Bachelor Artificial Intelligence

A proposed epistemic model representing human reasoning and knowledge

Creating a target logic that models logically non-omniscient, yet logically competent agents

Author:
Juul Petit

Supervisor:
Colin Caret
Second reader:
Daniel Cohnitz

Honours Thesis 15 ECTS
Faculty of Humanities
Utrecht University
June 25, 2021

Abstract

The aim of this paper is to find a logic that can model how agents reason about their knowledge. Accordingly, a model should be established that can solve logical omniscience and allow logical competence. Logical omniscience is the problem that assumes agents to have knowledge about all logical truths and about all consequences of their knowledge (Parikh 1987). In order to acquire the desired target model, existing models are tested on their capability to avoid logical omniscience while allowing logical competence. First, Kripke models are discussed, which fail to avoid logical omniscience (Fagin et al. 1995). Second, minimal models are introduced, also failing to circumvent logical omniscience completely (Chellas 1980). Third, two hyperintensional models are proposed; awareness logic and impossible worlds semantics (Fagin et al. 1995). Both models solve logical omniscience, but they assume agents to be logically incompetent. Still, human agents are capable of making some trivial inferences from their knowledge, making them logically competent (Cherniak 1981). Therefore, the target logic should add a concept that can simulate how agents reason about their knowledge. The target logic thus applies a dynamized version of the impossible worlds model, which models logically non-omniscient, yet logically competent agents (Bjerring and Skipper 2019). Furthermore, it models agents with different degrees of cognitive resources. Further research could be done to find other logics that can model human reasoning and knowledge.

Keywords: Logical omniscience - Logical competence - Hyperintensional models - Awareness logic - Impossible worlds - Dynamic epistemic logic

Contents

1 Introduction **3**
- 1.1 Relevance to Artificial Intelligence 4
- 1.2 Structure of the paper 5

2 Kripke semantics and logical omniscience **7**
- 2.1 Application to principles 8

3 Minimal models **12**
- 3.1 Application to principles 13

4 Hyperintensional models **16**
- 4.1 Awareness logic 18
  - 4.1.1 Application to principles 19
- 4.2 Impossible worlds semantics 21
  - 4.2.1 Application to principles 22

5 Logical competence **25**

6 Target logic **28**
- 6.1 Semantics of the target logic 29
- 6.2 Solving logical omniscience 35
- 6.3 Allowing logical competence 35

7 Conclusion and discussion **40**

8 Appendix **43**

9 Bibliography **44**

Chapter 1

Introduction

Epistemology is the branch of philosophy that studies the nature of knowledge (Sales and Pinto 2016). It is the foundation of a great number of current studies, among which Artificial Intelligence, and it is the patriarch of everything that is now studied in this field. The discipline of epistemic logic was originated by Hintikka in 1962 (Hintikka 1962), introducing theoretical reasoning tasks and formalizing knowledge.

The standard models used in issues concerning human knowledge nowadays are Kripke semantics. These models are based on possible world semantics, which contain a nonempty set of worlds W whose elements are considered to be possible worlds. There exist binary accessibility relations R between these worlds, such that a Kripke structure is as follows: $K = \langle W, R \rangle$ (Orlowska 1990). The theory of Kripke semantics was a major finding in the field of modal logic, since there did not exist a theory in this branch of logic yet.

Nevertheless, Kripke semantics give rise to a notion of knowledge, in which agents ought to be perfect logical reasoners. The model assumes that agents know all logical truths and all logical consequences of their knowledge (Parikh 1987). Thus, in Kripke semantics, agents are assumed to be logically omniscient (Fagin et al. 1995).

Logical omniscience is a well-known problem in logical epistemology. The word omniscience comes from Latin and is made up of the words 'omni', meaning 'all' and 'scientia', meaning 'knowledge', so omniscience actually means 'the knowledge of everything'. The main objection to logical omniscience is that it makes unrealistic assumptions about agents' reasoning ability. People are simply not logically omniscient; one can know a set of truths, without knowing all logical consequences that follow from these truths (Jago 2006).

The fact that people are not logically omniscient is due to a number of reasons. First, people are deficient in computational power. One cannot reason infinitely about the consequences of the facts of which he has knowledge. Human knowledge simply is finite, as are its reasoning powers (Hawke, Ozgun, and Berto 2019). Second, people are imperfect reasoners who can make false assumptions about facts they know or simply cannot see the obvious consequences of the

facts they know (Fagin et al. 1995). Therefore, the logical omniscience problem should be tackled in order to obtain a clear epistemic model that can explain the logic of human knowledge and human reasoning.

The goal of creating a model that avoids the problem of logical omniscience has been addressed in several ways. The impossible worlds semantics and awareness logic (Fagin et al. 1995), are examples of approaches that succeed in modelling agents who are not logically omniscient. Yet, these two models sacrifice too much trivial logical features agents are capable of, causing them to be logically incompetent. Logical competence is an important feature in understanding human reasoning. Ordinary humans are considered to be logically competent, since they are able to grasp the trivial logical consequences of their knowledge (Bjerring and Skipper 2019). For example, if someone knows ϕ and ψ, then he also knows ϕ and ψ individually. In chapter 5 this theory of logical competence will be further elaborated.

In this paper, the focus is on completing the difficult task of circumventing logical omniscience, without letting go of all properties of logical competence. Therefore, an attempt will be made to apply an epistemic model, the 'target logic', which models logically non-omniscient, yet logically competent agents. In order to achieve this goal, a number of models appropriate to tackle different aspects of the logical omniscience problem will be considered. These models will be analyzed to determine which properties are necessary to be preserved and which should be removed in the target logic. Eventually, a dynamic impossible worlds model, based on Bjerring and Skipper's model (2019), will be applied. The dynamic aspect of the model, which is missing in most epistemic models, is very important. It is this feature that enables the logic to model how reasoning can provide someone certain knowledge. By adding this component, the best applicable model is established that solves the problem of logical omniscience, without sacrificing every property of logical competence.

1.1 Relevance to Artificial Intelligence

The topic of epistemology is very important in the field of Artificial Intelligence. Artificial Intelligence is literally the study of artificially computing human intelligence. This discipline was originated by theorizing reasoning tasks and formalizing reasoning (Hintikka 1962). It is built upon epistemology, and always traceable to this theory of knowledge.

