Evidence-Based Belief Revision for Non-Omniscient Agents

MSc Thesis *(Afstudeerscriptie)*

written by

Kristina Gogoladze
(born May 11th, 1986 in Tbilisi, Georgia)

under the supervision of **Dr Alexandru Baltag**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:
August 26, 2016

Members of the Thesis Committee:
Dr Alexandru Baltag
Prof Dr Johan van Benthem
Prof Dr Dick de Jongh
Prof Dr Benedikt Löwe (Chair)
Dr Benjamin Rin

Institute for Logic, Language and Computation

Abstract

It has long been recognized that inconsistencies may easily occur in people's beliefs in real life. Even if one is rational, one may hold inconsistent beliefs due to receiving conflicting information along with the fact that our limited capacity for information processing (or limited memory) may make it hard to spot the inconsistency. A rational agent would, of course, like to revise his beliefs when he becomes aware of an inconsistency. However, the usual discussion in the Belief Revision literature on solving the contradiction involving old evidence and new evidence assumes that the agent is always aware of this contradiction (because of his logical omniscience).

In this thesis, we propose a logic that allows an agent to hold inconsistent beliefs if he has not noticed that they are inconsistent. The logic allows the agent to reason consistently, even if there are inconsistencies in the agent's beliefs. Furthermore, if the agent becomes aware of the inconsistency, he is able to correct his beliefs so that they are no longer explicitly contradictory. We first give the semantics of the logic, and then present a sound and complete axiomatization for the proposed logic, forming a formal basis for holding and fixing inconsistent beliefs in a rational agent.

Acknowledgements

I was told that no one will ever read my thesis except for the committee members, so I have no idea why I am writing these acknowledgements at all :)

Before continuing, I would like to draw the reader's attention to the essential details. Firstly, there are many people I would like to thank for a huge variety of reasons—more people than I can keep in my memory entirely :) So I hope they will forgive me if I forget to mention someone. Secondly, it is difficult to determine the order in which I would want to list all the people to whom I am immensely grateful. So, do not be offended by the order—in this writing, it must be linear (or, could I have written a conjunction?).

I would like to thank my supervisor, Alexandru Baltag—it was nice to work with someone whose level of rationality is not bounded by 1 :)

I am grateful for the financial support from the Evert Willem Beth Foundation making it possible for me to study at ILLC.

I would like to thank David Gabelaia, Mamuka Jibladze, Dick de Jongh, Revaz Kurdiani for all their support before and during my studies in Amsterdam.

My gratitude also goes to the programme directors, the Examination Board (they must know why :)), my thesis committee members and, of course, Tanja Kassenaar for her care and being always so helpful.

Living in Amsterdam was a very exciting experience. ILLC is a place, where the largest amount of people I love is concentrated.

Here, to fit into the society standards, I probably have to thank my parents :)

Contents

Introduction

1.1 Motivation

An important outstanding problem in epistemic and doxastic logics is the problem of logical omniscience, unrealistic assumptions with regard to the reasoning power of the agents. It would be nice, of course, to have perfect reasoners, but even in powerful computers the resources for reasoning are limited.

Epistemic logicians usually consider the following features as different issues involving logical omniscience:

Knowledge of all logical validities: If ϕ is valid, then the agent knows that ϕ. So, the agent knows anything that can be deduced using the given system. In particular, this means that the agent knows all propositional tautologies.

Monotonicity: Suppose that $\phi \to \psi$ is valid (but not necessarily in the set of formulas known by the agent). Then, if the agent knows that ϕ, then the agent knows that ψ. This means that the agent knows all the logical consequences of his knowledge.

Closure under logical equivalence: If ϕ and ψ are logically equivalent, then the agent knows that ϕ if, and only if, the agent knows that ψ. This means that the agent cannot distinguish between two different formulas that are logically equivalent.

Closure under conjunction: If the agent knows that ϕ and the agent knows that ψ, then the agent knows that $\phi \wedge \psi$. That is, the set of formulas that the agent knows is closed under conjunction.

Closure under known implication: If the agent knows that ϕ implies ψ and the agent knows that ϕ, then the agent knows that ψ. This Distribution Axiom means that the set of formulas that the agent knows is deductively closed (i.e., if ϕ and $\phi \to \psi$ is in the set of formulas known by the agent, then so is ϕ).

Introspection: Philosophers generally deny that we are omniscient about our mental states. A self-knowledge or the corresponding epistemic introspection axioms can be seen as omniscience about one's mental states.

Each of the above principles is a feature of idealized perfect reasoners that may not exist in a rational agent in real life.

Assuming that the agent is rational, the reasons he may be non-omniscient are typically the following: limited computational power, time constraints and insufficient memory. These restrictions may also cause the agent to believe some contradictory facts (in this case, he simply may not have noticed the contradiction yet). We are not aware of any previous work that deals with inconsistent beliefs and that has a framework that would allow agents to fix inconsistent beliefs later. There are so-called paraconsistent logics that allow reasoning about inconsistencies, but the underlying philosophy of these logics is that believing a contradiction may be rational and that, in principle, there is no need to resolve logical contradictions. So, if we want to be able to explain why agents can hold inconsistent beliefs, we need to think of something different.

We would like to introduce and investigate a model of belief formation that will be closer to real-life reasoning than existing models. In particular, we want to propose a model that enables agents to reason about inconsistent beliefs when they are not aware of the inconsistency due to some limitations by introducing more natural definition of beliefs. Even rational agents may happen to believe irrational things either because they read/were told something or have false evidence from other sources. An agent will never believe an explicit contradiction $\bot$. If he notices such inconsistency, he will have to revise his current beliefs to keep them consistent.

1.2 The Problem of Inconsistent Beliefs

As is well known, general epistemic logics, namely, logics of knowledge and belief, suffer from the problem of logical omniscience. The so-called logical omniscience problem usually means that an agent's knowledge and beliefs are closed under logical implication. Such properties as closure under conjunction and introspection are also examples of logical omniscience. There have been many attempts to solve the problem of logical omniscience (see, for example, Fagin and Halpern [1987], Levesque [1984], Vardi [1986], Moreno [1998]). Some approaches are syntactic, others focus on the refinement of possible world semantics. The general goal is to offer a formalism under which an agent's knowledge and beliefs need not to be closed under logical operations. As a consequence of being a non-omniscient agent, the agent may sometimes have inconsistent beliefs.

However, even for a non-omniscient agent, it is usually assumed that agent's beliefs are always consistent and he cannot hold inconsistent information. Such models are, of course, extremely idealized, as we all know that even reasonable people do have inconsistent beliefs.

One can find several approaches for dealing with inconsistent beliefs problem in the literature. There are logics that are used to deal with paradoxes (for example, with preface-like paradoxes). In paraconsistent logics [Priest, 2002] it is assumed that it might be rational to believe some inconsistent things. But then, there is no need to fix such inconsistencies, or, at least, one cannot

distinguish between the case when one would want to correct the contradiction and the case when it is "rational" to have inconsistent beliefs. Another approach is to allow agents to notice the inconsistency, but then they are not allowed to reason with any propositions that are involved in the inconsistency [Thimm, 2012], they are just aware that there are some inconsistent pieces of information and use only the consistent part.

Existing logical systems, therefore, do not solve the problem of inconsistent belief revision with the possibility of making beliefs consistent in the future. In this work, we would like to make the first steps towards modeling the situation we just described, adapting these solutions to less idealized agents.

Since one of the main reasons why people hold inconsistent beliefs is limited computational resources, as a possible solution, we, firstly, restrict agents to the usage of only finite amount of sentences at every given period of time. These are going to be (the agent's) *explicit* beliefs—a finite set of syntactically given formulas. We will use some of the ideas from Levesque's logic [1984] of implicit and explicit beliefs. *Implicit* beliefs will not have all the restrictions we impose on the explicit beliefs, but agents can reason only with their explicit belief sets. The explicit belief sets are not required to be closed under any of the logical operations, the only restriction will be that they do not contain an *explicit inconsistency* which we denote by $\bot$.

This idea seems to work for *static* models—explicit beliefs are represented as a list of formulas, they do not contain $\bot$, so the agent has not noticed an inconsistency yet. But the situation changes when the agent becomes aware of the contradiction in his beliefs: he has to decide which believed sentences to give up to remove the explicit inconsistency from his explicit beliefs. Since all explicit beliefs are equally plausible in this model, the agent may have problems deciding which formulas to exclude from the belief set. Moreover, he may go in loops by giving up and then adding again the same formula. Harman states some principles that should be valid for any resource-bounded agent[1986], including *Recognized Inconsistency Principle*:"One has a reason to avoid believing things one recognizes to be inconsistent". Since the agent may fail to perform some inferences, he may fail to realize that some of his beliefs are inconsistent. But once he realizes the inconsistency, the agent must try to resolve it.

Our solution is to go one level up and start with *explicit evidence pieces* instead of explicit beliefs by borrowing some ideas from van Benthem and Pacuit's work [2011] on evidence-based beliefs. The explicit beliefs of an agent will then be computed using his explicit evidences. This "computation" is defined in such a way that it does not allow an explicit inconsistency in the agent's explicit belief set even when his evidence set does contain $\bot$.

1.3 Outline of the Thesis

The remainder of this thesis is structured as follows.

Chapter 2 The first section of this chapter provides some key notions and results from Modal Logic that are used in the thesis. It may be skipped by readers who are familiar with the basic concepts of the field.

The subsequent sections describe some of the related frameworks.

Chapter 3 In the first section of this chapter we discuss a possible solution to the problem of agents holding inconsistent beliefs. We present a semantic model of explicit beliefs and provide some intuition.

Most of the second section is devoted to proving completeness of the proposed axiomatic system with respect to the semantics given in the first section.

In the third section of the chapter we generalize our models to a multi-agent case and argue how the completeness result can be generalized as well.

In the last section we discuss the problems that arise in cases of belief revision, the possible solutions and what needs to be changed in our proposed model.

Chapter 4 In the first section of this chapter we present a formal semantic model of explicit evidence pieces. We discuss the differences between this new model and the previously suggested model, and provide some examples.

The second section is concerned with the complete axiomatization of the proposed (static) logic.

In the third section, we define several dynamic operators that describe changes of models when agents modify their explicit evidences.

In the fourth section of this chapter we sketch some ideas for obtaining recursion axioms.

In section 5 we present a detailed example to show how our explicit evidence models deal with the problem of belief dynamics.

In the last section of the chapter we discuss some possible language extensions with additional evidence modalities.

Chapter 5 Finally, we give a brief summary of the main results obtained and make suggestions for future work.

Preliminaries

2.1 Background

In this section, we briefly introduce some of the basic definitions and results from Modal Logic that will be needed throughout the text.

Definition 1 (Relational Frame and Model)**.** A *relational frame* is a tuple (W, R) where W is a nonempty set (elements of W are called *states*), $R \subseteq W \times W$ is a relation on W. A *relational model* (also called a Kripke model) is a triple $\mathfrak{M} = (W, R, V)$ where (W, R) is a relational frame and $V : \mathsf{At} \to \mathscr{P}(W)$ is a *valuation function* assigning sets of states to atomic propositions. $\triangleleft$

Definition 2 (Truth of Modal Formulas)**.** Suppose that $\mathfrak{M} = (W, R, V)$ is a relational model. Truth of a modal formula $\phi \in \mathcal{L}(\mathsf{At})$ at a state w in $\mathfrak{M}$, denoted $\mathfrak{M}, w \models \phi$, is defined inductively as follows:

$\mathfrak{M}, w \models p$ iff $w \in V(p)$ (where $p \in \mathsf{At}$)

$\mathfrak{M}, w \models \top$ and $\mathfrak{M}, w \not\models \bot$

$\mathfrak{M}, w \models \neg\phi$ iff $\mathfrak{M}, w \not\models \phi$

$\mathfrak{M}, w \models \phi \wedge \psi$ iff $\mathfrak{M}, w \models \phi$ and $\mathfrak{M}, w \models \psi$

$\mathfrak{M}, w \models \Box\phi$ iff for all $v \in W$, if wRv then $\mathfrak{M}, v \models \phi$

$\mathfrak{M}, w \models \Diamond\phi$ iff there is $v \in W$, such that wRv and $\mathfrak{M}, v \models \phi$

A set of formulas $\Gamma \subseteq \mathcal{L}$ is *satisfiable* if there is some model $\mathfrak{M} = (W, R, V)$ and world $w \in W$ such that $\mathfrak{M}, w \models \phi$ for all $\phi \in \Gamma$. A formula $\phi \in \mathcal{L}$ is satisfiable when $\{\phi\}$ is satisfiable. $\triangleleft$

Definition 3 (Validity)**.** A modal formula $\phi \in \mathcal{L}$ is *valid on a relational structure* $\mathfrak{M} = (W, R, V)$, denoted $\mathfrak{M} \models \phi$, provided $\mathfrak{M}, w \models \phi$ for each $w \in W$. Suppose that $\mathfrak{F} = (W, R)$ is a relational frame. A model formula $\phi \in \mathcal{L}$ is valid on $\mathfrak{F}$, denoted $\mathfrak{F} \models \phi$, provided $\mathfrak{M} \models \phi$ for all models based on $\mathfrak{F}$. Suppose that $\mathbf{F}$ is a class of relational frames. A modal formula ϕ is *valid on* $\mathbf{F}$, denoted $\models_{\mathbf{F}} \phi$, provided $\mathfrak{F} \models \phi$ for all $\mathfrak{F} \in \mathbf{F}$. If $\mathbf{F}$ is the class of all relational frames, then one writes $\models \phi$ instead of $\models_{\mathbf{F}} \phi$. $\triangleleft$

Definition 4 (Semantic Consequence)**.** Suppose that Γ is a set of modal formulas and $\mathbf{F}$ is a class of relational frames. We say ϕ is a *semantic consequence* of Γ with respect to $\mathbf{F}$, denoted $\Gamma \models_{\mathbf{F}} \phi$, provided for all models $\mathfrak{M} = (W, R, V)$, based on a frame from $\mathbf{F}$ and all states $w \in W$, if $\mathfrak{M}, w \models \Gamma$, then $\mathfrak{M}, w \models \phi$. $\triangleleft$

Definition 5 (Soundness, Weak/Strong Completeness)**.** Suppose that $\mathbf{F}$ is a class of relational frames. A logic $\mathbf{L}$ is *sound* with respect to $\mathbf{F}$ provided, for all sets of formulas Γ, if $\Gamma \vdash_{\mathbf{L}} \phi$, then $\Gamma \models_{\mathbf{F}} \phi$. A logic $\mathbf{L}$ is *strongly complete* with respect to $\mathbf{F}$ provided for all sets of formulas Γ, if $\Gamma \models_{\mathbf{F}} \phi$, then $\Gamma \vdash_{\mathbf{L}} \phi$. Finally, a logic $\mathbf{L}$ is *weakly complete* with respect to $\mathbf{F}$ provided if $\models_{\mathbf{F}} \phi$, then $\vdash_{\mathbf{L}} \phi$. $\triangleleft$

Definition 6 (Normal Modal Logics)**.** A *normal modal logic* Λ is a set of formulas that contains all tautologies, $\Box(p \to q) \to (\Box p \to \Box q)$, and $\Diamond p \leftrightarrow \neg\Box\neg p$, and that is closed under *Modus Ponens*, *uniform substitution* and *generalization*. We call the smallest normal modal logic K. $\triangleleft$

Completeness theorems are essentially model existence theorems. Given a normal logic Λ, we prove its strong completeness with respect to some class of structures by showing that every Λ-consistent set of formulas can be satisfied in some suitable model. The idea is to use *canonical models* to build suitable satisfying models.

