neither knows p ($\Box p$ being logically stronger than $\Box p \vee \Box q$) nor q.

Essentially the same analysis can be made if the agent claims to only know some disjunction of possibilities he considers possible: if the agent only knows $\Diamond p \vee \Diamond q$, he does not know the stronger $\Diamond p$, nor $\Diamond q$, though $\Box\Diamond p \vee \Box\Diamond q$ follows in **L** from $\Box(\Diamond p \vee \Diamond q)$. And, indeed, intuition says that an agent may *know* some disjunctive information about facts for which he has some evidence, but this cannot be *all he knows*. This is why the current notion of honesty is too weak:

Observation 3.14 The formula $\Diamond p \vee \Diamond q$ is weakly honest[5].

Proof: We provide a semantical argument in observation 3.29; here we give a deductive one. It is easily seen that $\Diamond p \vee \Diamond q \not\vdash \Box\mathcal{L}_0$. By the separation lemma, there is a $\Gamma \in \mathcal{SAT}$ for which $\Diamond p \vee \Diamond q \in \Gamma$, and $\Gamma \cap \Box\mathcal{L}_0 = \emptyset$. By proposition 3.4, $\Box^-\Gamma$ is a stable set, that moreover contains $\Diamond p \vee \Diamond q$ (the latter is true by $\Diamond p \vee \Diamond q \vdash \Box(\Diamond p \vee \Diamond q)$). Since $\Box^-\Gamma \cap \mathcal{L}_0 = \emptyset$, $\Box^-\Gamma$ is obviously a $\subseteq_0$-minimal set for $\Diamond p \vee \Diamond q$. ∎

Analyzing the reasons for this observation, note that for weak honesty we did minimize the *objective* formulas in the stable set for φ, but not the *possibilities* contained in it. In fact, $\subseteq_0$-minimality is insufficiently restrictive: among the $\subseteq_0$-minimal stable sets, we want to single out those containing the least number of epistemic possibilities. This is achieved in our last notion of honesty:

Definition 3.15 **(Strong Honesty)** A formula φ is called strongly honest if there is a $\subseteq_\Diamond$-minimal element in the set $\{S \subseteq \mathcal{L} \mid S \text{ is } \subseteq_0 \text{-minimal in } \mathcal{ST}(\varphi)\ \}$.

Example 3.16 Now, $\Diamond p \vee \Diamond q$ is not strongly honest. As with weak honesty, for each objective formula $\pi \in \mathcal{L}_0$, the formulas π and $\Box\pi$ are strongly honest, but now, $\Diamond(p \vee q)$ is not strongly honest (Cf. section 3.2).

When characterizing the stable sets that contain a strongly honest formula, we need a lemma that we will not prove until section 3.2, and one more definition.

Lemma 3.17 Let S and S' be two stable sets such that $\Box^- S \subseteq_0 \Box^- S'$ and $\Diamond^- S \subseteq_0 \Diamond^- S'$. Then $S \subseteq S'$.

Proof: Postponed until corollary 3.23 ∎

Definition 3.18 For a formula ψ we define its *diamond remainder* $R^\Diamond_{\Box\varphi}$ as follows:

$$R^\Diamond_{\Box\varphi} = \{\Diamond\mu \in \Diamond(\mathcal{L}_0) \mid \Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0,\ \Diamond\mu\ \}$$

In words, $R^\Diamond_{\Box\varphi}$ contains $\Diamond$-formulas with a propositional argument, that are derivable from $\Box\varphi$, in disjunction with those $\Box$-formulas of which the argument is propositional and not a consequence of $\Box\varphi$.

Theorem 3.19 A set S is $\subseteq_\Diamond$-minimal in $\{S \subseteq \mathcal{L} \mid S \text{ is } \subseteq_0 \text{-minimal in } \mathcal{ST}(\varphi)\ \}$ iff $S_0 = (C_{\Box\varphi})_0$ and $S \cap \Diamond\mathcal{L}_0 = R^\Diamond_{\Box\varphi}$ and $S \in \mathcal{ST}(\varphi)$.

Proof:
($\Rightarrow$) Let S be $\subseteq_\Diamond$-minimal in $\{S \subseteq \mathcal{L} \mid S \text{ is } \subseteq_0 \text{-minimal in } \mathcal{ST}(\varphi)\ \}$. Clearly, S is $\subseteq_0$-minimal in $\mathcal{ST}(\varphi)$, hence, by 3.13, $S_0 = (C_{\Box\varphi})_0$. In order to show that $R^\Diamond_{\Box\varphi} \subseteq S \cap \Diamond\mathcal{L}_0$, suppose $\Diamond\mu \in R^\Diamond_{\Box\varphi}$. Then $\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\mu$. Let Γ be saturated such that $S = \Box^-\Gamma$. Then, $\Box\varphi \in \Gamma$ and, by definition of saturation, $\Gamma \cap (\Box(\overline{C_{\Box\varphi}})_0 \cup \{\Diamond\mu\}) \neq \emptyset$. Suppose $\Gamma \cap \Box(\overline{C_{\Box\varphi}})_0 \neq \emptyset$. Then there is some $\gamma \in \mathcal{L}_0$ with $\Box\gamma \in \Gamma, \Box\varphi \not\vdash \gamma$. But then $\gamma \in S$ and $\gamma \notin (C_{\Box\varphi})_0$, which contradicts the $\subseteq_0$-minimality of S (Cf. theorem 3.13). Thus, we conclude that $\Diamond\mu \in \Gamma$. Since $\Diamond\mu \vdash \Box\Diamond\mu$, we find $\Diamond\mu \in S \cap \Diamond\mathcal{L}_0$. To see that also $S \cap \Diamond\mathcal{L}_0 \subseteq R^\Diamond_{\Box\varphi}$, suppose for certain $\mu \in \mathcal{L}_0$ that both $\Diamond\mu \in S$, and $\Diamond\mu \notin R^\Diamond_{\Box\varphi}$. Then $\Box\varphi \not\vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\mu$. Using the separation lemma, we find a saturated set Γ for which $\Box\varphi \in \Gamma$ and $\Gamma \cap (\Box(\overline{C_{\Box\varphi}})_0 \cup \{\Diamond\mu\}) = \emptyset$. Clearly, $\Box^-\Gamma$ is stable for φ and also $\Box^-\Gamma \cap (\overline{C_{\Box\varphi}})_0 = \emptyset$. But then $\Box^-\Gamma \cap \mathcal{L}_0 \subseteq (C_{\Box\varphi})_0$. Since, by stability of $\Box^-\Gamma$ we also have $(C_{\Box\varphi})_0 \subseteq \Box^-\Gamma \cap \mathcal{L}_0$, we can apply theorem 3.13 to conclude that $\Box^-\Gamma$ is a $\subseteq_0$-minimal set for φ. However, this contradicts the fact that S is $\subseteq_\Diamond$-minimal amongst the $\subseteq_0$-minimal