If it is our goal to artificially recreate human intelligence in computers, then it is necessary to be able to use a clear epistemic model that can explain the logic of human knowledge and human reasoning. When we are unable to understand how human agents reason and how human intelligence is put together, then it will be impossible to recreate it in a computer. In addition, reasoning tasks often involve reasoning about other agents such as humans. This makes it of great importance to understand how other agents reason, since wrongly comprehending other's knowledge in such tasks could lead to huge mistakes (Rendsvig and Symons 2021). Therefore, it is of necessity to create a model that allows

logical competence, since humans can make some inferences of their knowledge, but avoids logical omniscience, since humans do not know all logical truths and all logical consequences of their knowledge.

1.2 Structure of the paper

The goal of this paper is to build a target logic that models agents who are both logically competent and logically non-omniscient. The paper consists of four chapters leading to chapter 6 in which the target logic is presented. The first three chapters are focused on avoiding logical omniscience and chapter 5 is dedicated to preserving the properties that belong to logical competence, which should therefore be retained.

First, Kripke semantics will be formally explained in chapter 2. It will also be shown why Kripke semantics cause logical omniscience. This will be done by proving that Kripke semantics validate principles that cause logical omniscience. Thus, making the agent logically omniscient in these models.

In chapter 3, minimal models (Chellas 1980) will be discussed. In the subsection of this chapter, it is explained that minimal models, also called Montague-Scott structures or neighborhood semantics, solve a substantial part of logical omniscience. They make some false principles, validated by Kripke semantics, invalid. It is also discussed that minimal models are nevertheless not a complete solution to the logical omniscience problem. Because there still is a principle causing logical omniscience that is validated in minimal models, it does not avoid the problem in its entirety (Sedlár 2019).

Chapter 4 introduces the hyperintensional models. In the first subsection the awareness logic will be explained. This approach assumes that an agent has to be aware of a formula in order to achieve knowledge about it. This logic will be tested by applying it to the invalidities of chapter 2 and 3, which will show that this model solves logical omniscience.The second subsection presents the impossible worlds semantics, which solve the logical omniscience problem using an impossible worlds structure. Again, the model will be applied to principles and this will confirm that it circumvents logical omniscience.

In chapter 5 logical competence will be further explained using principles that are trivially true for all agents. It will also be shown that the hyperintensional models do not allow logical competence and therefore do not meet the requirements of our target logic.

The target logic will be demonstrated in chapter 6. It is a dynamized version of the impossible worlds model, inspired by Bjerring and Skipper (2019). The dynamics of the model ensure that agents can reason about their knowledge and can therefore gain new information by applying some number of inference rules to their current knowledge. The subsections of this chapter will show that the framework successfully avoids logically omniscient agents, while allowing them to be logically competent. This will be proved by applying the target logic to the principles that cause logical omniscience and to the principles that allow logical competence.

Finally, chapter 7 provides some concluding remarks and discusses further research in this area.

Chapter 2

Kripke semantics and logical omniscience

Kripke semantics are very useful frameworks in a lot of issues in the field of epistemic logic. It was developed by philosopher Saul Kripke, who has contributed to different theories in logic (Burgess 2011). Kripke models are the standard models used in modal logic. It models knowledge by using a non-empty set of possible worlds and adding accessibility relations between the elements in this set (Blackburn, Rijke, and Venema 2014). In every possible world, certain formulas are true and others are false. The possible worlds can be interpreted as states of information and the accessibility relations between these possible worlds define knowledge to an agent. In order to understand the formal definition of Kripke models, we first need to understand the structure of Kripke frames (Blackburn, Rijke, and Venema 2014):

Definition 2.1 - Kripke frame
A Kripke frame is a tuple $F = \langle W, R \rangle$, where
a. W is a non-empty set of possible worlds, and
b. $R \subseteq W \times W$, is the accessibility relation. If wRv, it means that v is accessible from w, or that w can access v.

Now that the structure of Kripke frames is clear, the formal structure of Kripke models can be explained:

Definition 2.2 - Kripke model
A Kripke model is a tuple $M = \langle W, R, V \rangle$, where
a. $\langle W, R \rangle$ is a Kripke frame, and
b. $V : W \to P(At)$ is a function, called the valuation function.

A Kripke model is based on a Kripke frame $\langle W, R \rangle$, but is complemented by the valuation function. This function assigns a truth value to the formulas

represented by the worlds in W. It states that for world w if $p \in V(w)$, this indicates that p is true in w (Ditmarsh, Hoek, and Kooi 2008). If a formula p is true in a world w, it can also be denoted $w \vDash p$. In Kripke semantics, an agent knows a fact in world w if and only if this fact is true in all possible worlds that are accessible to the agent from w. Likewise, an agent does not know a fact in world w, if there is some world accessible to the agent from w in which the fact is not true (Blackburn, Rijke, and Venema 2014).
The following definition will show when formulas are true in a particular world w in Kripke semantics.

Definition 2.3 - Truth in Kripke models
Let w be a world in a Kripke model $M = \langle W, R, V \rangle$, then

$w \vDash p$	iff $p \in V(w)$ for atomic variables $p \in At$
$w \vDash \phi \wedge \psi$	iff $w \vDash \phi$ and $w \vDash \psi$
$w \vDash \phi \vee \psi$	iff $w \vDash \phi$ or $w \vDash \psi$
$w \vDash \neg\phi$	iff $w \nvDash \phi$
$w \vDash \phi \rightarrow \psi$	iff $w \nvDash \phi$ or $w \vDash \psi$
$w \vDash \phi \leftrightarrow \psi$	iff $w \vDash \phi$ if and only if $w \vDash \psi$
$w \vDash K\phi$	iff for every $x \in W$, such that wRx, then $x \vDash \phi$
$w \nvDash K\phi$	iff there is some $x \in W$, such that wRx and $x \nvDash \phi$

In the next section, the problem of logical omniscience in Kripke models will be proved on the basis of validity. Validity is defined as follows:

Definition 2.4 - Validity in Kripke models
Let $M = \langle W, R, V \rangle$ be a Kripke model, let $w \in W$ and let ϕ be any formula. We define ϕ to be valid in a Kripke model M, denoted $M \vDash \phi$ as follows:

$M \vDash \phi$	iff $w \vDash \phi$ for all worlds $w \in W$

Ever since 1960, Kripke semantics have been the most commonly used method to make assumptions about truths in the real world. It has operated as the basis of the semantics that are used in modal logic (Rendsvig and Symons 2021). As mentioned before, however, Kripke semantics are committed to an unrealistic image of human reasoning. It assumes that agents know all logical truths and all logical consequences of their knowledge, while this is, in reality, beyond our reasoning ability. In the next section it will be formally proved that Kripke semantics allow logical omniscience.

2.1 Application to principles

Now that the basic principles of Kripke semantics are explained, it can be shown why these semantics allow logical omniscience.