Definition 7 (Λ-MCSs)**.** A set of formulas Γ is *maximal Λ-consistent* if Γ is Λ-consistent, and any set of formulas properly containing Γ is Λ-inconsistent. If Γ is a maximal Λ-consistent set of formulas then we say it is a Λ-MCS. $\triangleleft$

Proposition 8 (Properties of MCSs)**.** *If Λ is a logic and Γ is a Λ-MCS then:*

(i) *Γ is closed under Modus Ponens;*

(ii) *$\Lambda \subseteq \Gamma$;*

(iii) *for all formulas ϕ: $\phi \in \Gamma$ or $\neg\phi \in \Gamma$;*

(iv) *for all formulas ϕ, ψ: $\phi \in \Gamma$ or $\psi \in \Gamma$.*

Lemma 9 (Lindenbaum's Lemma)**.** *If Σ is a Λ-consistent set of formulas then there is a Λ-MCS Σ^+ such that $\Sigma \subseteq \Sigma^+$.*

We are now ready to build models out of MCSs, and in particular, to build the very special models known as canonical models. With the help of these structures we will be able to prove the Canonical Model Theorem, a universal completeness result for normal logics.

Definition 10 (Canonical Model)**.** The *canonical model* $\mathfrak{M}^\Lambda$ for a normal modal logic Λ is the triple $(W^\Lambda, R^\Lambda, V^\Lambda)$ where:

(i) W^Λ is the set of all Λ-MCSs;

(ii) R^Λ is the binary relation on W^Λ defined by $R^\Lambda wu$ if for all formulas ψ, $\psi \in u$ implies $\Diamond\psi \in w$. R^Λ is called the *canonical relation*.

(iii) V^Λ is the valuation defined by $V^\Lambda(p) = \{w \in w^\Lambda : p \in w\}$. V^Λ is called the *canonical valuation*.

The pair $\mathfrak{F}^\Lambda = (W^\Lambda, R^\Lambda)$ is called the *canonical frame* for Λ. ◁

Lemma 11. *For any normal logic Λ, $R^\Lambda wv$ if, and only if, for all formulas ψ, $\Box\psi \in w$ implies $\psi \in v$.*

Lemma 12 (Existence Lemma)**.** *For any normal modal logic Λ and any state $w \in W^\Lambda$, if $\Diamond\phi \in w$ then there is a state $v \in W^\Lambda$ such that $R^\Lambda wv$ and $\phi \in v$.*

Lemma 13 (Truth Lemma)**.** *For any normal logic Λ and any formula ϕ, $\mathfrak{M}^\Lambda, w \models \phi$ if, and only if, $\phi \in w$.*

Theorem 14 (Canonical Model Theorem)**.** *Any normal modal logic is strongly complete with respect to its canonical model.*

Definition 15 (Canonicity)**.** A formula ϕ is *canonical* if, for any normal logic Λ, $\phi \in \Lambda$ implies that ϕ is valid on the canonical frame for Λ. A normal logic Λ is *canonical* if its canonical frame is a frame for Λ. ◁

The proofs of these classical results can be easily found in textbooks on modal logic, for example, see Blackburn at al.[2001]

Many important frame completeness results can be proved straightforwardly using canonical models. The key idea in such proofs is to show that the relevant canonical frame has the required properties. Such proofs boil down to the following task: showing that the axioms of the logic are canonical for the properties we want.

2.2 Awareness Models

How can someone say that he believes or does not believe about ϕ if ϕ is a concept he is completely unaware of? Awareness[Fagin and Halpern, 1987] is based on the intuition that an agent should be aware of a concept before he can have beliefs about it. The term "awareness" may have a number of interpretations. One of them is that an agent is aware of a formula if he can compute whether or not it is true in a given situation within a certain time or space bound. The formulas that an agent is aware of are represented syntactically.

Let us say that an agent is *aware* of a primitive proposition p, which we abbreviate Ap, if either the truth or falsity of p is known. Intuitively, this means that p is somehow relevant to the situation and that the agent is "aware" of p in that situation. Not every valid formula will be believed, though, it is the case that a valid formula is believed provided an agent is aware of all the primitive propositions that appear in it. We associate with every world w the set $A(w)$ of formulas that the agent is aware of.

Definition 16 (Language of Awareness Logic)**.** Let At be a set of atomic propositions. Formulas ϕ of the *epistemic awareness language* $\mathcal{L}'$ are given by

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid K\phi \mid A\phi$$

with $p \in \mathsf{At}$. Other Boolean connectives $\vee$, $\rightarrow$, $\leftrightarrow$, as well as the existential modal operator is defined as usual. $\triangleleft$

We will read formulas $A\phi$ as "the agent is aware of ϕ", and formulas $K\phi$ as "the agent knows ϕ implicitly". The language is interpreted in epistemic models assigning to each agent in each world a set of formulas, representing the information he is aware of.

Definition 17 (Semantic Awareness Model)**.** An *epistemic awareness model* is a tuple $\mathfrak{M} = (W, R, A, V)$ where

(W, R, V) is a standard epistemic model: a set or worlds W, an accessibility relation $R \subseteq W \times W$, and a valuation $V : \mathsf{At} \rightarrow \mathscr{P}(W)$.

$E : W \rightarrow \mathscr{P}(\mathcal{L}')$ is the "awareness" function giving the formulas that the agent "is aware of". $A(w)$ is the awareness set at w.

$\triangleleft$

The semantic interpretation of formulas in $\mathcal{L}'$ is entirely as expected:

Definition 18 (Truth in Awareness Model)**.** Let $\mathfrak{M} = (W, R, A, V)$ be an epistemic awareness model. Truth of a formula $\phi \in \mathcal{L}'$ is defined inductively as follows:

$\mathfrak{M}, w \models p \quad$ iff $\quad w \in V(p)$

$\mathfrak{M}, w \models \neg\phi \quad$ iff $\quad \mathfrak{M}, w \not\models \phi$

$\mathfrak{M}, w \models \phi \wedge \psi \quad$ iff $\quad \mathfrak{M}, w \models \phi \;$ and $\; \mathfrak{M}, w \models \psi$

$\mathfrak{M}, w \models A\phi \quad$ iff $\quad \phi \in A(w)$

$\mathfrak{M}, w \models K\phi \quad$ iff $\quad Rww' \;$ implies $\; \mathfrak{M}, w' \models \phi \;$ for all $\; w' \in W$

The truth set of ϕ is the set of worlds $[\![\phi]\!]_{\mathfrak{M}} = \{w : \mathfrak{M}, w \models \phi\}$. Standard logical notions of *satisfiability* and *validity* are defined as usual. $\triangleleft$

In these models we can impose standard epistemic assumptions about the accessibility relation, such as reflexivity, transitivity, and symmetry. One could also require that the awareness operator be introspective, or weakly introspective: $A\phi \rightarrow KA\phi$.

We have semantics for only implicit knowledge so far. Explicit knowledge can be introduced either as a primitive notion—when it is defined as a map assigning formulas to worlds, but not identified with any combination of operators K and A. Or, by combining implicit knowledge and the information the agent is aware of, which can be done in different ways.

These epistemic awareness models suggest natural *dynamic actions*. Though the agent is not logically omniscient anymore, he can get new information by various acts. For example, an agent may become aware of new facts.

Definition 19 (Consider Operation)**.** Let $\mathfrak{M} = (W, R, A, V)$ be an epistemic awareness model and χ any formula in $\mathcal{L}'$. The model $\mathfrak{M}^{+\chi} = (W, R, A', V)$ is $\mathfrak{M}$ with its awareness sets extended with χ, that is,

$$A'(w) = A(w) \cup \chi \quad \text{for every} \quad w \in W$$

◁

"Considering" extends the information an agent is aware of, but we can also define a neglecting operation with the opposite effect: reducing awareness sets. This fits with the operational idea that agents can expand and shrink the set of issues having their current attention.

Of course, one can also define explicit and implicit beliefs in a similar manner, and introduce multi-agent models, but we will not discuss it here.

2.3 Evidence Models and Evidence-Based Beliefs

A rational belief must be grounded in the evidence available to an agent. Van Benthem and Pacuit [2011] introduced evidence models that are, in their turn, based on neighborhood models [Pacuit, 2007].

Let W be a set of possible worlds one of which represents the "actual" situation. An agent gathers evidence about this situation from various sources. In line with our evidence interpretation, some constraints will be imposed on the evidences:

. No evidence set is empty (this means that evidences are not contradictory)

. The whole universe W is an evidence set (agents know their "space")

Additionally, the so-called property of "monotonicity" should hold:

If agent i has evidence X and $X \subseteq Y$, then i has evidence Y.

The language for reasoning about evidence and beliefs is defined as follows.

Definition 20 (Language of Evidence and Belief)**.** Let At be a set of atomic propositions. Formulas ϕ of the *evidence language* $\mathcal{L}''$ are given by

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid B\phi \mid \Box\phi \mid A\phi$$

with $p \in \mathsf{At}$. Other Boolean connectives $\vee$, $\rightarrow$, $\leftrightarrow$, as well as the existential modal operators are defined as usual. ◁

We will read formulas $\Box\phi$ as "the agent has evidence that implies ϕ", and formulas $B\phi$ as "the agent believes ϕ". We also include a universal modality $A\phi$ that means "ϕ is true in all states".

Having evidence for ϕ need not imply belief. In order to believe a proposition ϕ, an agent must consider all his evidence for or against ϕ.

Definition 21 (Evidence Model)**.** An *evidence model* is a tuple $\mathfrak{M} = (W, E, V)$ where

W is a non-empty set of worlds.

$E \subseteq W \times \mathscr{P}(W)$ is an evidence (neighborhood) relation.

$V : \mathsf{At} \to \mathscr{P}(W)$ is a valuation function.

A *pointed evidence model* is a pair $\mathfrak{M}, w$ with "actual world" w. When E is a constant function, we get a *uniform evidence model* $\mathfrak{M} = (W, \mathcal{E}, V), w$ with $\mathcal{E}$ the fixed family of subsets of W related to each state by E. ◁

We write $E(w)$ for the set $\{X : wEX\}$.

Even though evidence pieces are non-empty, their combination through the obvious operation of taking intersections need not yield consistent evidence: combining disjoint sets will lead to trouble. But importantly, even though an agent may not be able to consistently combine all of his evidence, there will be maximal collections that he can safely put together:

Definition 22 (Maximal Consistent Evidence)**.** A family $\mathcal{X}$ of subsets of W has the *finite intersection property* (f.i.p.) if $\bigcap \mathcal{X} \neq \varnothing$. We say that $\mathcal{X}$ has the *maximal f.i.p.* if $\mathcal{X}$ has the f.i.p., but no proper extension of $\mathcal{X}$ does. ◁

We now interpret the language $\mathcal{L}''$ on neighborhood models.

Definition 23 (Truth in Evidence Model)**.** Let $\mathfrak{M} = (W, E, V)$ be an evidence model. Truth of a formula $\phi \in \mathcal{L}''$ is defined inductively as follows:

$\mathfrak{M}, w \models p$ iff $w \in V(p)$

$\mathfrak{M}, w \models \neg\phi$ iff $\mathfrak{M}, w \not\models \phi$

$\mathfrak{M}, w \models \phi \wedge \psi$ iff $\mathfrak{M}, w \models \phi$ and $\mathfrak{M}, w \models \psi$

$\mathfrak{M}, w \models \Box\phi$ iff there is an X with wEX and for all $v \in X, \mathfrak{M}, v \models \phi$

$\mathfrak{M}, w \models B\phi$ iff for each maximal f.i.p. family$\mathcal{X} \subseteq E(w)$ and for all $v \in \bigcap \mathcal{X}, \mathfrak{M}, v \models \phi$

$\mathfrak{M}, w \models A\phi$ iff for all $v \in W, \mathfrak{M}, v \models \phi$

The truth set of ϕ is the set of worlds $[\![\phi]\!]_{\mathfrak{M}} = \{w : \mathfrak{M}, w \models \phi\}$. Standard logical notions of *satisfiability* and *validity* are defined as usual. ◁

Again, various extensions to the above modal language make sense, including conditional belief and conditional evidence.

Evidence is continually affected by new incoming information, and also by processes of internal re-evaluation. So, as with the awareness case, one can introduce different dynamic operators as well. These neighborhood models of evidence and belief suggest a new scope for these methods in dealing with more finely-structured evidence dynamics. We show one such operation here.

Definition 24 (Evidence Upgrade)**.** Let $\mathfrak{M} = (W, E, V)$ be an evidence model and ϕ a formula in $\mathcal{L}''$. The model $\mathfrak{M}^{\Uparrow\phi} = (W^{\Uparrow\phi}, E^{\Uparrow\phi}, V^{\Uparrow\phi})$ has $W^{\Uparrow\phi} = W$, $V^{\Uparrow\phi} = V$, and for all $w \in W$,

$$E^{\Uparrow\phi}(w) = \{X \cup [\![\phi]\!]_{\mathfrak{M}} : X \in E(w)\} \cup [\![\phi]\!]_{\mathfrak{M}}$$

$\triangleleft$

This is stronger than simply adding $[\![\phi]\!]_{\mathfrak{M}}$ as evidence, since one modifies each admissible evidence set. But it is weaker than publicly announcing ϕ, as the agent retains the ability to consistently condition on $\neg\phi$.

This setting can also model further phenomena. For instance, there is a natural notion of "reliable" evidence, based only on sets containing the actual world. We can imagine many other interesting notions that are not that important for this review.

2.4 Other Related Work

There are other frameworks that are related in one way or another to explicit beliefs or evidence pieces. Here is a short overview of some of them.

Learning Rules This is a framework [Velázquez-Quesada, 2009; Velázquez-Quesada, 2014] for representing implicit and explicit beliefs combined with the idea that the agent is able to perform inferences only if he explicitly knows the corresponding rules. The language has two components: formulas and rules. Where rules are pairs consisting of a set of formulas, the rule's premises, and a single formula, the rule's conclusion. Of course, for a non-omniscient agent, these rules cannot be given as schemes, the rules are defined as particular instantiations. Semantics is given on a pointed plausibility Kripke models.

Paraconsistent Logics Edwin Mares [2013] introduces a paraconsistent approach to belief change that is based on relevance logic [Anderson and Belnap, 1976]. In this approach, an agent has both a belief set consisting of the sentences he believes in and a reject set consisting of the sentences he rejects. Reject sets are duals of theories, which means that they are closed under disjunction and that if a reject set contains a certain sentence then it also contains all sentences that entail it. Both the belief set and the reject set can be subjected to the standard belief revision operations.

In this general framework Mares develops two systems that he calls the strong and the weak systems of paraconsistent belief change. In the strong system, an agent who accepts an inconsistent set of beliefs is committed to believing all conjunctions of elements of that set. In the weak system the agent can accept inconsistent beliefs without committing himself to believe in their conjunction.