[5] We remark that $\Diamond p \vee \Diamond q$ is also honest in the analysis of **S5** as given in [HM85].

stable sets for φ, since we have $\Diamond\mu \in S, \Diamond\mu \notin \Gamma$ hence $\Box\Diamond\mu \notin \Gamma$, thus $\Diamond\mu \notin \Box^-\Gamma$.
($\Leftarrow$) Suppose that both $S \cap \mathcal{L}_0 = (C_{\Box\varphi})_0$ and $S \cap \Diamond\mathcal{L}_0 = R^{\Diamond}_{\Box\varphi}$ for some $S \in \mathcal{ST}(\varphi)$. Then $S \cap (\mathcal{L}_0 \cup \Diamond\mathcal{L}_0) = (C_{\Box\varphi})_0 \cup R^{\Diamond}_{\Box\varphi}$. Let S' be an arbitrary stable set for φ that is $\subseteq_0$-minimal. We have to show that $S \subseteq_{\Diamond} S'$; by lemma 3.17 it is sufficient to show that both $\Box^- S \subseteq_0 \Box^- S'$ and $\Diamond^- S \subseteq_0 \Diamond^- S'$. Since $S \cap \mathcal{L}_0 = (C_{\Box\varphi})_0$, by theorem 3.13 we have $S \subseteq_0 S'$ so in particular, $\Box^- S \subseteq_0 \Box^- S'$. So suppose that we have some $\mu \in \mathcal{L}_0$ with $\Diamond\mu \in S$. Since $S \cap \Diamond\mathcal{L}_0 = R^{\Diamond}_{\Box\varphi}$, this means that $\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\mu$, and thus, $\Box S' \vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\mu$. Since S' is $\subseteq_0$-minimal for φ, we have, by theorem 3.13 that $S' \cap \mathcal{L}_0 = (C_{\Box\varphi})_0$, hence, $S' \cap (\overline{C_{\Box\varphi}})_0 = \emptyset$ and $\Gamma' \cap \Box(\overline{C_{\Box\varphi}})_0 = \emptyset$, for $S' = \Box^-\Gamma'$. But, then, since Γ' is saturated, we have $\Diamond\mu \in \Gamma'$ (by an argument similar to the one above), so that (by $\Diamond \Rightarrow \Box\Diamond$) $\Diamond\mu \in S'$. ∎

3.2 Minimal Models

Proposition 3.4 ties up the notion of stable set with a main semantic notion: recall from the construction of the canonical model, that saturated sets correspond to partial worlds. Let us make the following corollary of 3.4 relating stability directly to the knowledge in a balloon model.

Corollary 3.20 S is stable iff $S = \kappa(M)$ for some balloon model M.

In order to decide whether some formula is honest, we considered stable sets that were minimal in some sense. Combining this with corollary 3.20 gives rise to the following orders on models.

Definition 3.21 For any two models $M = \langle W, g, V\rangle$ and $M' = \langle W', g', V'\rangle$ we define:

- (Smyth order)
  $M \sqsubseteq_{\Box} M' \Leftrightarrow \forall w' \in W'\ \exists w \in W : V(w) \sqsubseteq V'(w')$
- (Hoare order)
  $M \sqsubseteq_{\Diamond} M' \Leftrightarrow \forall w \in W\ \exists w' \in W' : V(w) \sqsubseteq V'(w')$
- (Egli-Milner order)
  $M \sqsubseteq M' \Leftrightarrow M \sqsubseteq_{\Box} M'\ \&\ M \sqsubseteq_{\Diamond} M'$
- For any $\preceq \in \{\sqsubseteq_{\Box}, \sqsubseteq_{\Diamond}, \sqsubseteq\}$, we say that a model M is $\preceq$-minimal for φ if $\varphi \in \kappa(M)$ and for all M' with $\varphi \in \kappa(M')$ it holds that $M \preceq M'$. We then say that φ has a $\preceq$-minimal model.