First, it assumes that all agents have knowledge of all logical truths. This deduction is also known as the following invalidity:

(1) Omniscience Rule: If ϕ is valid, then so is $K\phi$

This implication is intuitively wrong because no agent can have knowledge about an infinite number of formulas and since there is an infinite number of logical truths, it follows that no agent can have knowledge of all logical truths.

The following proof shows that in Kripke semantics, **(1)** is nevertheless valid.

Theorem 2.1 - **(1)** is true for any logic defined from a collection of Kripke models.

Proof. Let C be any class of Kripke models.
Let $M \in C$ and $M = \langle W, R, V \rangle$
Let $C \vDash \phi$
Let $w \in W$
Then $x \vDash \phi$ for any x such that wRx
Then $w \vDash K\phi$
Therefore, $C \vDash K\phi$
So, if ϕ is valid in a class of Kripke models, then $K\phi$ is valid in this class of Kripke models.

It is shown in this proof that in any logic, defined from a collection of Kripke models, **(1)** is valid. This would imply that for every logical truth, an agent knows this truth. Intuitively, this is not true for ordinary humans.

Another objection to logical omniscience is that no agent can know all the deductive consequences of his own knowledge. In Kripke semantics, however, the following is a valid principle:

(2) Closure Under Known Implication: $K(\phi \to \psi) \to (K\phi \to K\psi)$

At first sight, this may look okay. Yet, an agent may know ϕ and $\phi \to \psi$, but he could still fail to put the two together (Hawke, Ozgun, and Berto 2019). For example, Don has two friends, Joe and Jeff. Don knows that Joe lives in Utrecht and that Jeff lives in Maastricht. Now, Don may know that Utrecht is North of Maastricht (ϕ) and she may know that if some city 1 is North of another city 2, then someone in city 1 has to travel South to go to city 2 ($\phi \to \psi$). Yet, Don may fail to understand the obvious implication, namely that Joe will have to travel South to visit Jeff.

In the following proof it will be formally shown that **(2)** is valid in Kripke models.

Theorem 2.2 - **(2)** is true for any logic defined from a collection of Kripke models.

Proof. Let C be any class of Kripke models.
Let $M \in C$ and $M = \langle W, R, V \rangle$
Let $w \in W$
If $w \vDash K(\phi \to \psi)$ and $w \vDash K\phi$
For any x such that wRx, then $x \vDash \phi \to \psi$ and $x \vDash \phi$
That means that $x \vDash \psi$
Therefore, $w \vDash K\psi$
Accordingly, for any world $w \in W$ we have shown that if $w \vDash K(\phi \to \psi)$, then $w \vDash K\phi \to K\psi$
Thus, $K(\phi \to \psi) \to (K\phi \to K\psi)$ is true in any class of Kripke models.

Theorem 2.2 proved that Kripke semantics validate **(2)**. Yet, as seen in the example of Don, Jeff and Joe, this is not a realistic assumption about human reasoning, since an agent can fail to make the obvious implication.

These are two examples of principles that create the problem of logical omniscience in Kripke semantics. Since one cannot have knowledge about all logical truths and one cannot indefinitely know the implications of his own knowledge, logical omniscience is impossible for ordinary humans.

As mentioned before, an attempt will be made to find a solution for the problem of logical omniscience. The goal is to create a model that models human reasoning and knowledge, at which logical omniscience will be removed, without completely removing logical competence. To achieve this goal, a dynamic model will be created in which some implications will be preserved, while others will be removed. Take a look at the following clauses:

(3) Conjunction Elimination: $K(\phi \wedge \psi) \to K\phi$

(4) Disjunction Introduction: $K\neg\phi \to K\neg(\phi \wedge \psi)$

In the target logic, **(3)** should be retained. If an agent has knowledge of two different formulas ϕ and ψ, then he has knowledge of these formulas separately from each other as well. This is an example of a principle that should be preserved in the target logic because of logical competence, which assumes that humans can make such simple trivial implications.

(4) on the other hand, will be removed from the target logic. It assumes that if an agent knows the negation of a formula, then he also knows the negation of the conjunction of that formula with another formula. Take a look at the following example. King William III knew that France would not go to war, but he did not know that France would not go to a nuclear war either, since nuclear weapons did not exist at that time. Therefore, it is true that no nuclear war would take place, but King William III had no knowledge of this (Hawke, Ozgun, and Berto 2019). For this reason, this clause will not be in the target logic that will solve the problem of logical omniscience, since it assumes that an

agent has certain knowledge, while this is not necessary.

In this chapter, it is discussed that Kripke frames are a good method for a lot of purposes. On the other hand, it also showed that it is not a good logic to simulate human reasoning and knowledge. It allows logical omniscience and agents cannot be logically omniscient.

In order to get to the target logic, other models will be discussed that solve the logical omniscience problem at some level. The next chapter will introduce a new type of model, which disposes of some logical omniscience, but does not solve the problem entirely.

Chapter 3

Minimal models

Minimal models, also called Montague-Scott semantics (Sedlár 2019) or neighborhood semantics (Pacuit 2017), are a more general semantics that is applied in modal logic. In these semantics, each world is associated to its so called 'neighborhoods'. Neighborhoods are sets of subsets of worlds, so that every world in the model has his own set of related propositions (Cerro, Herzig, and Mengin 2012). Specifically, the neighborhood of world w is a set of worlds that are associated with world w. The structure of minimal models is defined as follows:

Definition 3.1 - Minimal models
A minimal model is a tuple $M = \langle W, N, P \rangle$, where
a. W is a set of possible worlds,
b. N is a mapping from W to sets of subsets of W, and
c. P is a mapping from atomic formulas to subsets of W.

As mentioned before, the idea of minimal models is that each world w in W in a minimal model $M = \langle W, N, P \rangle$, is related to a set N_w of propositions. The mapping N is also called the neighborhood function, since it assigns a neighborhood N_w to every world w in W. A proposition in minimal models will be described as a set of possible worlds, which is a subset of W. This makes N_w a collection of subsets of W, and therefore a collection of propositions. Accordingly, a world w is associated with a set N_w, which represents a set of propositions that are true at w. The function P gives a truth value to every atomic formula in a world (Chellas 1980).

In order to apply minimal models to principles that allow logical omniscience, we need to understand the interpretation of knowledge in these models. For that, it is useful to know that a formula expressed by ϕ, also called the truth set of ϕ, is denoted $||\phi||$.