Justification Logic Justification logics first introduced by Sergei Artemov [2008] are epistemic logics which allow knowledge and belief modalities to be "explained" via *justification terms*. Justification logic is based on classical propositional logic augmented by justification assertions $t : F$ that read t is a justification for F. Later, Baltag at al. [2014] presented a logic for reasoning about a notion of completely trustworthy evidence and its relations to justifiable (implicit) belief and knowledge, as well as to their explicit justifications. This logic makes use of several evidence-related notions such as availability, admissibility, and "goodness" of a piece of evidence.

Topological Models There are a number of *topological models* that deal with explicit beliefs or evidences of agents. One of the latest frameworks [Baltag *et al.*, 2016] introduces a new topological semantics for evidence, evidence-based belief, knowledge and learning. They represent notions like *knowledge*, *basic pieces of evidence* and *evidence-based beliefs* in *topological spaces* where a family of (combined) evidence sets forms a topological basis.

Resource-Bounded Rationality Renata Wassermann's PhD thesis [2000] seems to be relevant to some notions of ours. Namely, the *Recognized Inconsistency Principle* which states that one has to avoid the inconsistency once it has been noticed is related to our notion of *quasi-consistency* and to the way beliefs are formed and revised in our proposed models, and *Recognized Implication Principle* which argues that one has a reason to believe ϕ if one recognizes that ϕ is implied by one's view, this principle describes closure of beliefs under Modus Ponens in case the agents infers the conclusion.

Answer Set Programming The proposal in this thesis bears some resemblance to "cautious reasoning" in Answer Set Programming [Lifschitz, 2008]. The "stable models" from Answer Set Programming are somewhat similar to our q-max sets of evidence formulas (see Section 4.1). In cautious reasoning, a ground atom a is accepted if it belongs to all stable models, which is similar to the way we form beliefs in our explicit evidence models by taking the intersection of all q-max sets of evidence. However, there are many differences. Stable models are formed only of ground atoms (since they are meant to be consistent "models"), while our q-max sets may consist of any evidence formulas and are not necessarily consistent (but only "quasi-consistent"). Also, stable models need to satisfy the required rules, which is different (though vaguely similar) to our

requirement that q-max sets are closed under Modus Ponens within the set $\mathcal{E}_s$ of all evidence. Overall, the intended motivation, technical details and behavior of our models are very different from the ones of Answer Set Programming, despite the few surface similarities.

Explicit Belief and Explicit Knowledge

To prepare the reader for the more complex model and to show the possible difficulties which may arise during the development of the model's wished properties, we will discuss more basic approach in this chapter. This section will also help us to make the proof of our main theorem simpler later.

3.1 Explicit Belief Models

We first introduce the basic logic that deals with having inconsistent beliefs for an agent. Let us start with the modal language.

Definition 25 (Language $\mathcal{L}$)**.** Let At be a set of atomic propositions. Formulas ϕ of language $\mathcal{L}$ are given by

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid B^e\phi \mid B^i\phi \mid K^e\phi \mid K^i\phi$$

with $p \in \mathsf{At}$. Other Boolean connectives $\vee$, $\rightarrow$, $\leftrightarrow$, as well as the existential modal operators for knowledge and belief are defined as usual. $\triangleleft$

We will read formulas $B^e\phi$ as "the agent believes ϕ explicitly", and formulas $B^i\phi$ as "the agent believes ϕ implicitly". Similarly, we will read formulas $K^e\phi$ as "the agent knows ϕ explicitly", and formulas $K^i\phi$ as "the agent knows ϕ implicitly".

The idea behind separating explicit and implicit knowledge and belief is agents' limited capacity for information processing (or limited memory): one can entertain only finite (explicit) amount of information at once—which is going to be explicit knowledge or belief.

The language is interpreted in suitable models:

Definition 26 (Semantic Model of Explicit Beliefs)**.** An *explicit belief model* (EB-model) is a tuple $\mathfrak{M} = (W, W_0, \mathcal{B}, \mathcal{K}, V)$ where

W is a non-empty set of possible worlds.

$W_0 \subseteq W$ represents (the set of worlds that are consistent with) the agent's *background beliefs* or "biases".

$\mathcal{B} \subseteq \mathscr{P}(\mathcal{L})$ represents the set of formulas that are "explicitly believed" by the agent.

$\mathcal{K}$ represents the set of sentences that are "explicitly known", subject to the requirement that $\mathcal{K} \subseteq \mathcal{B}$ (i.e. everything that is explicitly known is also explicitly believed).

$V : \mathsf{At} \to \mathscr{P}(W)$ is a valuation function.

Two conditions are imposed on these models to ensure the properties that should hold:

$$\bot \notin \mathcal{B}$$
$$\top \in \mathcal{K}$$

◁

The first condition says that the agent does not explicitly believe a contradiction. Even though we allow agents to have inconsistent explicit beliefs, we do not want them to believe $\bot$ because we assume the agents are rational. The second requirement means that the agent has some knowledge to start with.

The background beliefs of an agent represent his implicit assumptions, the "biases" that he has without necessarily being aware of having them.

We say that the agent *explicitly believes* a formula at some world if, and only if, that formula belongs to the set of his explicit beliefs $\mathcal{B}$. We will assume that the agent does not explicitly believe a contradiction $\bot \notin \mathcal{B}$. This notion will be used throughout the text:

Definition 27 (Quasi-consistency)**.** Let U be a set of formulas. We say that U is *quasi-consistent* if $\bot \notin U$. ◁

As for implicit belief, at first sight one may want the implicit beliefs to be exactly what is implied by the explicit beliefs of the agent. However, this cannot be the case because there may be sentences that are true in the set of possible worlds, but are not implied by what is believed. For example, there are basic, prior facts, or instincts such that the agent never thinks of them explicitly, but he subconsciously knows them and uses them. Actually, those things may be even the opposite to what is explicitly believed. Another example is introspection—it does not simply follow from what is explicitly believed. That is why we have the set W_0 in the model, this is the set of worlds that correspond to all the sentences that are basic, ground beliefs.

So there are reasons not to restrict implicit beliefs to only closure of the explicit ones. With these ground beliefs in mind, we say that the agent *implicitly believes* a formula if, and only if, the formula is logically entailed by his explicit beliefs together with his background (implicit) beliefs or "biases".

Of course, we also have the interpretation of knowledge which is defined as one might expect. We say that the agent *explicitly knows* a formula if, and only if, that formula belongs to the set of his explicit knowledge $\mathcal{K}$[1]. By assumption, all the formulas in $\mathcal{K}$ have to be true in every world. And, lastly, the agent *implicitly knows* a formula if, and only if, the formula is true in all possible worlds W.

Definition 28 (Truth)**.** Let $\mathfrak{M} = (W, W_0, \mathcal{B}, \mathcal{K}, V)$ be an explicit belief model. Truth of a formula $\phi \in \mathcal{L}$ is defined inductively as follows:

$$\mathfrak{M}, w \models p \quad \text{iff} \quad w \in V(p)$$

$$\mathfrak{M}, w \models \neg\phi \quad \text{iff} \quad \mathfrak{M}, w \not\models \phi$$

$$\mathfrak{M}, w \models \phi \wedge \psi \quad \text{iff} \quad \mathfrak{M}, w \models \phi \ \text{ and } \ \mathfrak{M}, w \models \psi$$

$$\mathfrak{M}, w \models B^e\phi \quad \text{iff} \quad \phi \in \mathcal{B}$$

$$\mathfrak{M}, w \models B^i\phi \quad \text{iff} \quad w' \models \phi \ \text{ for all } \ w' \in \{v : \mathfrak{M}, v \models \theta \text{ for all } \theta \in \mathcal{B}\} \cap W_0$$

$$\mathfrak{M}, w \models K^e\phi \quad \text{iff} \quad \phi \in \mathcal{K}$$

$$\mathfrak{M}, w \models K^i\phi \quad \text{iff} \quad w' \models \phi \ \text{ for all } \ w' \in W$$

The truth set of ϕ is the set of worlds $[\![\phi]\!]_{\mathfrak{M}} = \{w : \mathfrak{M}, w \models \phi\}$. Standard logical notions of *satisfiability* and *validity* are defined as usual. ◁

We will write simply $w \models \phi$ and $[\![\phi]\!]$ when the model is clear from the context.

Note that both explicit and implicit beliefs (as well as knowledge) are defined *globally*: if an agent believes something in some world, he believes it in every world that he considers to be possible. All the possible worlds are implicitly known by the agent, so the implicit knowledge modality is a universal modality.

With this semantics, implicit beliefs are those beliefs that can be potentially derived from the explicit ones, so if an agent happens to have an inconsistent set of explicit beliefs, then his set of implicit beliefs is inconsistent as well.

It is easy to visualize the sentences of explicit beliefs and what follows from them.

[1]There are other possible definitions of explicit knowledge, for example, to define it via explicit beliefs and implicit knowledge: $K^e\phi := B^e\phi \wedge K^i\phi$ [Fagin and Halpern, 1987] or $K^e\phi := B^eK^i\phi \wedge K^i\phi$ [van Benthem and Velázquez-Quesada, 2010]. However, there is a problem with such definitions: how does an agent come to know something? The agent would need to have some implicit knowledge in advance to learn something explicitly.

Example 29. Let $\mathcal{K} = \{\top\}$ and $\mathcal{B} = \{\top, \phi, \psi, \chi\}$ (note that $\mathcal{K} \subseteq \mathcal{B}$), if ϕ, ψ and χ are consistent, then they can be drawn as follows:

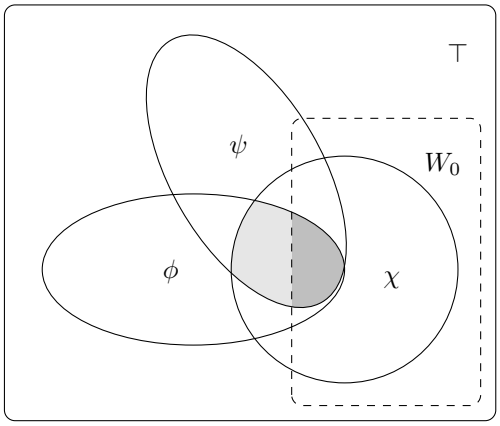

Figure 3.1: Example of consistent beliefs

Here, the lighter grayed area together with the darker grayed area (the intersection of the ellipses and the circle) represents what follows from the explicit beliefs. To obtain implicit beliefs we would need to intersect it with W_0, the obtained implicit beliefs are colored in dark gray in the picture.

If the explicit beliefs happen to be inconsistent, then the intersection will be empty and the implicit beliefs will be everything.

Example 30. Suppose now that $\mathcal{K} = \{\top\}$ and $\mathcal{B} = \{\top, \phi, \psi, \phi \to \bot\}$. First notice that since we do not require our explicit belief sets to be closed under Modus Ponens, this is a well-defined $\mathcal{B}$. Assuming that ϕ and ψ are again consistent, we could have a picture like Figure 3.2.

One can see that the intersection of ϕ and $\phi \to \bot$ is now empty, so independently of what W_0 is, the agent implicitly believes every formula.

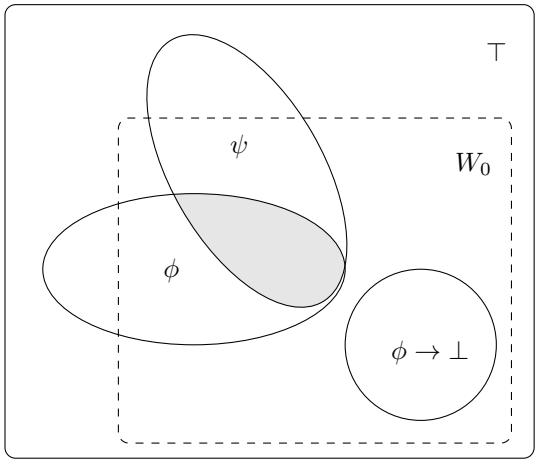

Figure 3.2: Example of inconsistent beliefs

Realistic agents can operate with only finitely many formulas explicitly, they can have some schemes of rules that potentially produce infinite amount of sentences, but at every given period of time, the agent considers only finitely many formulas.

Definition 31 (Feasible models)**.** Let $\mathfrak{M} = (W, W_0, \mathcal{B}, \mathcal{K}, V)$ be an explicit belief model. We say that $\mathfrak{M}$ is a *feasible model* if $|\mathcal{B}| < \omega$. $\triangleleft$

By definition, $\mathcal{K} \subseteq \mathcal{B}$, therefore, the number of explicitly known formulas will be finite as well. We will restrict our attention to models with finite amount of explicitly considered formulas, though all our results hold for the general models.

3.2 Axiomatization

Interestingly, models of explicit beliefs naturally validate axiom schemes that correspond to nice properties of knowledge and belief that one may want to have.

For convenience, we will stick with single-agent models, though most of our results are easily generalized to a multi-agent version. We will discuss what changes in a multi-agent version later.

Proposition 32 (Validities)**.** *The following formulas are valid on all explicit belief models.*

$$S5 \text{ axioms for } K^i \tag{3.1}$$
$$K45 \text{ axioms for } B^i \tag{3.2}$$
$$\neg B^e \bot \tag{3.3}$$
$$K^e \top \tag{3.4}$$
$$K^i\phi \to B^i\phi \tag{3.5}$$
$$K^e\phi \to B^e\phi \tag{3.6}$$
$$K^e\phi \to K^i\phi \tag{3.7}$$
$$B^e\phi \to B^i\phi \tag{3.8}$$
$$K^e\phi \to K^iK^e\phi \tag{3.9}$$
$$\neg K^e\phi \to K^i\neg K^e\phi \tag{3.10}$$
$$B^e\phi \to K^iB^e\phi \tag{3.11}$$
$$\neg B^e\phi \to K^i\neg B^e\phi \tag{3.12}$$
$$B^i\phi \to K^iB^i\phi \tag{3.13}$$
$$\neg B^i\phi \to K^i\neg B^i\phi \tag{3.14}$$

The smallest logic containing the above axioms and closed under specified rules will be denoted by EBL.

Proof. Let $\mathfrak{M} = (W, W_0, \mathcal{B}, \mathcal{K}, V)$ be an explicit belief model and $w \in W$.

We start with the $S5$ axioms for implicit knowledge. Reflexivity axiom is expressed with $K^i\phi \to \phi$, so assume $w \models K^i\phi$, we need to check that $w \models \phi$. But this trivially holds because ϕ is true in every world by the definition of implicit knowledge. For transitivity, suppose $w \models K^i\phi$, we already noticed that definitions of knowledge and belief are global in the sense that if the agent knows or believes something in some world, he believes it in each world, so we have that $\forall w' \in W, w' \models K^i\phi$. By the definition of implicit knowledge, this means that $w \models K^iK^i\phi$, so transitivity holds. To show Euclideaness, suppose $w \models \neg K^i\phi$, we need to show that $w \models K^i\neg K^i\phi$. If the agent does not know ϕ at a world, it also holds globally—otherwise there would exist a world in which the agent does believe ϕ and this would hold globally and would contradict the assumption. So, similarly to the previous proof, we have that $\forall w' \in W, w' \models \neg K^i\phi$ and, by the definition of implicit knowledge, $w \models K^i\neg K^i\phi$. It remains to show that the K axiom holds. Suppose $w \models K^i(\phi \to \psi)$ and $w \models K^i\phi$, we have to show that $w \models K^i\psi$. By the definition of implicit knowledge, we have that at every world w', $w' \models \phi \to \psi$ and $w' \models \phi$ respectively. Applying Modus Ponens, we get $w' \models \psi$. By definition of implicit knowledge, $K^i\psi$ holds at every world, in particular, it holds at w.