The above orders are familiar from *domain theory*; see e.g. [Sto77]. The orders do not specify anything about the root g of a model $M = \langle W, g, V\rangle$. Recall that $Th(M) = \{\varphi \in \mathcal{L} \mid M \models \varphi\}$, that $\kappa(M) = \Box^- Th(M)$ and $\pi(M) = \Diamond^- Th(M)$. This is how the relations $\subseteq_\star$ and $\sqsubseteq_\star$ are related:

Theorem 3.22 Let $M = \langle W, g, V\rangle$ and $M' = \langle W', g', V'\rangle$ be models. Then:

1. $M \sqsubseteq_{\Box} M' \Leftrightarrow Th(M) \cap \Box\mathcal{L}_0 \subseteq_{\Box} Th(M') \Leftrightarrow \kappa(M) \subseteq_0 \kappa(M')$
2. $M \sqsubseteq_{\Diamond} M' \Leftrightarrow Th(M) \cap \Diamond\mathcal{L}_0 \subseteq_{\Diamond} Th(M') \Leftrightarrow \pi(M) \subseteq_0 \pi(M')$
3. $M \sqsubseteq M'\ \&\ V(g) \sqsubseteq V'(g') \Leftrightarrow M'$ bisimulates $M \Leftrightarrow Th(M) \subseteq Th(M')$
4. $M \sqsubseteq M' \Leftrightarrow \kappa(M) \subseteq \kappa(M')$ and $\pi(M) \subseteq \pi(M') \Leftrightarrow \kappa(M) \subseteq \kappa(M')$

Proof: We only prove the first item *in extenso*, the second is proven similarly, whereas the third is already observed in theorem 2.9. The facts in the last item can be deduced from the others, using the degree 1 normal forms from the proof of theorem 2.16 for the first equivalence.
So, suppose that $M \sqsubseteq_{\Box} M'$ and let μ be some propositional formula for which $M \models \Box\mu$, i.e. for all $w \in W : M_w \models \mu$. Choose any $v' \in W'$. Since $M \sqsubseteq_{\Box} M'$ there is a $v \in W$ such that $V(v) \sqsubseteq V'(v')$, and

so, since $\mu \in \mathcal{L}_0$ we use the lemma about propositional persistence (lemma 2.6) to conclude $M'_{v'} \models \mu$. Since v' was arbitrary, we have $M' \models \Box\mu$. The opposite direction is proven using contraposition: if $M \not\sqsubseteq_\Box M'$, then there is some $w' \in W'$ such that for all $w \in W$: $V(w) \not\sqsubseteq V'(w')$. So, for each $w \in W$ there is a literal $\alpha_w \in \mathcal{P} \cup \neg\mathcal{P}$ such that $M_w \models \alpha_w$ and $M'_{w'} \not\models \alpha_w$. Now if $\alpha = \bigvee_{w \in W} \alpha_w$, obviously $\alpha \in \mathcal{L}_0$ and $M_w \models \alpha$ for all $w \in W$, so $M \models \Box\alpha$, yet $M'_{w'} \not\models \alpha$, so $M' \not\models \Box\alpha$. Therefore $\kappa(M) \cap \mathcal{L}_0 \not\subseteq \kappa(M')$, i.e. $\kappa(M) \not\subseteq_0 \kappa(M')$. ■

It is not hard to see that we really need the restrictions to $\mathcal{L}_0$; in the case of $\sqsubseteq_\Box$ for instance, let $M' = \langle W', g', V' \rangle$ such that $V'(w')(p) = 0$ for all $w' \in W'$. Consider $M = \langle W, g, V \rangle$ with $W = W' \cup \{x\}$ for some $x \notin W'$, $V(x)(p) = 1$ and $V(w') = V'(w')$ for all $w' \in W'$. Although $M \sqsubseteq_\Box M'$, we have $M \models \Box\Diamond p$, but at the same time $M' \not\models \Box\Diamond p$.

Corollary 3.23 **(Lemma 3.17, continued)** Let S and S' be two stable sets such that $\Box^- S \subseteq_0 \Box^- S'$ and $\Diamond^- S \subseteq_0 \Diamond^- S'$. Then $S \subseteq S'$.

Proof: Let M and M' be such that $S = \kappa(M)$, $S' = \kappa(M')$. Applying the first two items of theorem 3.22, we obtain $M \sqsubseteq_\Box M'$ and $M \sqsubseteq_\Diamond M$, hence $M \sqsubseteq M'$ and thus, by the last item of the same theorem, $S \subseteq S'$. ■

Corollary 3.24 **(Filtration)** For every model $M = \langle W, g, V \rangle$ there is a model $M' = \langle W', g', V' \rangle$ such that $V(g) = V'(g')$ and for all $u, v \in W'$ for which $u \neq v$, $V(u) \neq V(v)$ and $Th(M) = Th(M')$. We call M' a *distinctive* model, and denote it also with $D(M)$, the distinctive model that M gives rise to.

Proof: If $g \notin W$, put $g' = g$ and $V'(g') = V(g)$. For worlds in W we define the equivalence relation $x \equiv y \Leftrightarrow V(x) = V(y)$, denoting the equivalence class of x with $[x]$. Then, we define $W' = \{[x] \mid x \in W\}$ and $V'([x]) = V(x)$. By definition of V', all worlds in W' have a different valuation; moreover one easily checks that we have $V(g) \sqsubseteq V'(g') \sqsubseteq V(g)$ and $M \sqsubseteq M' \sqsubseteq M$, whence, by theorem 3.22, $Th(M) = Th(M')$. ■

Since we assumed $\mathcal{P}$ to be finite, our filtration result differs from the classical filtration lemma by the fact that we do not need to make a filtration *through a given formula.* Moreover, by identifying worlds, we only compare their propositional content. This is explained by the special structure of balloon models; essentially the same property guarantees a normal form with a modal degree at most 1. The way we will exploit the above filtration result is a more traditional one, however. To formulate it, let us write $M \cong M'$ for the equivalence defined by $D(M) = D(M')$.

Corollary 3.25

- $\mathbf{L}$ is sound and complete with respect to $\{ D(M) \mid M \text{ is a balloon model}\}$.
- For all models M and N, and any $\preceq \in \{\sqsubseteq_\Diamond, \sqsubseteq_\Box, \sqsubseteq\}$, we have $D(M) \preceq N \Leftrightarrow M \preceq N \Leftrightarrow M \preceq D(N)$.
- Every consistent formula φ has only finitely many finite models that are not $\cong$-equivalent.

Now we can characterize our different notions of honesty in semantic terms.

Theorem 3.26 φ is naïvely honest iff $\Box\varphi$ has a $\sqsubseteq$-minimal model.