Definition 3.2 - Truth in minimal models
Let w be a world in a minimal model $M = \langle W, N, P \rangle$, then

$w \vDash p$	iff $w \vDash P(p)$ for atomic variables
$w \vDash \phi \wedge \psi$	iff $w \vDash \|\|\phi\|\|$ and $w \vDash \|\|\psi\|\|$
$w \vDash \phi \vee \psi$	iff $w \vDash \|\|\phi\|\|$ or $w \vDash \|\|\psi\|\|$
$w \vDash \neg\phi$	iff $w \nvDash \|\|\phi\|\|$
$w \vDash \phi \rightarrow \psi$	iff $w \nvDash \|\|\phi\|\|$ or $w \vDash \|\|\psi\|\|$
$w \vDash \phi \leftrightarrow \psi$	iff $w \vDash \|\|\phi\|\|$ if and only if $w \vDash \|\|\psi\|\|$
$w \vDash K\phi$	iff $\|\|\phi\|\| \in N_w$
$w \nvDash K\phi$	iff $\|\|\phi\|\| \notin N_w$

In the last two lines, the definition states that an agent knows formula ϕ at world w in M if and only if the proposition expressed by ϕ is in the neighborhood of w. An agent at world w does not know a formula ϕ, however, when the proposition expressed by ϕ is not in the neighborhood of w.

Minimal models are a generalisation of Kripke semantics and solve some degree of logical omniscience. Yet, it still models some level of logical omniscience that should be avoided in the target logic. The next section provides proofs that show how minimal models make some principles that allow logical omniscience invalid. Subsequently, a proof will follow that shows how minimal models nonetheless validate a principle that causes logical omniscience.

3.1 Application to principles

Since it is clear how minimal models work, and how they model knowledge, **(1)** (Omniscience Rule) and **(2)** (Closure Under Known Implication) can be reviewed.

(1) stated that all agents have knowledge of all logical truths. More formally, if ϕ is valid, then so is $K\phi$. As explained in the preceding chapter, this deduction is naturally wrong. Because there are infinitely many logical truths and because humans cannot know an infinite number of formulas, **(1)** is irrational. The following proof shows that minimal models do not validate this implication.

Theorem 3.1 - **(1)** is not valid in minimal models.

Proof. Let $M = \langle W, N, P\rangle$ be a minimal model
Let $M \vDash \phi$, then $||\phi|| = W$
So, for any world $w \in W$, $w \vDash \phi$
Let $W \notin N_w$
Then $||\phi|| \notin N_w$
Therefore, $w \nvDash K\phi$
This means that $M \vDash \phi$, but $M \nvDash K\phi$
So, **(1)** is not valid in minimal models.

While Kripke semantics do make **(1)** valid, minimal models do not. Accordingly, Kripke semantics assume that agents know all logical truths, while minimal models do not accept this principle.

How about the following principle? **(2)** is validated by Kripke semantics, causing logical omniscience. This implication is formally expressed: $K(\phi \rightarrow \psi) \rightarrow (K\phi \rightarrow K\psi)$, meaning that if an agent knows the premises of a known implication, he also knows the inferred formula. Minimal models, in contrast to Kripke semantics, invalidate this principle as well.

Theorem 3.2 - **(2)** is not valid in minimal models.

Proof. Let $M = \langle W, N, P \rangle$ be a minimal model in which $W = \{x, y, z\}$
Let $P(p) = \{x\}$ and $P(q) = \{x, y\}$
Then $||p \rightarrow q|| = \{x, y, z\}$, $||p|| = \{x\}$ and $||q|| = \{x, y\}$
Let $\{x, y, z\} \in N_x$, $\{x\} \in N_x$ and $\{x, y\} \notin N_x$
Then, $||p \rightarrow q|| \in N_x$, $||p|| \in N_x$ and $||q|| \notin N_x$
Therefore, $x \vDash K(p \rightarrow q)$, but $x \nvDash K(p) \rightarrow K(q)$
Since, $K(p \rightarrow q) \rightarrow (K(p) \rightarrow K(q))$ is not true in this minimal model, we now have a counter model.
So, **(2)** is not valid in minimal models.

This proof shows that minimal models do not assume that agents know all logical consequences of the formulas they know. Since minimal models invalidate **(1)** as well as **(2)**, they seem to solve a significant part of logical omniscience.

It is one step in the right direction, but we are not there yet. There is, namely, another principle related to logical omniscience. It implies that if a double implication is valid, then an agent knows one side of the implication if and only if he knows the other side of the implication. Thus assuming that if agents know one side of a double implication, they will have knowledge of the other side. Formally,

(5) Equivalence Rule: If $\phi \leftrightarrow \psi$ is valid, then $K\phi \leftrightarrow K\psi$ is valid.

The following is an example of why **(5)** is unrealistic. Suppose that it is dark outside if and only if it is night. It could possibly be the case that an agent knows that it is dark outside, without realizing that it should therefore be night. In this instance, $\phi \leftrightarrow \psi$ is true, but $K\phi \leftrightarrow K\psi$ is not true. For this reason, **(5)** should not be maintained in the target logic. The following proof shows why minimal models are not an appropriate candidate for solving logical omniscience.

Theorem 3.3 - **(5)** is true for any logic defined from a collection of minimal models.

Proof. Let C be a class of minimal models

Let $M \in C$ and $M = \langle W, N, P \rangle$
Let $C \vDash \phi \leftrightarrow \psi$, so that the propositions expressed by ϕ and ψ are the same, i.e. $||\phi|| = ||\psi||$ for every minimal model in C
Then, for any world $w \in W$, $||\phi|| \in N_w$ iff $||\psi|| \in N_w$
Then $w \vDash K\phi$ iff $w \vDash K\psi$
Which means that $w \vDash K\phi \leftrightarrow K\psi$ for any world $w \in W$
Therefore, $C \vDash K\phi \leftrightarrow K\psi$
So, if $\phi \leftrightarrow \psi$ is valid, then $K\phi \leftrightarrow K\psi$ is valid for any class of minimal models.

This chapter showed that **(1)** and **(2)** are not valid in minimal models. This means that minimal models solve a part of the logical omniscience problem that could not be tackled by Kripke semantics. Nevertheless, **(5)** is valid in minimal models, suggesting that this semantics is not totally solving logical omniscience, leaving us halfway our quest for the target logic.

Chapter 4

Hyperintensional models

So what needs to be changed if we want to model logically non-omniscient agents? Kripke semantics were too specific, friendly allowing logical omniscience. Minimal models seemed to be heading in the right direction, but failed to circumvent total omniscience, giving in when the Equivalence Rule **(5)** came into play.