Axioms $\neg B^e\bot$ and $K^e\top$ hold because of the requirements that we forced on our models.

All the implicit introspection axioms (3.9)-(3.14) represent globality of knowledge and belief. They can be seen directly from the definitions. We will

discuss only $B^i\phi \to K^iB^i\phi$ because other proofs are similar. Suppose $w \models B^i\phi$, but $w \not\models K^iB^i\phi$. It would mean that $\exists w' \in W$, $w' \models \neg B^i\phi$. But B^i was defined globally and, thus, $B^i\phi$ cannot hold at w. This contradicts the initial assumption.

Before we proceed to the proof of $K45$ axioms for B^i, we need to show that $K^i\phi \to B^i\phi$ holds. So suppose $w \models K^i\phi$, this means that ϕ is true in every world, in particular, in all the worlds from the intersection of truth sets of explicit beliefs and W_0 that defines the implicit beliefs. So $w \models B^i\phi$.

Now we can continue and show that the $K45$ axioms hold for B^i. Axiom K holds because implicit beliefs are defined as what follows from the explicit beliefs together with the background beliefs. So if $\phi \to \psi$ and ϕ both follow, obviously, ψ follows as well by Modus Ponens. To show transitivity, suppose $w \models B^i\phi$. From the introspection axiom $B^i\phi \to K^iB^i\phi$, we have that $w \models K^iB^i\phi$. Using $K^i\phi \to B^i\phi$, we get $w \models B^iB^i\phi$ as required. It remains to show Euclideaness. Suppose $w \models \neg B^i\phi$, similarly to the previous reasoning, we first get $w \models K^i\neg B^i\phi$ and then $w \models B^i\neg B^i\phi$.

It is easy to see why the remaining (3.6)-(3.8) axioms hold. The two axioms $K^e\phi \to B^e\phi$ and $K^e\phi \to K^i\phi$ are valid by the definition of our model. Since implicit beliefs are defined as explicit ones together with the ground beliefs, $B^e\phi \to B^i\phi$ is valid by the definition of implicit belief. □

One may also want to have explicit introspection of knowledge or belief, or properties like $K^e\phi \to K^eK^i\phi$. Such properties do not hold unless we close $\mathcal{B}$ and $\mathcal{K}$ on these operations.

Theorem 33 (Completeness for EBL)**.** *The logic $\mathcal{L}$ is completely axiomatized (on explicit belief models) by the following system of axioms and rules:*

$$\text{S5 axioms and rules for } K^i \tag{3.15}$$
$$\text{K45 axioms and rules for } B^i \tag{3.16}$$
$$\neg B^e \bot \tag{3.17}$$
$$K^e \top \tag{3.18}$$
$$K^i\phi \to B^i\phi \tag{3.19}$$
$$K^e\phi \to B^e\phi \tag{3.20}$$
$$K^e\phi \to K^i\phi \tag{3.21}$$
$$B^e\phi \to B^i\phi \tag{3.22}$$
$$K^e\phi \to K^iK^e\phi \tag{3.23}$$
$$\neg K^e\phi \to K^i\neg K^e\phi \tag{3.24}$$
$$B^e\phi \to K^iB^e\phi \tag{3.25}$$
$$\neg B^e\phi \to K^i\neg B^e\phi \tag{3.26}$$
$$B^i\phi \to K^iB^i\phi \tag{3.27}$$
$$\neg B^i\phi \to K^i\neg B^i\phi \tag{3.28}$$

Proof. Soundness is straightforward—we already showed the validity of the axioms in Proposition 32.

As for completeness, it goes via a canonical *pseudo-model* which is in fact an EB-model. The idea is that we construct a model that will satisfy all the conditions and axioms we have in our models of explicit beliefs, but the interpretations of knowledge and belief will use the standard Kripke semantics, so that we can easily prove the completeness with respect to a relational model.

As usual [Blackburn *et al.*, 2001], we have to show that any consistent set of formulas is satisfiable on the canonical model. Before we proceed, let us introduce some notation. If T is a theory,

$$\mathcal{B}(T) = \{\phi : B^e\phi \in T\}$$

which is all the explicitly believed formulas that belong to the theory. And, accordingly,

$$\mathcal{K}(T) = \{\phi : K^e\phi \in T\}$$

Definition 34 (Pseudo-model)**.** Let us fix some consistent set of formulas Φ_0 in the language $\mathcal{L}$ and maximal consistent theory T_0 extending Φ_0. A *pseudo-model* for Φ_0 is a tuple $\overline{\mathfrak{M}} = (W, W_0, \mathcal{B}, \mathcal{K}, \to_0, \to^{\mathcal{B}}, \to^{\mathcal{K}}, \to, \sim, V)$ where

$W = \{\text{T maximal consistent theory} : \forall\phi(K^i\phi \in T \Leftrightarrow K^i\phi \in T_0)\}$

The set of possible worlds consists of the maximal EBL-consistent theories that agree with Φ_0 on implicit knowledge.

$\mathcal{B} = \mathcal{B}(T_0)$

The set of explicitly believed formulas agrees with the explicitly believed formulas of Φ_0.

$\mathcal{K} = \mathcal{K}(T_0)$

The set of explicitly known formulas agrees with the explicitly known formulas of Φ_0.

$\sim$ is a relation on $W \times W$: $T \sim S$ iff $\forall\phi(K^i\phi \in T \Rightarrow \phi \in S)$

This relation corresponds to the implicit knowledge.

$\rightarrow$ is a relation on $W \times W$: $T \rightarrow S$ iff $\forall\phi(B^i\phi \in T \Rightarrow \phi \in S)$

Similarly, this relation corresponds to the implicit belief.

$\rightarrow_0$ is a relation on $W \times W$: $T \rightarrow_0 S$ iff $T \rightarrow S$

The worlds in W_0 can be accessed via this relation.

$W_0 = \{S : \exists T, T \rightarrow_0 S\}$

All worlds (theories) that are reachable by $\rightarrow_0$ from some world.

$\sim^{\mathcal{K}}$ is a relation on $W \times W$: $T \sim^{\mathcal{K}} S$ iff $\forall\phi(\phi \in \mathcal{K}(T) \Rightarrow \phi \in S)$

This relation comes from the explicit knowledge.

$\rightarrow^{\mathcal{B}}$ is a relation on $W \times W$: $T \rightarrow^{\mathcal{B}} S$ iff $\forall\phi(\phi \in \mathcal{B}(T) \Rightarrow \phi \in S)$

In a similar way, this relation comes from the explicit belief.

V is a valuation function.

$\triangleleft$

Notice that since we may have inconsistent beliefs, there may not exist such a consistent S for $T \rightarrow S$, but this is not a problem because our models are not serial.

Semantics for knowledge and belief is defined using these relations, as it is usually defined in a Kripke model.

To prove completeness for the EB-models, we have to show completeness for the pseudo-models, and then justify why EB-models are pseudo-models. We must verify that all the conditions we had on our explicit belief models still hold, check the validity of axioms, prove Existence Lemmas for all modalities, and prove the Truth Lemma.

Let us first check that EB-models are pseudo-models. We start with checking that there is an accessibility relation arrow of implicit belief if, and only if, we have both background belief relation and explicit belief relation arrows.

Lemma 35. *Let $\overline{\mathfrak{M}}$ be a pseudo-model, then $T \rightarrow S$ iff $T \rightarrow_0 S$ and $T \rightarrow^{\mathcal{B}} S$.*

Proof. Let us first check that from $T \to S$ it also follows $T \to_0 S$ together with $T \to^{\mathcal{B}} S$. By definition, $\to_0$ coincides with $\to$, it remains to show the second part. Suppose $T \to S$, and suppose $\phi \in \mathcal{B}(T)$, which means $B^e\phi \in T$. By axiom (3.8), $B^i\phi$ is also in T, which by the definition of $\to$, implies $\phi \in S$. This finishes the left-to-right direction.

For another direction, suppose we have $T \to_0 S$ and $T \to^{\mathcal{B}} S$, since $\to_0$ and $\to$ are the same relations, $T \to S$ holds trivially. □

Proposition 36. *EB-models "are" pseudo-models.*

Proof. We have to show that the conditions on explicit and implicit knowledge and belief we had in an EB-model still hold in a pseudo-model. Roughly, one needs to check:

(i) $\to$ "is" the implicit belief relation (according to the properties defined for the EB-models)

(ii) $\sim$ "is" the implicit knowledge relation

(iii) $\to^{\mathcal{B}}$ "is" the explicit belief relation

(iv) $\to^{\mathcal{K}}$ "is" the explicit knowledge relation

The first condition follows from Lemma 35. The second condition follows directly from how we defined the possible worlds: the maximal consistent theories that agree with T_0 on the implicitly known formulas. From that definition, it also follows that these theories agree with T_0 on the explicit belief and knowledge as well: this $\forall\phi(K^e\phi \in T \Leftrightarrow K^e\phi \in T_0)$ holds because of the axiom $K^e \to K^i$, whereas fulfillment of this condition $\forall\phi(B^e\phi \in T \Leftrightarrow B^e\phi \in T_0)$ follows from the $K^e \to B^e$ axiom. This guarantees the conditions (iii) and (iv). □

We still have to check that Existence Lemmas hold for all modalities, and to prove the Truth Lemma. The Existence Lemmas go by the standard argument since all modalities we have are normal modalities. The proof of the Truth Lemma is also standard, by the same argument.

Now it remains to show that our canonical model validates the axioms of EBL logic. The axioms for K45 and S5 are known to be canonical[Blackburn *et al.*, 2001]. Validity of other axioms follows from the fact that worlds in the canonical model are MCS that are consistent with all the corresponding axioms. Consider, for example, the dual version of the axiom (3.19): $< B^i > \phi \to < K^i > \phi$ and suppose $T \to S$ and $\phi \in S$. Then, by the definition of $\to$, $< B^i > \phi \in T$, but T is a MCS that is consistent with all the axioms and, in particular, with $< B^i > \phi \to < K^i > \phi$, so by MP, $< K^i > \phi \in T$. The remaining axioms are valid by similar arguments.

As a consequence, $T_0 \models_{\overline{\mathfrak{M}}} \Phi_0$. The desired completeness result for the pseudo-models (as well as EB-models) follows.

□

3.3 Multi-Agent Models

We have seen how inconsistent beliefs can be represented in a single-agent structure. But multi-agent scenarios are also natural. As promised, we discuss what changes in the multi-agent case. Suppose $\mathcal{A} = \{1, \ldots, n\}$ is a finite set of agents. We define a multi-agent language by introducing modalities for each agent in the obvious way.

Definition 37 (Multi-agent Language)**.** Let At be a set of atomic propositions. Formulas ϕ of language $\mathcal{L}$ are given by

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid B^e_a\phi \mid B^i_a\phi \mid K^e_a\phi \mid K^i_a\phi$$

with $p \in \mathsf{At}$. Other Boolean connectives $\vee$, $\rightarrow$, $\leftrightarrow$, as well as the existential modal operators for knowledge and belief are defined as usual. $\triangleleft$

As for models, we have to add an equivalence relation to the previously defined model to represent *information clusters* of the agents. Another change is how we represent background beliefs: in a single-agent model those beliefs can be represented as a subset of possible worlds, but in the multi-agent situation, we need to distinguish between the background beliefs of different agents. We thus add a separate relation $\rightarrow^0_a$ for each agent that accesses his background beliefs. Multi-agent models are not much harder.

Definition 38 (Multi-agent Model of Explicit Beliefs)**.** A *multi-agent explicit belief model* is a tuple $\mathfrak{M} = (W, \{\sim_a\}_{a\in\mathcal{A}}, \{\rightarrow^0_a\}_{a\in\mathcal{A}}, \{\mathcal{B}_a\}_{a\in\mathcal{A}}, \{\mathcal{K}_a\}_{a\in\mathcal{A}}, V)$ where

W is a non-empty set of possible worlds.

$\sim_a \subseteq W \times W$ is an equivalence relation representing the implicit knowledge of agent a.

$\rightarrow^0_a \subseteq \sim_a$ represents (the accessible worlds that are consistent with) the agents a's background beliefs or "biases".

$\mathcal{B}_a : W \rightarrow \mathscr{P}(\mathscr{P}(\mathcal{L}))$ is a function mapping any world $w \in W$ to some set $\mathcal{B}_a(w) \subseteq \mathscr{P}(\mathcal{L})$, representing the set of sentences that are explicitly believed by agent a at world w.

$\mathcal{K}_a : W \rightarrow \mathscr{P}(\mathscr{P}(\mathcal{L}))$ is a function mapping any world $w \in W$ to some set $\mathcal{K}_a(w) \subseteq \mathscr{P}(\mathcal{L})$, representing the set of sentences that are explicitly known by agent a at world w. Subject to the requirement that $\mathcal{K}_a(w) \subseteq \mathcal{B}_a(w)$ (i.e. everything that is explicitly known is also explicitly believed).

$V : \mathsf{At} \rightarrow \mathscr{P}(W)$ is a valuation function.

The following conditions are forced on these models to ensure the wanted properties:

$$\begin{gathered}
\rightarrow_a^0 \text{ has to be transitive and Euclidean} \\
\bot \notin \mathcal{B}_a(w) \\
\top \in \mathcal{K}_a(w) \\
w \sim_a v \text{ implies } \mathcal{B}_a(w) = \mathcal{B}_a(v) \\
w \sim_a v \text{ implies } \mathcal{K}_a(w) = \mathcal{K}_a(v)
\end{gathered}$$

◁

Let us introduce some abbreviations before we continue

$$\begin{aligned}
w(a) &= \{w' : w \sim_a w'\} \\
w^0(a) &= \{w' : w \rightarrow_a^0 w'\}
\end{aligned}$$

Formulas will be interpreted similarly at a world, with the additional requirement that the world belongs to the information cell of an agent.

Definition 39 (Truth in Multi-agent Models)**.** Consider a multi-agent explicit belief model $\mathfrak{M} = (W, \{\sim_a\}_{a\in\mathcal{A}}, \{\rightarrow_a^0\}_{a\in\mathcal{A}}, \{\mathcal{B}_a\}_{a\in\mathcal{A}}, \{\mathcal{K}_a\}_{a\in\mathcal{A}}, V)$. Truth of a formula $\phi \in \mathcal{L}$ is defined inductively as follows:

$\mathfrak{M}, w \models p$ iff $w \in V(p)$

$\mathfrak{M}, w \models \neg\phi$ iff $\mathfrak{M}, w \not\models \phi$

$\mathfrak{M}, w \models \phi \wedge \psi$ iff $\mathfrak{M}, w \models \phi$ and $\mathfrak{M}, w \models \psi$

$\mathfrak{M}, w \models B_a^e\phi$ iff $\phi \in \mathcal{B}_a(w)$

$\mathfrak{M}, w \models B_a^i\phi$ iff $w' \models \phi$ for all $w' \in \{v : \mathfrak{M}, v \models \theta \text{ for all } \theta \in \mathcal{B}_a(w)\} \cap w^0(a)$

$\mathfrak{M}, w \models K_a^e\phi$ iff $\phi \in \mathcal{K}_a(w)$

$\mathfrak{M}, w \models K_a^i\phi$ iff $w' \models \phi$ for all $w' \in w(a)$

The truth set of ϕ is the set of worlds $[\![\phi]\!]_{\mathfrak{M}} = \{w : \mathfrak{M}, w \models \phi\}$. Standard logical notions of *satisfiability* and *validity* are defined as usual. ◁

The proof of completeness of the logic will go similarly, only minor changes need to be done to assure that we restrict agents to their information cells. Therefore, we will not bore the reader with a similar proof here.