Proof: Using corollary 3.20 and theorem 3.22, the argument is straightforward:

$$\begin{array}{lll}
\varphi \text{ is naively honest} & \Leftrightarrow & \text{(definition)} \\
\exists S \in \mathcal{ST}(\varphi)\ \forall S' \in \mathcal{ST}(\varphi) : S \subseteq S' & \Leftrightarrow & \text{(cor. 3.20)} \\
\exists M : \varphi \in \kappa(M)\ \&\ \forall M'(\varphi \in \kappa(M') \Rightarrow \kappa(M) \subseteq \kappa(M')) & \Leftrightarrow & \text{(def. } \kappa\text{, thm 3.22)} \\
\exists M : M \models \Box\varphi\ \&\ \forall M'(M' \models \Box\varphi \Rightarrow M \sqsubseteq M') & \Leftrightarrow & \text{(def. 3.21)} \\
\exists M \text{ which is } \sqsubseteq\text{-minimal for } \Box\varphi & &
\end{array}$$

■

Example 3.27 Figure 2 gives $\sqsubseteq$-minimal models for $p \wedge q$, $\Diamond(p \wedge q)$ and $\Box(p \wedge q)$, respectively. As was announced in example 3.6, the latter two formulas are thus naïvely honest by virtue of the two models. Let us check that M' is $\sqsubseteq$-minimal for $\Diamond(p \wedge q)$: suppose N is an arbitrary model for $\Diamond(p \wedge q)$. Since M' has an empty world, we immediately obtain $M' \sqsubseteq_\Box N$; moreover, since $N \models \Diamond(p \wedge q)$, there must be some world u satisfying both p and q, and this world obviously extends the two balloon worlds of M', therefore $M' \sqsubseteq_\Diamond N$.

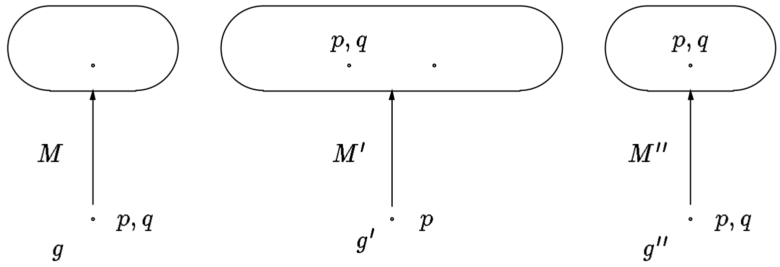

Figure 2: Three $\sqsubseteq$-minimal models

Theorem 3.28 φ is weakly honest iff $\Box\varphi$ has a $\sqsubseteq_\Box$-minimal model.

Proof: Again a direct argument is possible:

$$\begin{array}{lll}
\varphi \text{ is weakly honest} & \Leftrightarrow & \text{(def. weak honesty)} \\
\exists S \in \mathcal{ST}(\varphi)\ \forall S' \in \mathcal{ST}(\varphi) : S \subseteq_0 S' & \Leftrightarrow & \text{(cor. 3.20)} \\
\exists M : \varphi \in \kappa(M)\ \&\ \forall M'(\varphi \in \kappa(M') \Rightarrow \kappa(M) \subseteq_0 \kappa(M')) & \Leftrightarrow & \text{(def. } \kappa\text{, thm 3.22)} \\
\exists M : M \models \Box\varphi\ \&\ \forall M'(M' \models \Box\varphi \Rightarrow M \sqsubseteq_\Box M') & \Leftrightarrow & \text{(def. 3.21)} \\
\exists M \text{ which is } \sqsubseteq_\Box\text{-minimal for } \Box\varphi & &
\end{array}$$

■

Example 3.12 (continued)

The models M and M' of figure **??** are $\sqsubseteq_\Box$-minimal for $\Box(p \vee q)$ and $\Box p \vee \neg\Box p$, respectively. To see the latter, note first that M' is a model for $\neg\Box p$, and hence for $\Box p \vee \neg\Box p$. Moreover, $M' \sqsubseteq_\Box N$ for any N, due to the empty world in M'; thus M' is $\sqsubseteq_\Box$-minimal amongst the models for $\Box(\Box p \vee \neg\Box p)$.

We make the latter idea explicit, on the fly providing an alternative proof of observation 3.14:

Observation 3.29 Let us consider the following language $\mathcal{L}_\Diamond \supseteq \mathcal{L}_0$:

$$\pi \in \mathcal{L}_0, \varphi, \psi, \in \mathcal{L}_\Diamond \Rightarrow \pi \in \mathcal{L}_\Diamond, \Diamond\varphi \in \mathcal{L}_\Diamond \text{ and } (\varphi \wedge \psi) \in \mathcal{L}_\Diamond$$

Then, for all $\varphi \in \mathcal{L}_\Diamond$, we have: ($\varphi$ is satisfiable $\Leftrightarrow$ φ is weakly honest).

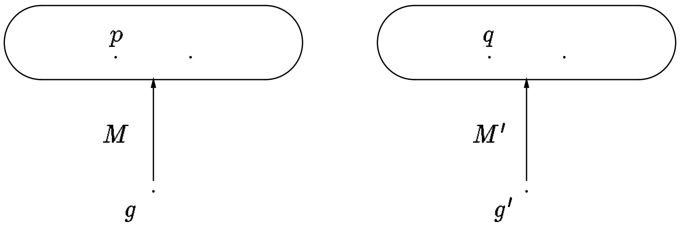

Figure 3: Two $\sqsubseteq_\Box$-minimal models

Proof: Note that formulas of $\mathcal{L}_\Diamond$ are satisfiable iff they are satisfiable in a model M with an empty world. But for such a model, we obviously have $M \sqsubseteq_\Box N$, for every model N. ∎

Connecting strong honesty with a semantic notion requires one more definition.