(5) is a commonly known problem for modal logic when it comes to modeling rational attitudes (Rendsvig and Symons 2021). As mentioned in previous chapter, the following situation is possible: two formulas p and q can be logically equivalent, while an agent could have knowledge of p without knowing q. In minimal models, a proposition is described as a set of possible worlds. So, if p and q are logically equivalent, they would also be identical. Therefore, if knowledge is a legitimate function, the above mentioned situation could not arise in minimal models. In other words, if knowledge is defined as a set of sets of possible worlds, then $p \leftrightarrow q$, but not $K(p) \leftrightarrow K(q)$ would be impossible. Cresswell called this the paradox of hyperintensional context (Cresswell 1975).

There are a lot of possible replies to this hyperintensionality concept [1]. Different responses are reflected in several hyperintensional models that attempt to model human reasoning and knowledge. Before coming to the presentation of specific models, the basic principle of hyperintensional models will be explained. In order to understand how hyperintensional models work, the pre-model will be introduced. This is a generalization of minimal models.

Definition 4.1 - Pre-model
A pre-model is a tuple $\rho = \langle W, C, O, N_c, I \rangle$, where
a. W is a non-empty set of possible worlds,
b. C is the non-empty set of semantic contents of sentences,
c. O is a function from formulas to C,
d. N_c is a function that assigns to every $w \in W$ a subset of C, and
e. I is a function that assigns to every $c \in C$ a proposition $I(c) \subseteq W$.

[1] More information in appendix A

Let's take a closer and more detailed look at this definition. It states that C is a set of semantic contents of sentences, which is represented by formulas. O is the content function, that assigns contents to these sentences. N_c assigns a set of contents $c \in C$ to every world $w \in W$. Accordingly, N_c describes a property of contents. Lastly, there is the intension function I that assigns to every content a proposition. This function shows that propositions are intuitively established by contents. There is a function $[\![\cdot]\!]$ which can be seen as the composition of the content function O and the intension function I, assigning propositions to formulas (Sedlár 2019). Figure 4.1 provides a demonstration of this idea.

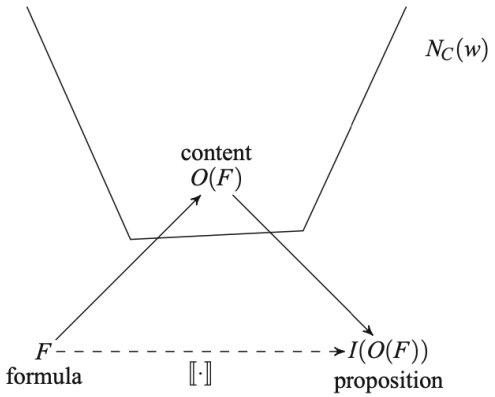

Figure 4.1: Pre-model (Sedlár 2019)

A hyperintensional model, is a pre-model ρ in which function $[\![\cdot]\!]$ is defined as seen in figure 4.1: $[\![\phi]\!] = I(O(\phi))$. In a hyperintensional model, a formula ϕ is true when $[\![\phi]\!]$. Knowledge is defined as follows: $[\![K\phi]\!] = \{w \mid O(\phi) \in N_c(w)\}$ (Sedlár 2019).

Now let's get back to Cresswell's paradox of hyperintensional context. It states that it is not necessary for the Equivalence Rule to be valid. Hyperintensional models agree. Suppose that $\phi \leftrightarrow \psi$ is true in a hyperintensional model. Then, $[\![\phi]\!] \leftrightarrow [\![\psi]\!]$ is true (i.e. $I(O(\phi)) \leftrightarrow I(O(\psi))$). Now suppose that $K\phi$, so $O(\phi) \in N_c$. Then it is possible for $O(\psi)$ not to be in N_c and therefore it is possible that $\neg K\psi$. This invalidates the Equivalence Rule, because $\phi \leftrightarrow \psi \rightarrow K\phi \leftrightarrow K\psi$ is not true in this hyperintensional model.

In the remainder of this chapter the focus will be on two hyperintensional models, introduced by Fagin et al. (1995).

4.1 Awareness logic

First, awareness logic will be considered. The idea of this semantics is that awareness about a formula is necessary to gain knowledge about this formula (Fagin et al. 1995). This will probably sound obvious. Intuitively, it is impossible for someone to have knowledge about ϕ without being aware of ϕ. After all, an agent is unaware of some formula ϕ if he doesn't know that he doesn't know ϕ. Consequently, this section will introduce a new modal operator A, such that $A\phi$ means that the agent is aware of formula ϕ.

To model knowledge in awareness logic, there is another modal operator defined. $X\phi$ means that the agent has *explicit knowledge* of ϕ. Explicit knowledge of a formula is gained by an agent if he is aware of the formula ($A\phi$) and if he has *implicit knowledge* of the formula ($K\phi$) (Schipper 2014). Formally, $X\phi \leftrightarrow A\phi \wedge K\phi$. Implicit knowledge can be considered the same way as it was in previous chapters: an agent knows a formula in world w if the formula is true in all worlds the agent considers doxastically possible from world w.

The formal definition of awareness logic is as follows:

Definition 4.2 - Awareness logic
An awareness logic is a tuple $M^a = \langle W, R, V, A(w)\rangle$, where
a. The tuple $\langle W, R, V\rangle$ is a Kripke structure, and
b. $A(w)$ is a function mapping sets of formulas to every world w in W.

Naturally, the awareness function A defines sets of formulas an agent is aware of at world w. Important to note is that $A\phi$ means that the agent is *aware* of ϕ, not necessarily that he *knows* ϕ. The formulas in the set defined by the awareness function can be arbitrary. Thus, both ϕ and $\neg\phi$ can be in $A(w)$, as well as one of the two or none of them. Furthermore, an agent can be aware of $\phi \vee \psi$, without being aware of $\psi \vee \phi$ (Fagin et al. 1995).

The semantics for awareness logic when it comes to implicit knowledge is the same as knowledge in Kripke semantics. When we get to the new modal operators $A\phi$ and $X\phi$, however, new clauses are needed:

$w \vDash A\phi$ iff $\phi \in A(w)$
$w \vDash X\phi$ iff $w \vDash A\phi$ and $w \vDash K\phi$

The first item states that an agent is aware of a formula ϕ in world w if and only if ϕ is in $A(w)$. The second clause explains that an agent has explicit knowledge of ϕ in world w if and only if (1) he is aware of ϕ in w and (2) he implicitly knows ϕ at world w.

Accordingly, agents do not need to have explicit knowledge of all logical truths or tautologies. For example, an agent may not explicitly know $p \vee \neg p$ because the agent could not be aware of the formula $p \vee \neg p$. In addition, an agent could have explicit knowledge of p and of $p \rightarrow q$, but he could still not explicitly

come to know q, since he could not be aware of q. Formally, $X\phi \wedge X(\phi \rightarrow \psi) \wedge \neg X\psi$ is possible in awareness logic.