Theorem 40 (Completeness for EBL_n)**.** *The logic* EBL_n *is completely axiomatized (on multi-agent explicit belief models) by the following system of axioms and rules:*

$$\textit{S5 axioms and rules for } K_a^i$$
$$\textit{K45 axioms and rules for } B_a^i$$
$$\neg B_a^e \bot$$
$$K_a^e \top$$
$$K_a^i \phi \to B_a^i \phi$$
$$K_a^e \phi \to B_a^e \phi$$
$$K_a^e \phi \to K_a^i \phi$$
$$B_a^e \phi \to B_a^i \phi$$
$$K_a^e \phi \to K_a^i K_a^e \phi$$
$$\neg K_a^e \phi \to K_a^i \neg K_a^e \phi$$
$$B_a^e \phi \to K_a^i B_a^e \phi$$
$$\neg B_a^e \phi \to K_a^i \neg B_a^e \phi$$
$$B_a^i \phi \to K_a^i B_a^i \phi$$
$$\neg B_a^i \phi \to K_a^i \neg B_a^i \phi$$

Proof. The proof proceeds as in Theorem 33. □

3.4 The Problem of Belief Dynamics

The aim of such explicit belief models is to allow reasoning about inconsistent beliefs if an agent is not explicitly aware of the inconsistency. Where believing in inconsistency means $B^e \bot$ (one could also have, for instance, $B^e(\phi \wedge \neg\phi)$).

But now other questions arise. What is the "last step" before the contradiction? And how to correct it? In the sense that which part of the contradiction causes the problem. There might be not enough structure on the model yet to deal with this problem. The sentences are "equal" in a sense and we are stuck in loops. If we remove the actual last step in our reasoning, we go back to the previous situation. Consider the following example:

Example 41. Let $\mathfrak{M} = (W, W_0, \mathcal{B}, \mathcal{K}, V)$ be an explicit belief model, suppose also that $\mathcal{B} \coloneqq \{\top, \phi_1, \ldots, \phi_n, \phi_1 \wedge \ldots \wedge \phi_n, \phi_1 \wedge \ldots \wedge \phi_n \to \bot\}$. Assume for a moment that explicit beliefs are closed under finite conjunctions. The only rational decision would be to give up one of the ϕ_i. But there is also a dual problem—we cannot give up one of the conjuncts and keep the conjunction unless the agent is not aware of the rule that conjunction implies its conjuncts.

So we might want to have not only closure under finite conjunctions, but also a requirement that $\phi \wedge \psi \in \mathcal{B}$ iff $\phi \in \mathcal{B}$ and $\psi \in \mathcal{B}$ for a rational enough agent and if $\mathcal{B}$ set is entirely in the agent's *working memory* (with preventing formation of infinite amount of equivalent conjunctions). Then, one will have

to give up a conjunct if, and only if, he gives up a conjunction. If the agent is certainly sure $K^e(\phi_1 \wedge \ldots \wedge \phi_n \to \bot)$, he must give up one of the ϕ_i together with the conjunction. In other words, the agent keeps everything he knows. Nevertheless, it is still not entirely clear which conjunct to give up if we do not have a plausibility relation on the sentences.

Conjunction is not the only rule that leads to obvious contradictions, another problem is the "last step" before Modus Ponens. We will write $+_{\mathcal{K}}\phi$ for *hard updates* in the sense that formulas are added to the agent's explicit knowledge, and $+_{\mathcal{B}}\phi$ for *soft update*, when the agent adds the formula only to his explicit beliefs. The following example illustrates another problem.

Example 42. Let $\mathfrak{M} = (W, W_0, \mathcal{B}, \mathcal{K}, V)$ be an explicit belief model, and, say, the agent believed $\mathcal{B} = \{\top, p, p \to q\}$, then, after applying MP, $\mathcal{B} = \{\top, p, p \to q, q\}$, also, from some other source he learned $+_{\mathcal{K}}(q \to \bot)$, so $(q \to \bot) \in \mathcal{K}$ for simplicity. We cannot just undo the result of applying MP, we have to drop some assumptions. The problem is, if we drop q, we will get $\mathcal{B} = \{\top, p, p \to q, q \to \bot\}$ and after another application of MP we will encounter the same problem.

A possible solution could be if an agent could learn instances of rules, for example, $(\{p, q\} \mapsto p \wedge q) \in \mathcal{K}$ or $(\{p, p \to q\} \mapsto q) \in \mathcal{K}$. Then we can require closure of $\mathcal{B}$ under these rules. There will be only finitely many instances of such rules, so $\mathcal{B}$ remains finite. With these assumptions, if the agent gives up, he gives up on both q and one of $\{p, p \to q\}$ (if they are not in $\mathcal{K}$). More generally, if $\{(\phi_1 \to (\phi_2 \to \ldots \to (\phi_n \to \phi)\ldots)), \neg\phi, \neg\phi \to (\phi \to \bot)\} \subseteq \mathcal{K}$ and $\phi_1, \ldots, \phi_n \in \mathcal{B}$, we have to give up one of $\phi_1, \ldots, \phi_n$ (if they are not in $\mathcal{K}$).

For the static part of our logic, we would have to require the closure under such known instances of rules for both $\mathcal{K}$ and $\mathcal{B}$. Then, for the dynamic part, when the agent becomes aware of a contradiction, he keeps the $\mathcal{K}$ and revises his belief set such that the new belief set $\mathcal{B}'$ is again closed under all the constraints and $\mathcal{K} \subseteq \mathcal{B}' \subseteq \mathcal{B}$. It is important that agents know only instances of rules and sets of sentences are not closed under substitution, otherwise agents become logical omniscient.

However, there are still open questions: How to resolve an inconsistency if an agent does not know one of the explicit rules? How does he come to know these instances? Without some kind of *patterns* and simple substitutions this type of knowledge is a bit mystical. On another hand, deciding which substitutions are simple enough to become aware of is also just a dynamic step and it can be reduced to becoming aware of one of the instances of some rule.

With these new constraints, the semantics of upgrades needs to be justified as well. If we started with the model $\mathfrak{M} = (W, W_0, \mathcal{B}, \mathcal{K}, V)$, after the hard update $+_{\mathcal{K}}\phi$ we first close $\mathcal{K} \cup \{\phi\}$ and $\mathcal{B} \cup \{\phi\}$ under all rules, suppose we got $\mathcal{K}'$ and $\mathcal{B}'$ respectively. If $\bot$ is in the newly obtained $\mathcal{B}'$, then choose some $\mathcal{B}''$ such that $\mathcal{K} \cup \{\phi\} \subseteq \mathcal{B}'' \subseteq \mathcal{B} \cup \{\phi\}$ and $\mathcal{B}''$ satisfies all the constraints, otherwise go to $\mathfrak{M}' = (W, W_0, \mathcal{B}', \mathcal{K}', V)$. Similarly for the $+_{\mathcal{B}}\phi$, except that $\mathcal{K}$ remains the same.

Depending on the type of operation we want to model with $+_{\mathcal{B}}\phi$ upgrade, ϕ does not necessarily remain in $\mathcal{B}$ after the consistency check according to

the constraints. And, what is even more unpleasant, there is no clear way of keeping all the explicit beliefs that do not participate in the contradiction (in the sense that they can be excluded while the contradiction still holds) when choosing $\mathcal{B}''$. We could consider one of the maximal (in the sense that it cannot be extended any longer) consistent $\mathcal{B}''$s. In case humans sometimes really decide non-deterministically, this will model their unconscious choices. Although, there is a problem with determining the so-called *recursion axioms* for such non-deterministic choices. There could be another possibility—by introducing and modifying the plausibility relation on the explicit beliefs set, one could see how different decisions lead to different scenarios of developing of a human being.

But let us introduce a modified model that will deal with the problems during the updates discussed above.

Explicit Evidence Models

To avoid the above-mentioned problems during the dynamic steps, we will look at the similar model, but instead of having explicit beliefs, we will have a similar syntactic set of *evidence pieces*. This is in the spirit of van Benthem and Pacuit's work [2011], but more syntactic, logical non-omniscient version: we do not want it to be monotone. An agent does not necessarily believe those evidences, they are just some facts that the agent gathers from various sources. One way to do so is to replace $\mathcal{B}$ with the evidence set $\mathcal{E}$. The idea is that the explicit beliefs of the agent are defined via his evidence pieces and, thus, are *recomputed* automatically after every update. Let us for convenience and simplicity focus on the single-agent model only in this section.

4.1 A Static Logic

We leave the syntax for this refined version as before.

Definition 43 (Explicit Evidence Language)**.** Let At be a set of atomic propositions. Formulas ϕ of language $\mathcal{L}$ are given by

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid B^e\phi \mid B^i\phi \mid K^e\phi \mid K^i\phi$$

with $p \in \mathsf{At}$. Other Boolean connectives $\vee$, $\rightarrow$, $\leftrightarrow$, as well as the existential modal operators for knowledge and belief are defined as usual. $\triangleleft$

The model and semantics, however, changes in the following way.

Definition 44 (Semantic Model of Explicit Evidence)**.** An *explicit evidence model* (EE-model) is a tuple $\mathfrak{M} = (W, W_0, \mathcal{E}_s, \mathcal{E}_h, V)$ where

W is a set of possible worlds.

$W_0 \subseteq W$ is a set of worlds that represents the agent's background beliefs or "biases".

$\mathcal{E}_s \subseteq \mathscr{P}(\mathcal{L})$ is a set of formulas that represent agent's (soft) evidence pieces.

$\mathcal{E}_h \subseteq \mathcal{E}_s$ is a set of hard evidence pieces.

$V : \mathsf{At} \rightarrow \mathscr{P}(W)$ is a valuation function.

The following condition is imposed on the models:

$$\top \in \mathcal{E}_h(w)$$

◁

Note that unlike the restriction on the previous models, here the soft evidence set may contain $\bot$ (but the explicit beliefs will not). We will use $\bot$ to "mark" a contradiction, but this is only a convention because we did not want to restrict ourselves—it could have been any formula.

The idea is that we can now think of $\mathcal{K}$ also as of an evidence set—it is hard evidence set $\mathcal{E}_h$ that is infallibly true, whereas $\mathcal{E}_s$ is a soft evidence set: an agent is not absolutely certain about those evidence pieces and they may even be inconsistent with each other.

Definition 45 (Closed Evidence)**.** A set $F \subseteq \mathcal{E}_s$ of (soft) evidence pieces is said to be *closed* if, and only if, it includes all the hard evidence (i.e. $\mathcal{E}_h \subseteq F$) and it is closed under Modus Ponens within $\mathcal{E}_s$ (i.e. if ϕ and $\phi \to \psi$ belong to F and ψ belongs to $\mathcal{E}_s$, then ψ belongs to F). ◁

Definition 46 (Q-max Evidence)**.** A set $F \subseteq \mathcal{E}_s$ of (soft) evidence pieces is said to be *maximal closed quasi-consistent set* (or *q-max*, for short) if it is (1) closed (in the above sense), (2) quasi-consistent, and (3) maximal with respect to properties (1) and (2) (i.e. for every other closed quasi-consistent set F', if $F \subseteq F' \subseteq \mathcal{E}_s$, then $F' = F$). ◁

Since we do not have explicit beliefs in the model anymore, we have to define them. Soft evidences are on the more abstract level than beliefs, the evidence pieces play a role of the derivations an agent made so far, and beliefs are encoded there. We say that the agent *explicitly believes* a formula at some world if, and only if, that formula belongs to the intersection of all maximal closed quasi-consistent sets:

$$\mathcal{B} := \bigcap \{F : F \text{ is q-max}\}$$

Let us use this abbreviation for the explicit belief set from now on.

The choice of such definition naturally arises from our line of research—since we assume the agent has the fast "working" memory where he can easily compute even exponential things. According to this definition, the agent stays safe and cautious and sticks with what is included to every maximal closed quasi-consistent set of evidence pieces. This can be seen as the appropriate syntactic counterpart of the van Benthem and Pacuit's definition of Maximal Consistent Evidence [2011].

Definition 47 (Truth in the Explicit Evidence Models)**.** Consider an explicit evidence model $\mathfrak{M} = (W, W_0, \mathcal{E}_s, \mathcal{E}_h, V)$. Truth of a formula $\phi \in \mathcal{L}$ is defined inductively as follows:

$$\mathfrak{M}, w \models p \quad \text{iff} \quad w \in V(p)$$

$\mathfrak{M}, w \models \neg\phi \quad \text{iff} \quad \mathfrak{M}, w \not\models \phi$

$\mathfrak{M}, w \models \phi \wedge \psi \quad \text{iff} \quad \mathfrak{M}, w \models \phi \ \text{ and } \ \mathfrak{M}, w \models \psi$

$\mathfrak{M}, w \models B^e\phi \quad \text{iff} \quad \phi \in \mathcal{B}$

$\mathfrak{M}, w \models B^i\phi \quad \text{iff} \quad w' \models \phi \ \text{ for all } \ w' \in \{v : \mathfrak{M}, v \models \theta \text{ for all } \theta \in \mathcal{B}\} \cap W_0$

$\mathfrak{M}, w \models K^e\phi \quad \text{iff} \quad \phi \in \mathcal{E}_h$

$\mathfrak{M}, w \models K^i\phi \quad \text{iff} \quad w' \models \phi \ \text{ for all } \ w' \in W$

Where $\mathcal{B}$ is defined as discussed above. The truth set of ϕ is the set of worlds $[\![\phi]\!]_{\mathfrak{M}} = \{w : \mathfrak{M}, w \models \phi\}$. Standard logical notions of *satisfiability* and *validity* are defined as usual. ◁

As before, we assume that

$$\phi \in \mathcal{E}_h \Rightarrow [\![\phi]\!] = W$$

required $\bot \notin \mathcal{B}$ will hold automatically by construction.

Example 48. Let $\mathcal{E}_h = \{\top\}$ and $\mathcal{E}_s = \{\top, \phi, \phi \to \bot, \neg\phi\}$. To compute $\mathcal{B}$, we have to find all the maximal consistent subsets of $\mathcal{E}_s$ closed under MP within it and check the intersection. In this example, the only maximal quasi-consistent subset of $\mathcal{E}_s$ is $\mathcal{E}_s$ itself (remember quasi-consistency means $\bot \notin \mathcal{E}_s$), so $\mathcal{B} = \{\top, \phi, \phi \to \bot, \neg\phi\}$. Evidences can be pictured as follows:

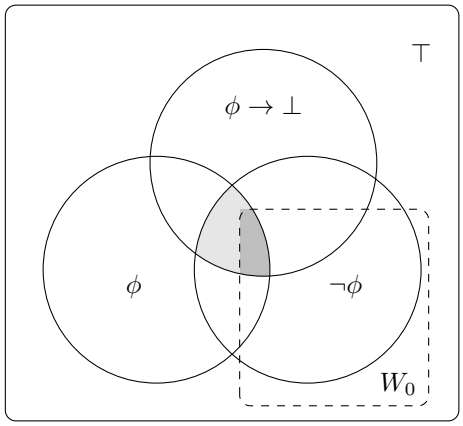

Figure 4.1: Example of consistent beliefs

Firstly, note that in this picture there are evidences, not beliefs as before. The only maximal consistent subset of these pieces of evidence is represented as

the union of the lighter grayed area and the darker grayed area (the intersection of the circles). This union is also what follows from the explicit beliefs of the agent. As before, to obtain implicit beliefs we would need to intersect it with W_0, the obtained implicit beliefs are colored in dark gray in the picture.