Definition 3.30 A model M is called *strongly minimal for* φ if M is $\sqsubseteq_\Diamond$-minimal in the set $\{ M' \mid M' \text{ is } \sqsubseteq_\Box \text{-minimal for } \varphi \}$.

Note that strongly minimal models for φ are by definition $\sqsubseteq_\Box$-minimal for φ. Also note, however, that a strongly minimal model (for φ) need not be $\sqsubseteq_\Diamond$-minimal.

Theorem 3.31 φ is strongly honest iff $\Box\varphi$ has a strongly minimal model.

Proof:
($\Rightarrow$) Assume S to be $\subseteq_\Diamond$-minimal amongst the $\subseteq_0$-minimal stable sets for φ, and let M be a model for which $S = \kappa(M)$ (corollary 3.20). We will show that M is strongly minimal for φ, i.e. $M \sqsubseteq_\Diamond M'$ for any M' that is $\sqsubseteq_\Box$-minimal for φ. Consider such an M'. Then, as in the proof of theorem 3.28, $\kappa(M')$ is $\subseteq_0$-minimal in $\mathcal{ST}(\varphi)$. So, by assumption, $\kappa(M) \subseteq_\Diamond \kappa(M')$. To draw the required conclusion $M \sqsubseteq_\Diamond M'$, it suffices, because of theorem 3.22 (2), to show that $\pi(M) \subseteq_0 \pi(M')$. This is straightforward: $\alpha \in \pi(M) \cap \mathcal{L}_0 \Rightarrow M \models \Diamond\alpha \Rightarrow M \models \Box\Diamond\alpha \Rightarrow \Diamond\alpha \in \kappa(M) \Rightarrow \Diamond\alpha \in \kappa(M') \Rightarrow M' \models \Box\Diamond\alpha \Rightarrow M' \models \Diamond\alpha \Rightarrow \alpha \in \pi(M')$.
($\Leftarrow$) Let M be strongly minimal for $\Box\varphi$, and $S = \kappa(M)$. If S' is some $\subseteq_0$-minimal stable set for φ, the proof of theorem 3.28 shows that the model M' for which $S' = \kappa(M')$ is $\sqsubseteq_\Box$-minimal for $\Box\varphi$. We have to show that $S \subseteq_\Diamond S'$. Since M was strongly minimal for $\Box\varphi$, we have that $M \sqsubseteq_\Box M'$ and $M \sqsubseteq_\Diamond M'$, i.e. $M \sqsubseteq M'$. Hence, by theorem 3.22 $\kappa(M) \subseteq \kappa(M')$, i.e. $S \subseteq S'$, and so $\Box\Diamond\psi \in Th(M)$, then $S \subseteq_\Diamond S'$. ∎

Example 3.16 (continued)

We argue that $\Diamond p \vee \Diamond q$ is not strongly honest: consider the two models M and M' of figure 4. Both models verify $\Box(\Diamond p \vee \Diamond q)$ and contain an empty balloon world, whence both are $\sqsubseteq_\Box$-minimal for $\Box(\Diamond p \vee \Diamond q)$. But then we also see that there can be no model N for $\Box(\Diamond p \vee \Diamond q)$ for which both $N \sqsubseteq_\Diamond M$ and $N \sqsubseteq_\Diamond M'$: such a model N has to contain at least a p- or a q-world, if it has a p-world then $N \not\sqsubseteq_\Diamond M'$, if it has a q-world, then $N \not\sqsubseteq_\Diamond M$.

3.3 Disjunction Properties

One might want to have an even more direct condition providing honesty, without interference of the notion of stability. Here, we will provide several syntactic, of perhaps rather deductive characterizations for honesty[6]. Inspecting the properties of saturated and stable sets, one good candidate for this is the

[6] In fact, it is highly questionable whether a real syntactic/morphological criterion for circumscriptive knowledge exists.

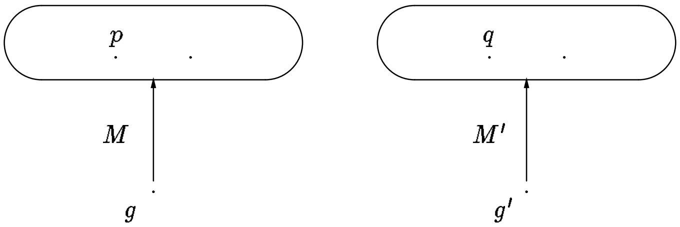

Figure 4: Two $\sqsubseteq_\Box$-minimal models

disjunction property, defined below. In fact, this property is already mentioned in [HC84], be it that there it is a property of logical systems, rather than of formulas. In partial logic this property should be slightly reformulated.

Definition 3.32 Disjunction Properties

Let $\varphi \in \mathcal{L}$. The following conditions define when φ is said to have the *disjunction property* (DP), the *propositional disjunction property* (PDP) or *propositional diamond disjunction property* (PDDP), respectively.

DP	$\forall \Sigma \subseteq \mathcal{L}$:	$\Box\varphi \vdash \Box\Sigma \Rightarrow \exists\sigma \in \Sigma : \Box\varphi \vdash \sigma$
PDP	$\forall \Pi \subseteq \mathcal{L}_0$:	$\Box\varphi \vdash \Box\Pi \Rightarrow \exists\pi \in \Pi : \Box\varphi \vdash \pi$
PDDP	$\forall \Pi \subseteq \mathcal{L}_0$:	$\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\Pi \Rightarrow \exists\pi \in \Pi : \Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\pi$