Now that (explicit) knowledge in awareness logic is explained, the logic can be applied to the logical omniscience problem and its additional principles.

4.1.1 Application to principles

In this section it will be proved that awareness logic solves the problem of logical omniscience for the concept of explicit knowledge. This will be done by proving that awareness logic does not allow **(1), (2), (4)** and **(5)**. In the following paragraphs, if the word knowledge, or a variation thereof, is used it implies explicit knowledge. If it concerns implicit knowledge, this will be clearly stated.

Let's start with **(1)**, the Omniscience Rule. It states that if a formula is true in the real world, then an agent should have knowledge of this formula. A statement that obviously assumes the agent to be logically omniscient. The awareness logic correctly invalidates this rule.

Theorem 4.1 - **(1)** is not valid in awareness logic.

Proof. Let ϕ be a true formula in the real world
$X\phi$ is not necessary, because it could be possible that $\neg A\phi$
So, **(1)** is not valid in awareness logic.

The simple proof above showed that an agent does not necessarily need to know every logically true formula, because he could not be aware of the formula.

The following proof looks similar. In awareness logic, **(2)**, Closure Under Known Implication, is not valid. The principle states that if an agent knows some implication $\phi \rightarrow \psi$ then, if the agent knows ϕ he also knows ψ.

Theorem 4.2 - **(2)** is not valid in awareness logic.

Proof. Suppose that $X(\phi \rightarrow \psi)$ and $X\phi$
It is not necessary that $X\psi$, because it could be possible that $\neg A\psi$
So, **(2)** is not valid in awareness logic.

In the previous section the above proof was already stated. An agent can have knowledge of p and of $p \rightarrow q$ in world w, but he could still fail to know q in that world, since q could not be in the set of formulas $A(w)$.

Another principle that allows logical omniscience is **(4)**, Disjunction Introduction. It assumes that if an agent knows that some formula is not true, then he also knows that the conjunction of that formula and another formula is not true. The awareness logic does not allow this principle.

Theorem 4.3 - **(4)** is not valid in awareness logic.

Proof. Suppose that $X\neg\phi$
Then not necessarily $X\neg(\phi \wedge \psi)$, since it could be possible that $\neg A\psi$
So, **(4)** is not valid in awareness logic.

Let's bring back the example of chapter 2. It described a situation in which King William III knows that France is not going to war ($X\neg\phi$), but he doesn't know that it is not true that France is going to war *and* that France is going to a nuclear war ($X\neg(\phi \wedge \psi)$). This is because the King is not aware of the existence of nuclear weapons ($\neg A\psi$). So principle **(4)** is correctly invalidated by awareness logic.

The last invalid principle that will be discussed in this section is the Equivalence Rule, **(5)**. It is the rule that prevented the minimal models from totally solving logical omniscience. Let's see whether awareness logic allows the Equivalence Rule.

Theorem 4.4 - **(5)** is not valid in awareness logic.

Proof. Suppose that $\phi \leftrightarrow \psi$ is true
Also, suppose that $X\phi$
Then, it is not necessarily true that $X\psi$, since $\neg A\psi$ could be true.
So, **(5)** is not valid in awareness logic.

Thus, awareness logic comes closer to solving logical omniscience than minimal models, since it doesn't allow **(5)**.

In chapter 2 the Conjunction Elimination rule **(3)** was introduced. It is the only principle that should be retained in the target logic. It states that if an agent knows two formulas p and q, then he also knows the two formulas separately. The awareness logic correctly invalidated previous principles, but it falsely invalidates **(3)** too.

Theorem 4.5 - **(3)** is not valid in awareness logic.

Proof. Suppose that $X(\phi \wedge \psi)$
Then, it is not necessarily true that $X\phi$, since $\neg A\phi$ could be true
So, **(3)** is not valid in awareness logic.

At first sight, awareness logic seems to be doing really good in solving logical omniscience. It invalidates all rules that allow logical omniscience, therefore solving the problem. However, awareness logic also invalidates the Conjunction Elimination, which should be retained in the target logic. Accordingly, awareness logic can be seen as a semantics that solves logical omniscience, but at the same time allows total logical incompetence. It assumes that agents cannot perceive any logical implication of their knowledge.

The goal is to create a semantics that can model human reasoning, avoiding logical omniscience, while preserving logical competence. Now, we have seen two extremes: semantics allowing too much logical omniscience and a semantics allowing too much logical incompetence. The target logic will be something in between, such that human reasoning and knowledge can be approached as accurately as possible. Before getting there, another hyperintensional model will be discussed.

4.2 Impossible worlds semantics

The problem of logical omniscience arises when knowledge is defined as truth in all possible worlds. In chapter 2 we have seen that Kripke semantics allow logical omniscience because of it's interpretation of knowledge. In impossible worlds semantics, the notion of knowledge is different. Particularly, the possible worlds that we have seen before are expanded by adding impossible worlds (Fagin et al. 1995). In these worlds, the ordinary logical rules do not persist. It could be true in these worlds, for example, that ϕ is true and $\phi \rightarrow \psi$ is true, but ψ is not true.

These appended worlds are logically impossible. They make principles true, that are inconceivable in modern logic. Yet, ordinary humans are not perfect reasoners, decently following all the logically implemented rules. Therefore, human agents can consider impossible worlds to be possible, causing them to make invalid assumptions (Hintikka 1975). Bjerring and Skipper (2019) explain that in impossible worlds semantics, for agents with limited reasoning capabilities, there are more doxastic possibilities (formulas the agent considers possible) than logical possibilities (formulas that are actually true). This is because agents can consider a significant space of impossible worlds, containing a lot of doxastic possibilities.

Formally, impossible worlds semantics are structured as follows:

Definition 4.3 - Impossible worlds semantics
An impossible worlds semantics is a tuple $M^i = \langle W, R, P, \sigma \rangle$, where
a. The tuple $\langle W, R \rangle$ is a Kripke frame, and
b. $P \subseteq W$ is the set of *possible* worlds, and
c. σ is a function that assigns truth values to all formulas in all worlds.

Remember that R is the accessibility relation on worlds $w \in W$, such that if wRv, it means that w can access v. The syntactic assignment σ works as usual on formulas in possible worlds. This means that if $w \in P$, then:

$\sigma(w)(\phi \wedge \psi)$ iff $\sigma(w)(\phi)$ and $\sigma(w)(\psi)$
$\sigma(w)(\neg\phi)$ iff $\sigma(w)(\phi)$ is not true
$\sigma(w)(K\phi)$ iff $\sigma(v)(\phi)$ for all v, such that wRv

For impossible worlds, on the other hand, σ works arbitrarily. The only, obvious, rule is that for all w (whether $w \in P$ or $w \notin P$) it applies that $w \vDash \phi$ iff $\sigma(w)(\phi)$ (Fagin et al. 1995).