It is also interesting to look at what happens in case when $\mathcal{E}_s$ has several maximal subsets.

Example 49. Suppose now that $\mathcal{E}_s = \{\top, \phi \to \psi, \phi, \psi, \psi \to \bot, \bot\}$ and $\mathcal{E}_h = \{\top\}$. There are three maximal consistent subsets of $\mathcal{E}_s$:

$$\mathcal{E}_1 = \{\top, \phi \to \psi, \phi, \psi\}$$
$$\mathcal{E}_2 = \{\top, \phi \to \psi, \psi \to \bot\}$$
$$\mathcal{E}_3 = \{\top, \phi, \psi \to \bot\}$$

It is easy to see what their (syntactic) intersection is: $\mathcal{B} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 = \{\top\}$. So the implicit beliefs coincide with the ground beliefs in this case.

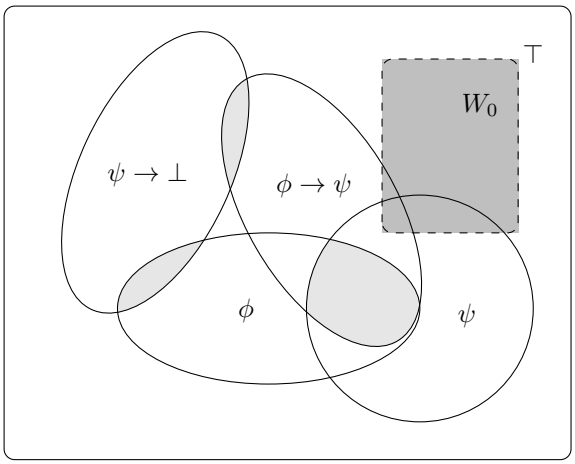

Figure 4.2: Computing beliefs

The crucial difference between the EB and EE models is that in the EB-models agents could only consciously believe some sentences and know that they are not obviously inconsistent, whereas in the EE-models agents not only have basis for their beliefs, but can temporarily store the information they know is inconsistent.

4.2 Axiomatization

If our syntax does not talk about evidence pieces, then completeness for this new evidence model is not hard to show.

Proposition 50. *Every* EB*-model is a* EE*-model.*

Proof. Suppose $\mathfrak{M} = (W, W_0, \mathcal{B}, \mathcal{K}, V)$ is an EB-model. One gets an EE-model defining $\mathcal{E}_h = \mathcal{K}$ and $\mathcal{E}_s = \mathcal{B}$. In this case, if we calculate the explicit belief set for the newly obtained EE-model, it will coincide with $\mathcal{B}$ in the model we started with because $\mathcal{B}$ was quasi-consistent by definition, and by taking $\mathcal{E}_s$ to be $\mathcal{B}$, we assure that the only maximal quasi-consistent subset of $\mathcal{E}_s$ is $\mathcal{E}_s$ itself. □

Definition 51 (Model Transformations)**.** We define two functions (model transformators) that transform EB-models to EE-models and vice versa.

Let $\mathfrak{M}' = (W, W_0, \mathcal{B}, \mathcal{K}, V)$ be an EB-model. The corresponding EE-model is obtained by $t_1(\mathfrak{M}') = (W, W_0, \mathcal{B}, \mathcal{K}, V)$.

Let $\mathfrak{M}'' = (W, W_0, \mathcal{E}_s, \mathcal{E}_h, V)$ be an EE-model and $\mathcal{B}''$ the explicit belief set calculated from $\mathcal{E}_s$ according to the definition. Then the corresponding EB-model is obtained by $t_2(\mathfrak{M}'') = (W, W_0, \mathcal{B}'', \mathcal{E}_h, V)$.

◁

Proposition 52 (Truth Preservation)**.** *The two model transformators given in Definition 51 preserve truth values of all formulas in language* $\mathcal{L}$ *at all worlds.*

Proof. Let $\phi \in \mathcal{L}$.

Using Proposition 50, the statement follows for t_1 immediately since every EB-model is also an EE-model.

The proof for t_2 is by induction on the complexity of ϕ.

. The Boolean connectives and the base case are straightforward because the transformators are defined in such a way that the possible worlds, background belief sets and valuations remain the same.

. In case of modalities, it is also easy to see from the definition of how the explicit belief set is defined in EE-models and the definitions of truth of both models that there is a direct correspondence.

□

Theorem 53 (Completeness for EEL)**.** *The logic $\mathcal{L}$ is completely axiomatized (on explicit evidence models) by the following system of axioms and rules:*

$$\begin{gathered}
\textit{S5 axioms and rules for } K^i \\
\textit{K45 axioms and rules for } B^i \\
\neg B^e \bot \\
K^e \top \\
K^i\phi \to B^i\phi \\
K^e\phi \to B^e\phi \\
K^e\phi \to K^i\phi \\
B^e\phi \to B^i\phi \\
K^e\phi \to K^iK^e\phi \\
\neg K^e\phi \to K^i\neg K^e\phi \\
B^e\phi \to K^iB^e\phi \\
\neg B^e\phi \to K^i\neg B^e\phi \\
B^i\phi \to K^iB^i\phi \\
\neg B^i\phi \to K^i\neg B^i\phi
\end{gathered}$$

The smallest logic containing the above axioms and closed under specified rules will be denoted by EEL*.*

Proof. All the axioms are still sound because the map t_2 from Definition 51 is truth preserving, by Proposition 52, and together with the soundness of the axioms for EB-models (as proved in Theorem 33), gives us soundness for EE-models.

For completeness, we use that the map t_1 from EB-models to EE-models is truth-preserving. This, with the completeness of the system for EB-models (which we had already proved in Theorem 33), gives us completeness for EE-models. This completes the proof. □

4.3 Evidence Dynamics

One of our main intentions in this work was to allow agents to resolve the inconsistency once they become aware of it. Where by awareness we assume that the agent gets an explicit contradiction as an evidence. To model this, we have to express some *actions* that describe how the agent becomes aware of new information.

In this section, we will focus on some of the possible evidence dynamics which are also called *updates*. The updates are informational actions that *change* the original model. Some of these changes may remove the possible worlds of the agent, another—will just modify the explicit information of the agent.

After the update, one has to check that the model remains *persistent* under this operation which means that the model still satisfies all the conditions we imposed on the models. We are aware of Moore-like paradoxes when knowledge is defined as a primitive notion. The problem is that when one defines explicit knowledge as a separate set, each formula in that set has to be true by assumption. But then, updates can change the truth value of an epistemic assertion. For instance, consider $\mathcal{E}_h = \{T, \neg K^e\phi\}$, the formula $\neg K^e\phi$ says that the agent does not explicitly know that ϕ, which is true. But if after the update he learns ϕ by simply adding it to his prior knowledge, $\mathcal{E}_h$ becomes inconsistent.

There are different ways of fixing the described problem. One could have in the agent's knowledge set only sentences that do not contain the explicit knowledge modalities [van Benthem and Velázquez-Quesada, 2010], this is a significant restriction on the formulas the agent can learn, though. Or, we could introduce a new modality Y in our language which would represent "one step in past", and before each update prefix the K^e modalities with this new modality [Baltag *et al.*, 2014], this solution also seems to add a bunch of nested Y's already after several updates even if they were not needed.

So, let us for simplicity assume that $\mathcal{E}_h$ can only contain purely factual sentences. Moreover, we will update our models with only propositional, non-epistemic information (There is a possibility of a more relaxed constraint, when in all evidence formulas the negation symbols should only occur in front of atomic letters. But we choose the stricter one here to avoid the need to prove persistency).

Definition 54 (Propositional Fragment of $\mathcal{L}$)**.** Let $\mathcal{L}_0$ be the fragment of $\mathcal{L}$ formed only by using Boolean connectives and atomic sentences. ◁

First of all, one could look at the usual DEL updates. Consider, for example, the operation of update. This type of update models the situation when the agent receives hard piece of evidence from an infallible source. The standard way of modeling this is via model restriction. Since by assumption the information source is infallible, after the update with ϕ only the worlds that satisfy ϕ are left.

Definition 55 (Update with Hard Evidence)**.** Let $\mathfrak{M} = (W, W_0, \mathcal{E}_s, \mathcal{E}_h, V)$ be an explicit evidence model, and ϕ a formula in $\mathcal{L}_0$. We define the model $\mathfrak{M}^{!\phi} = (W^{!\phi}, W_0^{!\phi}, \mathcal{E}_s^{!\phi}, \mathcal{E}_h^{!\phi}, V^{!\phi})$ as follows.

$$W^{!\phi} = \{w \in W : \mathfrak{M}, w \models \phi\}$$

$$W_0^{!\phi} = W_0 \cap W^{!\phi}$$

$$V^{!\phi}(p) = V(p) \cap W^{!\phi}$$

$$\mathcal{E}_h^{!\phi} = \mathcal{E}_h \cup \{\phi\}$$

$$\mathcal{E}_s^{!\phi} = \mathcal{E}_s \cup \{\phi\}$$

◁

This operation is described by a dynamic modality $[!\phi]\psi$ stating that "ψ is true after the update of ϕ". The truth condition is as usual:

$$\mathfrak{M}, w \models [!\phi]\psi \quad \text{iff} \quad \mathfrak{M}, w \models \phi \;\; \text{implies} \;\; \mathfrak{M}^{!\phi}, w \models \psi$$

Here, the precondition is that ϕ is true at w, otherwise it would be obvious that the source is lying.

Proposition 56. *The update with hard evidence operation transforms* EE*-models into* EE*-models.*

Proof. All the axioms were valid naturally on the EE-models, from the definition of the model. The only conditions we have are $\top \in \mathcal{E}_h \subseteq \mathcal{E}_s$. They will still hold because the evidence sets of the agents are given separately and do not depend on possible worlds. □

Another, more natural for our models, scenario is when the agent learns new pieces of evidence. They may be consistent or inconsistent with the previously learned information. It can be even explicit inconsistency $\bot$. There are two possibilities: either the agent adds ϕ to his explicit knowledge $\mathcal{E}_h$, or he just accepts ϕ as a piece of evidence and adds it only to $\mathcal{E}_s$.

Definition 57 (Hard Evidence Addition)**.** Let $\mathfrak{M} = (W, W_0, \mathcal{E}_s, \mathcal{E}_h, V)$ be an explicit evidence model, and ϕ a formula in $\mathcal{L}_0$. We define the modified model $\mathfrak{M}^{+_h\phi} = (W^{+_h\phi}, W_0^{+_h\phi}, \mathcal{E}_s^{+_h\phi}, \mathcal{E}_h^{+_h\phi}, V^{+_h\phi})$ as follows.

$$W^{+_h\phi} = W$$

$$W_0^{+_h\phi} = W_0$$

$$V^{+_h\phi} = V$$

$$\mathcal{E}_h^{+_h\phi} = \mathcal{E}_h \cup \{\phi\}$$

$$\mathcal{E}_s^{+_h\phi} = \mathcal{E}_s \cup \{\phi\}$$

◁

This operation is described by a dynamic modality $[+_h\phi]\psi$ stating that "ψ is true after ϕ is accepted as a hard evidence". The truth condition is straightforward:

$$\mathfrak{M}, w \models [+_h\phi]\psi \quad \text{iff} \quad \forall w' \in W, \mathfrak{M}, w' \models \phi \;\; \text{implies} \;\; \mathfrak{M}^{+_h\phi}, w \models \psi$$

Here, the precondition is that ϕ is true everywhere because the agent already knew ϕ implicitly. By the action, he only becomes explicitly aware of something that was already implicitly known.

One can think of this operation as of becoming aware of hard evidence that was already implicitly known. It is, in a way, an action of introspection because it assumes the agent already implicitly knew ϕ, and now he just becomes aware of it. His implicit knowledge of ϕ becomes explicit knowledge. In a way, this is a special case of update with hard evidence: it is just $!\phi$ in the special case that ϕ was already implicitly known.

Proposition 58. *The hard evidence addition operation transforms* EE*-models into* EE*-models.*

Proof. The conditions to be checked are $\top \in \mathcal{E}_h \subseteq \mathcal{E}_s$. They obviously remain true from the definition of the hard evidence addition. □

Another version of evidence addition is when the agent becomes aware of some piece of evidence.

Definition 59 (Soft Evidence Addition)**.** Let $\mathfrak{M} = (W, W_0, \mathcal{E}_s, \mathcal{E}_h, V)$ be an explicit evidence model, and ϕ a formula in $\mathcal{L}_0$. We define the modified model $\mathfrak{M}^{+_s\phi} = (W^{+_s\phi}, W_0^{+_s\phi}, \mathcal{E}_s^{+_s\phi}, \mathcal{E}_h^{+_s\phi}, V^{+_s\phi})$ as follows.

$$W^{+_s\phi} = W$$

$$W_0^{+_s\phi} = W_0$$

$$V^{+_s\phi} = V$$

$$\mathcal{E}_h^{+_s\phi} = \mathcal{E}_h$$

$$\mathcal{E}_s^{+_s\phi} = \mathcal{E}_s \cup \{\phi\}$$

◁

This operation is described by a dynamic modality $[+_s\phi]\psi$ stating that "ψ is true after ϕ is accepted as a soft evidence". The interpretation of this formula is as follows:

$$\mathfrak{M}, w \models [+_s\phi]\psi \quad \text{iff} \quad \mathfrak{M}^{+_s\phi}, w \models \psi$$

Note that there is no precondition here—the agent simply adds a new piece of evidence to his soft evidence set. Besides the fact that our model is a kind of "syntactic" version of van Benthem and Pacuit's evidence models, there are other subtle differences. They assume evidence sets are non-empty. In contrast, we do not assume that evidence formulas have non-empty extension, and not even that they are different from $\bot$ (so they can be "obviously" inconsistent). As a result, our definition of $+_s\phi$ modality is simpler, since we do not need the precondition that ϕ is true at some state. Indeed, the case of $+_s\bot$ is very important in our setting: it can be used to model becoming aware of an inconsistency.

As before, the assumptions on the model are preserved under this action.

Proposition 60. *The soft evidence addition operation transforms* EE*-models into* EE*-models.*

Proof. The arguments are as in Proposition 58. □

Note that we allow an update with $+\bot$. By adding $+\bot$ an agent may become aware of some of the inconsistencies. Since we do not have justifications [Artemov, 2008] in our logic, adding $\bot$ means that an agent becomes aware of

all inconsistencies such that both $\phi, \phi \to \bot \in \mathcal{E}_s$, since all such constructions are "equal" in a sense. This is the price of not having "derivations" in the logic.

It is interesting to distinguish between the two types of evidence addition. Hard evidence addition happens only when ϕ is already known implicitly, such an upgrade can happen when the agent becomes aware of some fact (learns it for sure). In case of soft evidence addition, this can be truly new information that was not known even implicitly. It may, of course, happen that ϕ was already implicitly known for the soft evidence addition as well, then it could be seen as becoming aware of a piece of evidence, this implicitly known evidence may not be even believed.