Observation 3.33

- All disjunction properties imply consistency. Take $\Pi, \Sigma = \emptyset$ for the arguments Π, Σ in the rules of the definition above.
- Note that $\Box\varphi \vdash \psi \Leftrightarrow \Box\varphi \vdash \Box\psi$ for all $\psi \in \mathcal{L}$. Furthermore, $\Box\varphi \vdash \Box\Delta, \Diamond\psi \Leftrightarrow \Box\varphi \vdash \Box\Delta, \Box\Diamond\psi$ for every formula ψ. These are simple consequences of corollary 2.15. This observation facilitates proving the next theorem on the relation between different notions of honesty and the various disjunction properties which were presented in the definition above.
- Note that PDDP implies PDP. To see this implication, suppose that φ is a formula which does not have PDP. This means there exists $\Pi \subseteq \mathcal{L}_0$ such that $\Box\varphi \vdash \Box\Pi$ (1) and $\Box\varphi \nvdash \pi$ for all $\pi \in \Pi$. This means $\Pi \subseteq (\overline{C_{\Box\varphi}})_0$, and by application of R-MON to (1) we derive $\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0$. This implies that φ does not have the PDDP, for there exists a Σ (viz. $\emptyset$) such that $\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\Sigma$ and $\Box\varphi \nvdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\sigma$ for all $\sigma \in \Sigma$.

Theorem 3.34 (Disjunction properties and honesty)

φ has the DP	$\Leftrightarrow$	φ is naïvely honest
φ has the PDP	$\Leftrightarrow$	φ is weakly honest
φ has the PDDP	$\Leftrightarrow$	φ is strongly honest

Proof: We start by proving the $\Leftarrow$-direction for the stated equivalences. These are in fact almost immediate consequences of the modal saturation property of stable sets (Cf. proposition 3.2 item 3) and the various characterizations of minimal stable sets.

- Let φ be naïvely honest. This means it has a $\subseteq$-minimal stable set S. Now suppose $\Box\varphi \vdash \Box\Sigma$, then $S \vdash \Box\Sigma$. By modal saturation we know $S \cap \Sigma \neq \emptyset$. According to theorem 3.7, $S = C_{\Box\varphi}$, so for some $\sigma \in \Sigma$, $\Box\varphi \vdash \sigma$. In other words, φ has the disjunction property.

- If φ is weakly honest, there is a similarly straightforward proof that φ has the PDP, since by theorem 3.13: $\exists S \in \mathcal{ST}(\varphi) : S_0 = (C_{\Box\varphi})_0$.

- Let φ be strongly honest. Suppose $\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\Pi$ for certain $\Pi \subseteq \mathcal{L}_0$. The second item in observation 3.33 tells us that

  $$\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Box\Diamond\Pi.$$

  Let S be $\subseteq_\Diamond$-minimal amongst the $\subseteq_0$-minima of $\mathcal{ST}(\varphi)$. Again modal saturation shows that there exists a $\varrho \in (\overline{C_{\Box\varphi}})_0 \cup \Diamond\Pi$ such that $\varrho \in S$. On account of theorem 3.19 we know that $S_0 = (C_{\Box\varphi})_0$, so ϱ must be some $\Diamond\pi$ in $\Diamond\Pi$. We also know from 3.19 that $S \cap \Diamond\mathcal{L}_0 = R^\Diamond_{\Box\varphi}$, so $\Diamond\pi \in R^\Diamond_{\Box\varphi}$. By definition of the diamond remainder of $\Box\varphi$, we conclude that $\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\pi$, the PDDP.

The $\Rightarrow$-direction of the proof is accounted for by the saturation lemma and the relation between stability and saturation as formulated in proposition 3.4. Then following three claims provide the desired results.

$a.$ φ has the DP $\quad\Rightarrow\quad \{\Box\varphi\} \trianglelefteq \Lambda_1 := \Box C_{\Box\varphi} \cup \overline{\Box\mathcal{L}},$

$b.$ φ has the PDP $\quad\Rightarrow\quad \{\Box\varphi\} \trianglelefteq \Lambda_2 := \Box(C_{\Box\varphi})_0 \cup \overline{\Box\mathcal{L}_0},$

$c.$ φ has PDP & PDDP $\quad\Rightarrow\quad \{\Box\varphi\} \trianglelefteq \Lambda_3 := \Box(C_{\Box\varphi})_0 \cup \Diamond R^\Diamond_{\Box\varphi} \cup \overline{(\Box\mathcal{L}_0 \cup \Diamond\mathcal{L}_0)}.$

The following arguments show that these implications are sufficient.

1. φ has DP $\Rightarrow$ (Saturation Lemma, claim a above)

   $\exists\Theta \in \mathcal{SAT} : \{\Box\varphi\} \subseteq \Theta \subseteq \Box C_{\Box\varphi} \cup \overline{\Box\mathcal{L}} \Rightarrow$ $(S = \Box^-\Theta,$ proposition 3.4)

   $\exists S \in \mathcal{ST}(\varphi) : S = C_{\Box\varphi} \quad\Rightarrow$ (corollary 3.8)

   φ is naïvely honest.

2. φ has PDP $\Rightarrow$ (Saturation Lemma, claim b above)

   $\exists\Theta \in \mathcal{SAT} : \{\Box\varphi\} \subseteq \Theta \subseteq \Box(C_{\Box\varphi})_0 \cup \overline{\Box\mathcal{L}_0} \Rightarrow$ $(S = \Box^-\Theta,$ proposition 3.4)

   $\exists S \in \mathcal{ST}(\varphi) : S_0 = (C_{\Box\varphi})_0 \quad\Rightarrow$ (theorem 3.13

   φ is weakly honest.

3. φ has PDDP $\Rightarrow$ (Saturation Lemma, claim c)

   $\exists\Theta \in \mathcal{SAT} : \{\Box\varphi\} \subseteq \Theta \subseteq \Box(C_{\Box\varphi})_0 \cup \Diamond R^\Diamond_{\Box\varphi} \cup \overline{\Box\mathcal{L}_0 \cup \Diamond\mathcal{L}_0} \Rightarrow$ $(S = \Box^-\Theta)$

   $\exists S \in \mathcal{ST}(\varphi) : S_0 = (C_{\Box\varphi})_0$ & $\Diamond(\Diamond^- S)_0 = R^\Diamond_{\Box\varphi} \quad\Rightarrow$ (theorem 3.19)

   φ is strongly honest.