Thus, the notion of knowledge in the impossible worlds model is expressed as follows:

Definition 4.4 - Knowledge in impossible worlds semantics
An agent knows a formula ϕ if and only if ϕ is true in all worlds that are doxastically possible for the agent.

Note that the worlds the agent considers possible can be either possible or impossible.

Impossible worlds semantics are useful to explain why humans consider false possibilities to be true. We are modelled to inconsistently reason about knowledge (Berto and Jago 2018). We could consider, for example, an impossible world in which $\phi \wedge \neg\phi$ is true to be doxastically possible, while this is obviously logically impossible. The impossible worlds give an explanation of how humans reason and the next section will show that it avoids logical omniscience as well.

4.2.1 Application to principles

Impossible worlds semantics are a good option for solving logical omniscience. By reason of the usage of impossible worlds, the flaws of human reasoning can be represented. Again, it will be shown that impossible worlds solve logical omniscience by proving that it invalidates principles **(1), (2), (4)** and **(5)**.

To show: $\phi \rightarrow K\phi$ is not valid

Theorem 4.6 - **(1)** is not valid in impossible worlds semantics.

Proof. Let ϕ be a true formula in the real world
$w \vDash K\phi$ is not necessary because there could be some world v, such that $\sigma(v)(\phi)$ is false and wRv
So, **(1)** is not valid in impossible worlds semantics.

An agent only gains knowledge of a formula if this formula is true in *all* worlds the agent considers doxastically accessible. Therefore, the Omniscience Rule is not valid in impossible worlds, since the agent can consider an impossible world to be true in which ϕ is not true.

To show: $K(\phi \rightarrow \psi) \rightarrow (K\phi \rightarrow K\psi)$ is not valid

Theorem 4.7 - **(2)** is not valid in impossible worlds semantics.

Proof. Suppose that $w \vDash K(\phi \to \psi)$ and $w \vDash K\phi$
It is not necessary that $w \vDash K\psi$, since there could be some world v, such that $\sigma(v)(\psi)$ is false and wRv
So, **(2)** is not valid in impossible worlds semantics.

The above theorem shows that an agent can know p and $p \to q$, but may still be unknown of q, since q could be false in an impossible world the agents considers doxastically possible.

To show: $K\neg\phi \to K\neg(\phi \wedge \psi)$ is not valid

Theorem 4.8 - **(4)** is not valid in impossible worlds semantics.

Proof. Suppose that $w \vDash K\neg\phi$
Then it is not necessarily true that $w \vDash K\neg(\phi \wedge \psi)$, since it could be possible that there is some world v, such that $\sigma(v)(\phi \wedge \psi)$ is true and wRv
So, **(4)** is not valid in impossible worlds semantics.

The Disjunction Introduction is proved to be invalid, since the conjunction of the false formula and another formula could be true in an impossible world the agent considers doxastically possible.

To show: $\phi \leftrightarrow \psi \to K\phi \leftrightarrow K\psi$ is not valid

Theorem 4.9 - **(5)** is not valid in impossible worlds semantics.

Proof. Suppose that $\phi \leftrightarrow \psi$ is true
Also, suppose that $w \vDash K\phi$
Then, it is not necessary that $w \vDash K\psi$, since it could be possible that there is some world v, such that $\sigma(v)(\psi)$ is false and wRv
So, **(5)** is not valid in impossible worlds semantics.

The above Equivalence Rule is a principle we have seen a lot in preceding sections, because of the different outcomes of the considered semantics. Minimal models failed to invalidate this rule, but the awareness logic and impossible worlds model prove to be competent in resolving it.

For the Conjunction Elimination, in contrast to the previous rules, we want the target logic to allow it. This is something the awareness logic failed to do, therefore tolerating too much logical incompetence. Let's see what impossible

worlds do with **(3)**.

To show: $K(\phi \wedge \psi) \rightarrow K\phi$ is not valid

Theorem 4.10 - **(3)** is not valid in impossible worlds semantics.

Proof. Suppose that $w \vDash K(\phi \wedge \psi)$
Then it is not necessary for $w \vDash K\phi$ to be true, since it could be possible that there is some world v, such that $\sigma(v)(\phi)$ is false and wRv
So, **(3)** is not valid in impossible worlds semantics.

Now, we run into the same problem as with awareness logic. Both models allow too much logical incompetence, both not validating **(3)**. They are evidently a solution to logical omniscience, but they bring their own problems concerning human competence. We are less competent than Kripke semantics and minimal models claim us to be, but we are definitely capable of more than awareness logic and impossible worlds semantics assert.

For the target logic, there should be a golden mean between too much logical omniscience and too little logical competence. The hyperintensional models are a useful basis, but there is something missing that enables these semantics to model human reasoning properly. In the target logic there will be an added feature, giving the opportunity to simulate how people reason more accurately. This will allow the logic to model agents who are not logically omniscient, while they will still able to generate a certain number of logical implications, making them logically competent.

The next chapter will provide some necessary information on logical competence, which will be used in the target logic.

Chapter 5

Logical competence

Logical competence states that agents know the trivial consequences of their knowledge (Bjerring and Skipper 2019). Formally,

Definition 5.1 - Logical competence
An agent is logical competent if he does not miss out on any trivial consequences of his knowledge.

If we reject logical omniscience completely, this would mean that a person could never know q, while knowing p and $p \rightarrow q$. Obviously, there are occasions in which people really know things by making logical inferences (Cherniak 1981). In order to create a clear vision of this idea, let's consider the following instance. Suppose that an agent knows p and let q be a trivial logical consequence of p. In some occasions, when this agent is asked whether q is true, he will answer that q is indeed true. This means he can make some logical inferences and therefore is logically competent. For example, an agent knows that if someone drives too fast on the highway, that person will get a fine. He also knows that Emily is driving too fast on the highway. Now, if someone asks him whether Emily will get a fine, and he answers 'yes', he knows the trivial consequence of his knowledge and therefore is logically competent. The reason why logical competence is so important in this matter, is because of the relation between logical omniscience and logical competence. Again, if the aim is to model real-world agents, such as humans, it is necessary to understand how they reason. Humans are not omniscient, but they are very intelligent agents, with the ability to perform simple trivial reasoning tasks (Bjerring and Skipper 2019).