With these operations agents now can change their beliefs, even from ϕ to $\neg\phi$:

Example 61 (Belief Revision)**.** Suppose $\mathfrak{M}$ is an EE-model and $\mathcal{E}_s = \{\top, \phi\}$. Since the only maximal quasi-consistent subset of $\mathcal{E}_s$ is $\mathcal{E}_s$ itself, it follows that $\mathcal{B} = \mathcal{E}_s = \{\top, \phi\}$. Which means the agent explicitly believes ϕ. Now suppose the following sequence of evidence addition was performed: $+_s(\neg\phi)$, $+_s(\phi \to \bot)$, $+_s\bot$. As a result, the agent's (soft) evidence pieces became $\mathcal{E}'_s = \{\top, \phi, \neg\phi, \phi \to \bot, \bot\}$. Let us calculate what the explicit beliefs are. There are two maximal quasi-consistent subsets of $\mathcal{E}'_s$:

$$\mathcal{E}_1 = \{\top, \phi, \neg\phi\}$$
$$\mathcal{E}_2 = \{\top, \neg\phi, \phi \to \bot\}$$

So $\mathcal{B} = \mathcal{E}_1 \cap \mathcal{E}_2 = \{\top, \neg\phi\}$ by definition. One can see that the agent's explicit belief has changed from ϕ to $\neg\phi$.

We would like to model belief revision of realistic agents, and realistic agents cannot hold all the information they learn forever. In real life, agents do forget some things from time to time. It, therefore, makes perfect sense to consider evidence removal operation.

Definition 62 (Evidence Removal)**.** Let $\mathfrak{M} = (W, W_0, \mathcal{E}_s, \mathcal{E}_h, V)$ be an explicit evidence model, and ϕ a formula in $\mathcal{L}$. We will define the modified model $\mathfrak{M}^{-\phi} = (W^{-\phi}, W_0^{-\phi}, \mathcal{E}_s^{-\phi}, \mathcal{E}_h^{-\phi}, V^{-\phi})$ as follows.

$$W^{-\phi} = W$$

$$W_0^{-\phi} = W_0$$

$$V^{-\phi} = V$$

$$\mathcal{E}_h^{-\phi} = \mathcal{E}_h \setminus \{\phi\}$$

$$\mathcal{E}_s^{-\phi} = \mathcal{E}_s \setminus \{\phi\}$$

$\triangleleft$

This operation is described by a dynamic modality $[-\phi]\psi$ stating that "ψ is true after ϕ is removed as an evidence". The interpretation of this formula is as follows:

$$\mathfrak{M}, w \models [-\phi]\psi \quad \text{iff} \quad \mathfrak{M}^{-\phi}, w \models \psi$$

We would like to stress that even though a rational agent should not remove his hard evidences, we had in mind realistic agents who do not have "perfect memory". These explicit sets will grow very fast and become too unrealistic (at least for humans) if the agent remembers everything. So one could think of this operation more like of "forgetting things".

Proposition 63. *The evidence removal operation transforms* EE*-models into* EE*-models.*

Proof. As before, holds almost trivially. □

With the above-described dynamic operations, the agent can become aware of the inconsistency and is able to fix it (this happens automatically). This means that both explicit and implicit beliefs of the agent can become consistent.

4.4 Towards Recursion Axioms: variants of conditional belief

As in the previous section, we restrict ourselves to *propositional-evidence models* only, and restrict our dynamic operators $!\phi$, $+_s\phi$, $+_h\phi$ to formulas ϕ that are *purely propositional* (i.e. Boolean combinations of propositional letters). To obtain recursion axioms for our dynamic operators, we will need to follow van Benthem and Pacuit [2011] in introducing various *conditional* versions of our (explicit and implicit) operators. We sketch here the first steps. Note that in this section we use subscripts B_i and B_e (instead of superscripts as in previous sections) for the indices i (for implicit belief) and e (for explicit belief). We do this to avoid the clash with the superscripts for conditional beliefs, thus being able to write B_i^ϕ and B_e^ϕ without any clashes.

A set $F \subseteq \mathcal{E}_s \subseteq \mathcal{L}_0$ is said to be *closed with respect to a sentence* $\phi \in \mathcal{L}_0$ (or *ϕ-closed*, for short) if, and only if, it includes all the hard evidence (i.e. $\mathcal{E}_h \subseteq F$) and it is closed under Modus Ponens within $\mathcal{E}_s \cup \{\phi\}$ (i.e. if ψ and $\psi \to \theta$ belong to F and θ belongs to $\mathcal{E}_s$, then θ is in F).

The *ϕ-closure* of a set $F \subseteq \mathcal{E}_s$ is the least ϕ-closed set $F^{\phi-\mathrm{cl}} \subseteq \mathcal{E}_s \cup \{\phi\}$ with $F \subseteq F^{\phi-\mathrm{cl}}$.

A set $F \subseteq \mathcal{E}_s$ is said to be *quasi-consistent with* ϕ (or *ϕ-qcons*, for short) if the ϕ-closure of $F \cup \{\phi\}$ is quasi-consistent (i.e. $\bot \notin (F \cup \{\phi\})^{\phi-\mathrm{cl}}$).

A set $F \subseteq \mathcal{E}_s$ is said to be *maximal ϕ-closed ϕ-qcons* (or *ϕ-qmax*, for short) if it is (1) ϕ-closed, (2) quasi-consistent with ϕ, and (3) a maximal subset of $\mathcal{E}_s$ with respect to properties (1) and (2).

We put now

$$\mathcal{B}^{+\phi} = \bigcap \{(F \cup \{\phi\})^{\phi-\mathrm{cl}} : F \subseteq \mathcal{E}_s \text{ is } \phi\text{-qmax } \}.$$

The set of worlds $[\![\mathcal{B}^{+\phi}]\!] := \bigcap \{[\![\theta]\!] : \theta \in \mathcal{B}^{+\phi}\} = \{w \in W : \mathfrak{M} \models \theta \text{ for all } \theta \in \mathcal{B}^{+\phi}\}$ will give us *implicit beliefs conditional on ϕ*.

Given these notations, we can introduce operators $B_i^{+\phi}\psi$ and $B_e^{+\phi}\psi$ for (implicit and explicit) conditional belief:

$$\mathfrak{M}, w \models B_e^{+\phi}\psi \quad \text{iff} \quad \psi \in \mathcal{B}^{+\phi}$$

and

$$\mathfrak{M}, w \models B_i^{+\phi}\psi \quad \text{iff} \quad [\![\mathcal{B}^{+\phi}]\!] \cap W_0 \subseteq [\![\psi]\!].$$

Following again van Benthem and Pacuit [2011], we can also introduce *negatively-conditioned beliefs* $B^{-\phi}\psi$, based on *incompatibility with ϕ*:

A set $F \subseteq \mathcal{E}_s$ is said to be *ϕ-incompatible* (or *ϕ-inc*, for short) if it is not quasi-consistent with ϕ; i.e. if $\bot \in (F \cup \{\phi\})^{\phi-\mathrm{cl}}$.

We put now

$$\mathcal{B}^{-\phi} = \bigcap \{F \subseteq \mathcal{E}_s : F \text{ qmax with } \bot \in (F \cup \{\phi\})^{\phi-\mathrm{cl}}\}$$

for the intersection of all sets that are ϕ-incompatible. The set of worlds $[\![\mathcal{B}^{-\phi}]\!] := \bigcap\{[\![\theta]\!] : \theta \in \mathcal{B}^{-\phi}\}$ will model *implicit negatively-conditioned beliefs*:

$$\mathfrak{M}, w \models B_i^{-\phi}\psi \quad \text{iff} \quad [\![\mathcal{B}^{-\phi}]\!] \cap W_0 \subseteq [\![\psi]\!],$$

while explicit *negatively-conditioned beliefs* are defined by putting

$$\mathfrak{M}, w \models B_e^{-\phi}\psi \quad \text{iff} \quad \psi \in \mathcal{B}^{-\phi}.$$

Lemma 64. *We have the following equivalencies:*

(1) *$F \subseteq \mathcal{E}_s$ is ϕ-qmax iff $(F \cup \{\phi\})^{\phi-cl} \subseteq \mathcal{E}_s^{+\phi}$ is q-max (within $\mathcal{E}_s^{+\phi}$);*

(2) *$F \subseteq \mathcal{E}_s$ is ϕ-incompatible q-max iff $(F \cup \{\phi\})^{\phi-cl} \subseteq \mathcal{E}_s^{+\phi}$ is q-max (within $\mathcal{E}_s^{+\phi}$) s.t. $\phi \notin F$;*

(1') *$F' \subseteq \mathcal{E}_s^{+\phi}$ is q-max (within $\mathcal{E}_s^{+\phi}$) with $\phi \in F'$ iff $F' = (F \cup \{\phi\})^{\phi-cl}$ for some q-max $F \subseteq \mathcal{E}_s$;*

(2') *$F' \subseteq \mathcal{E}_s^{+\phi}$ is q-max (within $\mathcal{E}_s^{+\phi}$) with $\phi \notin F'$ iff $F' \subseteq \mathcal{E}_s$ is ϕ-incompatible and q-max (within $\mathcal{E}_s$).*

Proposition 65. *For every $\phi \in \mathcal{L}_0$ and $\psi \in \mathcal{L}$, the formulas*

$$[!\phi]K_i\psi \;\leftrightarrow\; (\phi \to K_i[!\phi]\psi)$$

$$[+_h\phi]K_i\psi \;\leftrightarrow\; (K_i\phi \to K_i[+_h\phi]\psi)$$

$$[+_s\phi]K_i\psi \;\leftrightarrow\; K_i[+_s\phi]\psi$$

$$[!\phi]B_i\psi \;\leftrightarrow\; \Big(\phi \to B_i^{+\phi}[!\phi]\psi\Big)$$

$$[+_h\phi]B_i\psi \;\leftrightarrow\; \Big(K_i\phi \to B_i^{+\phi}[+_h\phi]\psi\Big)$$

$$[+_s\phi]B_i\psi \;\leftrightarrow\; \Big(B_i^{+\phi}[+_s\phi]\psi \wedge B_i^{-\phi}[+_s\phi]\psi\Big)$$

are valid on all propositional-evidence models.

Proof. Most of these validities are similar to the ones in van Benthem and Pacuit [2011], and their proof is similar. We prove here only the last one:

For the *left-to-right direction* assume that $\mathfrak{M}, w \models [+_s\phi]B_i\psi$, and so $\mathfrak{M}^{+_s\phi}, w \models B_i\psi$. Using part (1) of Lemma 64, we have that

$$\{(F \cup \{\phi\})^{\phi-\mathrm{cl}} : F \subseteq \mathcal{E}_s \text{ is } \phi\text{-qmax }\} \subseteq \{F' \subseteq \mathcal{E}_s^{+\phi} : F' \text{ is qmax }\}.$$

Using the fact that the set-intersection operator reverses inclusions (i.e. it is anti-monotonic with respect to inclusion), we obtain that

$$\bigcap\{(F \cup \{\phi\})^{\phi-\mathrm{cl}} : F \subseteq \mathcal{E}_s \text{ is } \phi\text{-qmax }\} \supseteq \bigcap\{F' \subseteq \mathcal{E}_s^{+\phi} : F' \text{ is qmax }\}.$$

Using again the anti-monotonicity of $\bigcup$, we get that

$$[\![\bigcap\{(F\cup\{\phi\})^{\phi-\mathrm{cl}} : F\subseteq\mathcal{E}_s \text{ is } \phi\text{-qmax }\}]\!] =$$

$$=\bigcap\{[\![\theta]\!] : \theta\in\bigcap\{(F\cup\{\phi\})^{\phi-\mathrm{cl}} : F\subseteq\mathcal{E}_s \text{ is } \phi\text{-qmax }\}\}\subseteq$$

$$\subseteq\bigcap\{[\![\theta]\!] : \theta\in\bigcap\{F'\subseteq\mathcal{E}_s^{+\phi} : F' \text{ is qmax }\}\} = [\![\bigcap\{F'\subseteq\mathcal{E}_s^{+\phi} : F' \text{ is qmax }\}]\!],$$

and we conclude that

$$[\![\bigcap\{(F\cup\{\phi\})^{\phi-\mathrm{cl}} : F\subseteq\mathcal{E}_s \text{ is } \phi\text{-qmax }\}]\!]\cap W_0\subseteq$$

$$\subseteq[\![\bigcap\{F'\subseteq\mathcal{E}_s^{+\phi} : F' \text{ is qmax }\}]\!]\cap W_0\subseteq[\![\psi]\!]_{\mathfrak{M}^{+_s\phi}}$$

(where we used the fact that $\mathfrak{M}^{+_s\phi}, w\models B_i\psi$), and thus we obtain $\mathfrak{M}, w\models B_i^{+\phi}[+_s\phi]\psi$. In a completely similar way, we can show that $\mathfrak{M}, w\models B_i^{-\phi}[+_s\phi]\psi$ using part (2) of Lemma 64, and thus prove the right-hand side of the desired equivalence. The *right-to-left direction* uses in a similar manner parts (3) and (4) of Lemma 64. □

Proposition 66. *For every $\phi,\psi\in\mathcal{L}_0$, the formulas*

$$[!\phi]K_e\psi \;\leftrightarrow\; (\phi\to K_e\psi), \quad \textit{for } \psi\neq\phi$$
$$[!\phi]K_e\phi$$
$$[+_h\phi]K_e\psi \;\leftrightarrow\; (K_i\phi\to K_e\psi), \quad \textit{for } \psi\neq\phi$$
$$[+_h\phi]K_e\phi$$

$$[!\phi]B_e\psi \;\leftrightarrow\; \big(\phi\to B_e^{+\phi}\psi\big)$$
$$[+_h\phi]B_e\psi \;\leftrightarrow\; \big(K_i\phi\to B_e^{+\phi}\psi\big)$$
$$[+_s\phi]B_e\psi \;\leftrightarrow\; \big(B_e^{+\phi}\psi\wedge B_e^{-\phi}\psi\big)$$

are valid on all propositional-evidence models. (Note that on propositional-evidence models, the syntax of $K_e\psi$ and $B_e\psi$ has *to be restricted to $\psi\in\mathcal{L}_0$, since only purely propositional formulas can be evidence in these models!)*

To explain the absence of dynamic modalities from the right-hand sides of the last two validities, note that purely propositional formulas $\psi\in\mathcal{L}_0$ are persistent under updates and evidence-addition, so that $[!\phi]\psi$ is equivalent to $\phi\to\psi$, and $[+\phi]\psi$ is equivalent to ψ, for such formulas.

To obtain full reductions, the reduction axioms have to be extended to conditional beliefs $B^\psi\theta$ after an update or evidence addition. This can be easily done (as usual) for updates and hard addition. But for soft addition, one would need to follow again van Benthem and Pacuit [2011] in further extending the language to a more complex conditional $B^{\phi,\alpha}\psi$ that combines q-consistency with ϕ and incompatibility with ψ. We leave the details for further work.

4.5 Solving the Problem of Belief Dynamics

Back to our original problem of becoming aware of a contradiction, the explicit evidence models solve the problem of deciding which step to consider the last one and how not to go in loops after removing the problematic sentences. As we already mentioned above, this happens automatically, because the explicit belief set is recomputed every time depending on what the explicit evidence pieces are.