What remains to be shown are the three claims $a - c$ above. Recall that this boils down to showing $\Sigma \cap \Lambda_i \neq \emptyset$, for each Σ for which $\Box\varphi \vdash \Sigma$ $i \leq 3$).

a Suppose φ has the DP and $\Box\varphi \vdash \Sigma$.

- If $\Sigma \cap \overline{\Box\mathcal{L}} \neq \emptyset$, we immediately obtain $\Sigma \cap \Lambda_1 \neq \emptyset$.

- If $\Sigma \cap \overline{\Box\mathcal{L}} = \emptyset$ then $\Sigma \subseteq \Box\mathcal{L}$, which means that $\Sigma = \Box\Sigma'$ for certain $\Sigma' \subseteq \mathcal{L}$. DP guarantees the existence of a $\sigma' \in \Sigma'$ such that $\Box\varphi \vdash \sigma'$. Since this means that $\Box\sigma' \in \Box C_{\Box\varphi}$, we may conclude $\Sigma \cap \Box C_{\Box\varphi} \neq \emptyset$, hence $\Sigma \cap \Lambda_1 \neq \emptyset$.

b Suppose φ has the PDP, and $\Box\varphi \vdash \Sigma$.

If $\Sigma \subseteq \Box\mathcal{L}_0$ then, according to PDP, there exists $\Box\sigma \in \Sigma$ such that $\Box\varphi \vdash \sigma$. This means $\Sigma \cap \Box(C_{\Box\varphi})_0 \neq \emptyset$. If $\Sigma \not\subseteq \Box\mathcal{L}_0$, then $\Sigma \cap \overline{\Box\mathcal{L}_0} \neq \emptyset$. Consequently, in all cases $\Sigma \cap \Lambda_2 \neq \emptyset$.

c Suppose φ has the PDDP, and $\Box\varphi \vdash \Sigma$.

- If $\Sigma \cap \overline{(\Box\mathcal{L}_0 \cup \Diamond\mathcal{L}_0)} \neq \emptyset$ then $\Sigma \cap \Lambda_3 \neq \emptyset$.
- Suppose $\Sigma \subseteq \Box\mathcal{L}_0 \cup \Diamond\mathcal{L}_0$.
  - If $\Sigma \cap \Box(C_{\Box\varphi})_0 \neq \emptyset$, then also $\Sigma \cap \Lambda_3 \neq \emptyset$.
  - Take $\Sigma \cap \Box\mathcal{L}_0 \subseteq \Box(\overline{C_{\Box\varphi}})_0$. In this remaining case all formulas of Σ are either of the form $\Diamond\pi$ with $\pi \in \mathcal{L}_0$ or $\Box\varrho$ with $\varrho \notin (C_{\Box\varphi})_0$. Application of the rule R-MON yields

    $$\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Sigma \cap \Diamond\mathcal{L}_0.$$

    According to PDDP this means that there exist $\sigma \in \Sigma \cap \Diamond\mathcal{L}_0$ such that $\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \sigma$ and therefore $\sigma \in \Diamond R^{\Diamond}_{\Box\varphi}$. We conclude, also in this last case, $\Sigma \cap \Lambda_3 \neq \emptyset$. ■

Since the disjunction properties are purely inferential and strictly related to the possibly honest formula under inspection, and neither involves extension to a stable set that is minimal in some sense, nor minimization in a class of models, they provide a very convenient tool for testing honesty. In particular, to prove that some formula is *dishonest*, disjunction properties may be a great help, as is illustrated below.

Example 3.12 (continued)

Using the PDP it easily follows that $\Box p \vee \Box q$ is not weakly honest: $\Box(\Box p \vee \Box q) \vdash \Box p \vee \Box q$, so $\Box(\Box p \vee \Box q) \vdash \Box\{p, q\}$, yet $\Box(\Box p \vee \Box q) \not\vdash p$ and $\Box(\Box p \vee \Box q) \not\vdash q$ (where non-derivability is shown by providing a countermodel, as usual). That $\Box p \vee \neg p$ is not weakly honest has a similar proof, again by taking $\Sigma = \{p, q\}$, thus contradicting the PDP.

Some of the earlier proofs can also be simplified. For example, observation 3.9 now has a very easy proof: $\Box(p \vee q) \vdash \Box p \vee \Box\Diamond q$, yet $\Box(p \vee q) \not\vdash p$ and $\Box(p \vee q) \not\vdash \Diamond q$, and thus DP shows that $p \vee q$ is not naïvely honest.

Though less comfortable, PDDP can be used for an alternative proof of example 3.16: $\varphi = \Diamond p \vee \Diamond q$ is not strongly honest since $\Box\varphi \vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond\{p, q\}$, $\Box\varphi \not\vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond p$, and $\Box\varphi \not\vdash \Box(\overline{C_{\Box\varphi}})_0, \Diamond q$. To show the latter, consider the model M from the proof at the end of section 3.2: M verifies $\Box\varphi$, but does not verify $\Diamond q$, nor any element of $\Box(\overline{C_{\Box\varphi}})_0$. To make this last point, suppose that $M \models \Box\alpha$ for some $\alpha \in \mathcal{L}_0$. Then in the empty balloon world ϵ: $M_\epsilon \models \alpha$, thus (by propositional persistence) $\models \alpha$, and therefore $\alpha \in (C_{\Box\varphi})_0$.

4 Conclusion

We have described a new epistemic logic with the remarkable feature that on the one hand knowledge implies truth, yet on the other hand truth does not imply epistemic possibility, thus avoiding at least one type of logical omniscience. The logic is shown to be sound and complete for so-called balloon models with partial interpretation.

This logic is then used as a vehicle to study circumscription of knowledge. We have introduced different notions of honesty, each of which can be equivalently described in a number of ways. This results in a hierarchy of honesty, since we can easily prove

φ is naïvely honest $\Rightarrow$ φ is strongly honest $\Rightarrow$ φ is weakly honest.

We summarize our hierarchy of honesty in the following table:

type	*in* $\mathcal{ST}(\varphi)$	*w.r.t.* $C_{\Box\varphi}$ *and* $R^{\Diamond}_{\Box\varphi}$	*for models of* $\Box\varphi$	*DP
weak	$\exists \subseteq_0$-minimum	$\exists S \in \mathcal{ST}(\varphi) : S =_0 C_{\Box\varphi}$	$\exists \sqsubseteq_{\Box}$-minimum	PDP
strong	$\exists \subseteq_{\Diamond}$-min. of $\subseteq_0$-minima	$\exists S \in \mathcal{ST}(\varphi) : S =_0 C_{\Box\varphi}$ & $\Diamond(\Diamond^{-} S)_0 = R^{\Diamond}_{\Box\varphi}$	$\exists$ strong minimum	PDDP
naïve	$\exists \subseteq$-minimum	$C_{\Box\varphi}$ is stable	$\exists \sqsubseteq$-minimum	DP

Table 1: Criteria for φ being honest

As we have illustrated on a number of examples, naïve honesty is too strong (i.e. it yields too many dishonest formulas), whereas weak honesty is indeed too weak — *strong honesty* is the preferable option. By means of honesty we can also define the semantics of the operator 'Agent only knows'.

Finally, we can evidently define a non-monotonic preferential entailment relation $\mid\!\sim$ by:

$$\varphi \mid\!\sim \psi \Leftrightarrow \varphi \text{ is strongly honest and for all strongly minimal } S \in \mathcal{ST}(\varphi) : \psi \in S.$$

This relation intuitively denotes that, if φ is only known, then ψ is also known. Due to the partial background logic, we find no entailment of irrelevant possibilities, e.g.:

$$p \mid\!\not\sim \Diamond q.$$

Notice that though many non-monotonic entailments that were valid for the classical system **S5** do not qualify for our partial system **L**, such entailment still differs from (partial) consequence and derivability: we have e.g. that

$$(p \vee q) \;\not\vdash \Diamond p \;\;\&\;\; (p \vee q) \mid\!\sim \Diamond p.$$

We can extend the latter example to show that '$\mid\!\sim$ ' is indeed a *non-monotonic relation*: we have for instance

$$(p \vee q) \wedge \neg p \mid\!\not\sim \Diamond p.$$

References

[Ben90] Benthem, J. van, 'Modal logic as a theory of information', in: J. Copeland (ed.) *Proceedings of the Prior Memorial Colloquium, Christchurch 1989*, to appear with Oxford University Press, Oxford UK

[Bla86] Blamey, S., 'Partial Logic', in: Gabbay & Günthner (eds) *Handbook of Philosophical Logic*, volume 3, Reidel, Dordrecht, 1986.

[HM85] Halpern, J. & Y. Moses, 'Towards a theory of knowledge and ignorance', in Kr. Apt (ed.) *Logics and Models of Concurrent Systems*, Springer–Verlag, Berlin, 1985

[HC84] Hughes, G. & M. Cresswell - *A Companion to Modal Logic*, Methuen, London, 1984

[Jas91a] Jaspars, J., 'Theoretical circumscription in partial modal logic', in: J. van Eijck (ed.), *Logics in AI*, Proceedings JELIA'90, pp. 301–316, LNCS 478, Springer–Verlag, Berlin, 1991

[Jas91b] Jaspars, J., 'A generalization of stability and its application to circumscription of positive introspective knowledge', *Proceedings of the Ninth Workshop on Computer Science Logic* (CSL'90), Springer–Verlag, Berlin, 1991

[Jas93] Jaspars, J. , 'Normal forms in partial modal logic', in: C. Rauszer (ed.), *Algebraic Methods in Logic and in Computer Science*, Banach Center Publications, volume 28, Institute of Mathematics, Polish Academy of Sciences, Warsaw, 1993

[JT93] Jaspars, J. & E. Thijsse, 'Fundamentals of Partial Modal Logic', in P. Doherty & D. Driankov (eds.), *Partial Semantics and Non-monotonic Reasoning for Knowledge Representation* (provisional title), based on workshop Linköping (Sweden) 1992, to appear

[Lan88] Langholm, T., *Partiality, Truth and Persistence*, CSLI Lecture Notes No. 15, Stanford CA, 1988.

[Moo85] Moore, R., 'Semantical considerations on non-monotonic logic', *Artificial Intelligence* 25, pp. 75–94, 1985

[ST92] Schwarz, G. & M. Truszczyński, 'Modal logic S4F and the minimal knowledge paradigm', in Y. Moses (ed.), *Proceedings of* TARK *5*, (Monterey CA), Morgan Kaufmann, Palo Alto CA, 1992

[Sta] Stalnaker, R., *A note on non-monotonic modal logic*, unpublished manuscript, Department of Philosophy, Cornell University

[Sto77] Stoy, J., *Denotational Semantics: The Scott-Strachey Approach to Programming Language Theory*, The M.I.T. Series in Computer Science, M.I.T. Press, Cambridge MA, 1977

[Thi90] Thijsse, E. - 'Partial propositional and modal logic: the overall theory', M. Stokhof & L. Torenvliet (eds.) *Proceedings of the* 7^{th} *Amsterdam Colloquium*, volume 2, pp. 555–579, ITLI, Amsterdam, 1990

[Thi92] Thijsse, E., *Partial logic and knowledge representation*, doctoral dissertation, Eburon Publishers, Delft, 1992