Here is a simple example, showing a possible chain of updates of the model and how the explicit (and, thus, implicit) beliefs change at each step.

Example 67 (Belief Dynamics)**.** Let $\mathfrak{M} = (W, W_0, \mathcal{E}_s, \mathcal{E}_h, V)$ be an explicit evidence model.

Step 1 Suppose that initially the agent's explicit evidence pieces form a quasi-consistent set: $\mathcal{E}_s = \{\top, \phi \wedge \neg\phi\}$, where $\mathcal{E}_h = \{\top\}$. By definition, this means that the explicitly believed formulas will be the same as the evidence pieces $\mathcal{B} = \mathcal{E}_s = \{\top, \phi \wedge \neg\phi\}$ since there is only one maximal closed quasi-consistent subset of $\mathcal{E}_s$—$\mathcal{E}_s$ itself. Note that the implicit beliefs of the agent are already inconsistent at this stage because a contradiction follows from $\phi \wedge \neg\phi$, hence, $B^i\bot$.

Step 2 Now suppose the agent learns a piece of hard evidence (that is, knowledge) $+_h(\phi \wedge \neg\phi \to \bot)$. This is a validity and is implicitly known, so the precondition holds and one can consider this action. The new explicit evidence sets will become $\mathcal{E}_s = \{\top, \phi \wedge \neg\phi, \phi \wedge \neg\phi \to \bot\}$ and $\mathcal{E}_h = \{\top, \phi \wedge \neg\phi \to \bot\}$. The soft evidence set is still quasi-consistent, so the explicitly believed formulas will again coincide with $\mathcal{E}_s$:

$$\mathcal{B} = \{\top, \phi \wedge \neg\phi, \phi \wedge \neg\phi \to \bot\}$$

The implicit beliefs remain inconsistent.

Step 3 When the agent suddenly becomes aware of the inconsistency $+_s\bot$, the situation noticeably changes. First, look at the evidence sets $\mathcal{E}_s = \{\top, \phi \wedge \neg\phi, \phi \wedge \neg\phi \to \bot, \bot\}$ and $\mathcal{E}_h = \{\top, \phi \wedge \neg\phi \to \bot\}$, the soft evidence set is no longer quasi-consistent since it contains $\bot$. Which implies that the explicit beliefs cannot be equal to the whole evidence set anymore. There are two maximal closed quasi-consistent subsets of $\mathcal{E}_s$:

$$\mathcal{E}_1 = \{\top, \phi \wedge \neg\phi\}$$
$$\mathcal{E}_2 = \{\top, \phi \wedge \neg\phi \to \bot\}$$

We cannot extend $\mathcal{E}_1$ with $\phi \wedge \neg\phi \to \bot$ because, by the closure within $\mathcal{E}_s$ condition, it would become inconsistent, and, by the same argument, $\mathcal{E}_2$ also cannot be extended further. Thus, $\mathcal{B} = \mathcal{E}_1 \cap \mathcal{E}_2 = \{\top\}$. There are several things to notice. The first, and the most important one, is that even though the explicit evidence set was inconsistent, the explicit beliefs

remain quasi-consistent (in this case even consistent). The second one is that if the background beliefs were consistent (and we can always assume this if we need it), then the implicit beliefs became consistent as well.

So, the explicit evidence models give us a natural way to perform explicit belief revision when obvious contradictions are derived in the evidence set of an agent.

The next step would be to find the so-called *recursion axioms* for the newly introduced dynamic modalities. These axioms relate the epistemic and doxastic modalities before, and after the updates and allow to push the dynamic modalities inside recursively. But since we have neither conditional belief nor evidence modalities in the language, this seems to be problematic for the belief modalities. So, we might want to think of adding additional modal operators to obtain a sound and complete logic for the extended dynamic language.

4.6 Additional modal operators

One may want to talk about the evidence sets themselves. Even if we do not want to do this for some reason, evidence modalities might be useful for defining reduction axioms for dynamic logic of upgrades. For this purpose, one can introduce several modalities, possibly again, explicit and implicit ones.

Definition 68. We say that a formula ϕ is *factive* at w if $\mathfrak{M}, w \models \phi$. $\triangleleft$

Let $\Box$ denote the evidence modality. Then the following modalities can be added to the language of EE-models.

Definition 69 (Explicit Evidence)**.** An *explicit evidence* is just any evidence the agent is aware of. $\triangleleft$

This notion is described with a formula $\Box^e\phi$ that we will read as "ϕ is an explicit evidence". The interpretation of the modality is straightforward:

$$\mathfrak{M}, w \models \Box^e\phi \quad \text{iff} \quad \phi \in \mathcal{E}_s$$

Definition 70 (Implicit Evidence)**.** An agent has an *implicit evidence* of a formula if it can be derived from the explicit evidences. $\triangleleft$

This will be denoted as $\Box^i\phi$, and we will read it as "ϕ is an implicit evidence". The interpretation is as follows:

$$\mathfrak{M}, w \models \Box^i\phi \quad \text{iff} \quad \exists \theta_1, \ldots, \theta_n \in \mathcal{E}_s \ \text{ such that } \ [\![\theta_1 \wedge \ldots \wedge \theta_n]\!] \subseteq [\![\phi]\!]$$

Definition 71 (Explicit Factive Evidence)**.** An *explicit factive evidence* at a world is a factive evidence the agent is aware of. $\triangleleft$

This type of evidence is described with a formula $\Box_0^e\phi$ that we will read as "ϕ is an explicit factive evidence". The interpretation of the modality is as for explicit evidence, but with additional requirement that ϕ is true at the world of interest:

$$\mathfrak{M}, w \models \Box_0^e\phi \quad \text{iff} \quad \phi \in \mathcal{E}_s \ \text{ and } \ w \in [\![\phi]\!]$$

Definition 72 (Implicit Factive Evidence)**.** An agent has an *implicit factive evidence* of a formula at a world if it can be derived from the explicit evidences and, in addition, the formula is factive. ◁

This will be denoted as $\Box_0^i\phi$, and we will read it as "ϕ is an implicit factive evidence". The interpretation is as follows:

$$\mathfrak{M}, w \models \Box_0^i\phi \quad \text{iff} \quad \exists\theta_1, \ldots, \theta_n \in \mathcal{E}_s \ \text{ such that } \ [\![\theta_1 \wedge \ldots \wedge \theta_n]\!] \subseteq [\![\phi]\!] \ \text{ and } \ w \in [\![\phi]\!]$$

The notion of evidence we have is not necessarily factive since evidences may be false (and even inconsistent). But if we add factive modalities to our language, it will allow us to express more powerful statements.

Conclusions

5.1 Summary of the Thesis

In this thesis we aimed to provide the first steps towards modeling one of the consequences of being a non-omniscient agent. Namely, the problem of holding inconsistent beliefs due to non-awareness of a contradiction. Working with the assumption that the agent remains rational, and that he does not find it rational to believe in explicit inconsistencies, we provided a model that in a sense corrects the explicit inconsistencies themselves.

Our proposed solution addresses this problem by providing a basis for agent's beliefs—syntactic pieces of evidence that an agent uses to justify his beliefs. Then, the explicit beliefs of an agent are computed using his explicit evidence set. The explicit belief set is purely syntactic as well, which allows an agent to hold any kind of sentences without identifying them with an inconsistency, unless it is indeed an explicit inconsistency.

In our models, implicit belief is a defined notion: it is defined as a closure of agent's explicit beliefs together with his prior background biases. From this it follows that implicit beliefs may be inconsistent (in the usual sense). We think, that defining implicit beliefs via explicit ones is more natural than treating them as an independent notion, and it makes perfect sense that the implicit beliefs of an agent may happen to be inconsistent at some point in time. Of course, these beliefs can become consistent if the agent manages to resolve the inconsistencies in his explicit beliefs.

Since we allow our agents to operate only with (finite) syntactic explicit information, our proposed explicit evidence models happen to resolve all the omniscience problems that epistemic logicians are usually concerned with. Closure properties need not hold at all for the explicit sets of knowledge and belief. As well as explicit introspection.

It is worthwhile to mention that neither B^e nor B^i satisfy the standard $KD45$ axioms for belief. Implicit beliefs do not satisfy the seriality axiom, as explained above, whereas explicit beliefs do not satisfy the K-axiom. Interestingly, B^e does satisfy the D-axiom in the sense that $\neg B^e \bot$ (but not $B^e\phi \to \neg B^e \neg\phi$). Consequently, one could argue that the standard notion of belief is a mixture of these two.

We have given an axiomatization for the (static) logic of explicit evidence

which is complete. The proof proceeded using canonical pseudo-models. Next, we presented the dynamic actions that describe change of models due to modifications in the evidence pieces. We saw various examples that illustrate how our models work. In the key part of this thesis, we showed how the problem of inconsistent belief revision is solved with the help of our models. Lastly, we discussed some possible extensions of the language of explicit evidence in order to provide sound and complete system for the extended dynamic language.

5.2 Open Questions and Further Work

In this thesis, we have motivated, defined, and begun to analyse a new logic. But we will not stop on this—there is still a lot of work to be done to get the desired results. In the following paragraphs, we list a number of open problems and put forward what we feel would be important areas of further investigation.

The axiomatization of static explicit evidence logic alone is not entirely satisfactory. Ideally, we would like to be able to give a complete axiomatization for our logic that also covers dynamic modalities. Thus, our first open question thus is to find a complete axiomatization of dynamic explicit evidence logic. This will probably require the definition of further modalities of conditional beliefs.

Another problem concerns Moore-like paradoxes if we do not restrict the set of explicit knowledge to purely factual sentences. One of the possible solutions would be to introduce some kind of history on the evidence set by prefixing the old evidences with suitable modalities during the update. Or, another option would be to find a way of redefining the set of explicit knowledge in case of updates.

As we have already mentioned, the drawback of the method which is used for fixing inconsistencies is that it takes care of all the present inconsistencies. Given that the idea is the agent always maintains only small amount of evidences at each given point in time, this is still acceptable. But one might want to allow an agent to become aware of some particular contradiction. In this case, we would need to add a notion of a derivation to our logic, this would help us to rule out inconsistencies that are characterized by specific derivations.

One could also look at versions of the explicit evidence logic where there is some kind of plausibility relation. This relation should be defined not on evidence pieces themselves because in finite case, one will not be able to have inconsistent beliefs (since there will be always a maximal plausible element).

Given the social nature of agents, we need extensions to group notions like common and distributed knowledge and belief. With this new approach we might be able to rationally hold and maintain inconsistent aggregated beliefs of agents. Even if all the agents are perfect reasoners and always have only consistent beliefs, the common beliefs might very well be inconsistent.

We will follow these lines of research in future work.

Bibliography

[Anderson and Belnap, 1976] A. R. Anderson and N. D. Belnap. *Entailment: The Logic of Relevance and Necessity*. Princeton University Press, 1976.

[Artemov, 2008] S. Artemov. The logic of justification. *The Review of Symbolic Logic*, 1(4):477–513, 2008.

[Baltag and Smets, 2008] A. Baltag and S. Smets. *In R. van Rooij and K. Apt (eds.), Texts in Logic and Games, Special Issue on New Perspectives on Games and Interaction*, volume 4, chapter The Logic of Conditional Doxastic Actions, pages 9–31. Amsterdam University Press, 2008.

[Baltag *et al.*, 2014] A. Baltag, B. Renne, and S. Smets. The logic of justified belief, explicit knowledge, and conclusive evidence. *Annals of Pure and Applied Logic*, 165(1):49–81, 2014.

[Baltag *et al.*, 2016] A. Baltag, N. Bezhanishvili, A. Özgün, and S. Smets. Justified belief and the topology of evidence. In *Proceedings WoLLIC 2016*. Springer, 2016.

[van Benthem and Pacuit, 2011] J. van Benthem and E. Pacuit. Dynamic logics of evidence-based beliefs. *Studia Logica*, 99:61–92, 2011.

[van Benthem and Velázquez-Quesada, 2010] J. van Benthem and F. R. Velázquez-Quesada. The dynamics of awareness. *Synthese*, 177(1):5–27, 2010.

[Blackburn *et al.*, 2001] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2001.

[van Ditmarsch *et al.*, 2007] H. van Ditmarsch, W. van Der Hoek, and B. Kooi. *Dynamic epistemic logic*. Springer, 2007.

[Fagin and Halpern, 1987] R. Fagin and J.Y. Halpern. Belief, awareness, and limited reasoning. *Artificial intelligence*, 34(1):39–76, 1987.

[Halpern and Pucella, 2011] J.Y. Halpern and R. Pucella. Dealing with logical omniscience: Expressiveness and pragmatics. *Artificial intelligence*, 175(1):220–235, 2011.

[Hansson (eds.), 2013] S. O. Hansson (eds.). *David Makinson on Classical Methods for Non-Classical Problems*. Springer Netherlands, 2013.

[Harman, 1986] G. Harman. Change in view: Principles of reasoning. MIT Press, 1986.

[Huang and Kwast, 2005] Z. Huang and K. Kwast. *In J. van Eijck (eds.), Logics in AI*, volume 478, chapter Awareness, negation and Logical omniscience, pages 282–300. Springer Berlin Heidelberg, 2005.

[Levesque, 1984] H.J. Levesque. A logic of implicit and explicit belief. In *Proceedings AAAI-84*, pages 198–202. AAAI Press, 1984.

[Lifschitz, 2008] V. Lifschitz. What is answer set programming? In *AAAI*, pages 1594–1597, 2008.

[Mares, 2013] E. Mares. *David Makinson on Classical Methods for Non-Classical Problems*, volume 3, chapter Liars, Lotteries, and Prefaces: Two Paraconsistent Accounts of Belief Change, pages 119–141. Springer Netherlands, 2013.

[Moreno, 1998] A. Moreno. Avoiding logical omniscience and perfect reasoning: a survey. *AI Communications*, 11(2):101–122, 1998.

[Pacuit, 2007] E. Pacuit. Neighborhood semantics for modal logic: An introduction. ESSLLI 2007 course notes, 2007.

[Priest, 2002] G. Priest. *Handbook of Philosophical Logic*, volume 6, chapter Paraconsistent logic, pages 287–393. Kluwer Academic Publishers, 2002.

[Stalnaker, 1968] R. C. Stalnaker. *In N. Rescher (ed.), Studies in Logical Theory*, chapter A Theory of Conditionals, pages 98–112. Oxford, Blackwell, 1968.

[Thimm, 2012] M. Thimm. *Probabilistic reasoning with incomplete and inconsistent beliefs*. IOS Press, 2012.

[Vardi, 1986] M. Y. Vardi. On epistemic logic and logical omniscience. In *TARK '86*, pages 293–305. Morgan Kaufmann Publishers Inc., 1986.

[Velázquez-Quesada, 2009] F. R. Velázquez-Quesada. Inference and update. *Synthese*, 169(2):283–300, 2009.

[Velázquez-Quesada, 2011] F. R. Velázquez-Quesada. *Small steps in dynamics of information*. PhD thesis, University of Amsterdam, 2011.

[Velázquez-Quesada, 2014] F. R. Velázquez-Quesada. Dynamic epistemic logic for implicit and explicit beliefs. *Journal of Logic, Language and Information*, 23(2):107–140, 2014.

[Wassermann, 2000] Renata Wassermann. *Resource-Bounded Belief Revision*. PhD thesis, University of Amsterdam, 2000.