There are two important competencies that human agents can perform. First, there is the Conjunction Elimination, which we encountered before. It states that if a person knows a conjunction of two formulas, then he also knows the formulas individually. Formally, $K(\phi \wedge \psi) \rightarrow K\phi$. Intuitively, this principle is true for logically competent agents such as humans. An agent knows that Sarah is a doctor and a singer if and only if the agent knows that Sarah is a

doctor. Naturally, knowing the latter (that Sarah is a doctor) is a part of the former (that Sarah is a doctor and a singer) (Hawke, Ozgun, and Berto 2019). Suppose that someone knows that Sarah is a doctor and a singer, but does not know that Sarah is a doctor. It is very clear to see that this is an improbable situation and that no logically competent agent could make this statement. Thus, since the left part of the inference automatically implies the right part, **(3)** is a trivial principle.

Second, logically competent agents know simple tautologies. A tautology is a compound proposition which is always true (Chellas 1980). The following principle is an example of a tautology that should be retained in the target logic.

(6) DeMorgan's Theorem: $K\neg(\phi \wedge \psi) \rightarrow K(\neg\phi \vee \neg\psi)$

(6) states that if an agent knows that a conjunction is not true, then he also knows that either one of the formulas in the conjunction, or both, are not true (assuming that we use an inclusive-or, which is always the case in modern logic). One knows that Julia is not a mother *and* a cook if and only if one knows that Julia is not a mother or that Julia is not a cook. Obviously, this is a valid principle in the logic that models human reasoning. Suppose, in an implausible situation, that someone knows that Julia is not a mother *and* a cook, but also knows that Julia is a mother and that Julia is a cook. This obviously is an unrealistic assumption that only logically incompetent agents could make. Again, knowing the latter part of the implication is a part of knowing the former. Therefore, if an agent has knowledge of $K\neg(\phi \wedge \psi)$ then he consistently knows $K(\neg\phi \vee \neg\psi)$ too.

In both **(3)** and **(6)** the implications follow naturally from the premises. Because the agent has knowledge of the first part, he automatically has knowledge of the latter. In for example **(4)**, however, this is not the case. The Disjunction Introduction states that if someone knows that a formula is not true, then he also knows that the conjunction of that formula and another formula is not true. In contrast to before mentioned rules, **(4)** is not intuitively correct. It could for instance be possible that one knows that Max is not a fireman, but that he does not know that Max is not a fireman *and* a surf teacher for dogs, since he may not even know that the latter is an existing job. In this case, the latter part of the implication is not a part of the premise. Therefore, there is a non-trivial inference needed that can provide knowledge of the added formula to the agent.

As a result, the target logic should validate the principles **(3)** and **(6)** in order to model a logically competent agent. Principles **(1), (2), (4)** and **(5)**, on the other hand, are not intuitively true and should therefore be invalidated by the target logic.

We have seen that the awareness logic and impossible worlds model solve logical omniscience, but they fail in allowing agents to be logically competent. In chapter 4 we have already seen how they invalidate **(3)**, now it will be proved that they invalidate **(6)** as well.

To show: $K\neg(\phi \wedge \psi) \rightarrow K(\neg\phi \vee \neg\psi)$ is not valid.

Theorem 5.1 - **(6)** is not valid in awareness logic.

Proof. Suppose that $K\neg(\phi \wedge \psi)$
Then, it is not necessary that $X\neg\phi$ or $X\neg\psi$, since $\neg A\neg\phi$ and $\neg A\neg\psi$ could be true
Thus, $X(\neg\psi \vee \neg\psi)$ is not necessarily true
So, **(6)** is not valid in awareness logic.

Theorem 5.2 - **(6)** is not valid in impossible worlds semantics.

Proof. Suppose that $w \vDash K\neg(\phi \wedge \psi)$
Then it is not necessary for $w \vDash K(\neg\phi \vee \neg\psi)$ to be true, since it could be possible that there is some world v, such that $\sigma(v)(\neg\phi \vee \neg\psi)$ is false and wRv
So, **(6)** is not valid in impossible worlds semantics.

Above proofs show that both models fail to allow the principles that model logically competent agents. Still, people are competent reasoners and the goal is to be able to model people with different degrees of logical competence. Therefore, we need to create an epistemic logic that is not too extreme in one of the two ways. It should not assume that all agents are omniscient, nor it should say that all agents are incompetent.

In the following chapter, an attempt will be made to apply a target logic in which this ability of logical competence is preserved and logical omniscience is removed in order to simulate human reasoning and knowledge as precise as possible.

Chapter 6

Target logic

Awareness logic and impossible worlds semantics are two popular solutions to logical omniscience (Cresswell 1975), (Hintikka 1962), (Fagin et al. 1995). In the previous chapter we have seen, however, that these models assume agents to be logically incompetent. In this chapter the target logic will be presented, which models logically non-omniscient, yet competent agents. In order to reach this goal, a dynamized impossible worlds model will be applied, inspired by Bjerring and Skipper (2019). This enables us to capture what agents know, as well as what agents can infer from their knowledge. First, the formal details of the model will be defined. The subsections show how the target logic avoids logical omniscience and how it allows logical competence.

In previous chapter logical competence was defined as not missing out on any trivial consequences of your knowledge. We assume that ordinary human agents are capable of some degree of logical reasoning (Cherniak 1981) and this ability is exactly what should be captured in the target logic. This way, the logic will model agents who do not overlook any trivial consequences of their knowledge and who therefore are logically competent.

Intuitively, the definition of a trivial consequence is different per person, depending on ones cognitive assets. For a professional logician, for example, $\neg\phi \rightarrow \neg\psi$ entails $\neg(\phi \wedge \neg\psi)$ probably is a trivial inference, whereas for a first year Artificial Intelligence student, this may be non-trivial (Bjerring and Skipper 2019).

In order to capture these different levels of cognitive skills, a step-based method will be introduced, inspired by Drapkin's work in step-logic (1999). It suggests that agents apply rules from a set R of trivial inference rules to reason about knowledge. This logic succeeds in modelling someone's cognitive resources by means of a number n of steps of reasoning, at which one step corresponds to applying one inference rule from R. The number of steps one can take, depends on its cognitive resources. As a result, this method enables us to model agents with different cognitive means. When n is equal to 0, one has no cognitive resources and no logical consequences are trivial to this agent. On the other hand, when n reaches infinity, the agent has unlimited cognitive

resources and all logical consequences will be trivial to him. In between these extremities, there is a wide spectrum of numbers n of steps, representing agents with different reasoning skills (Bjerring and Skipper 2019). Thus, a proposition is true for an agent, if he can infer it from his knowledge in a certain number n of steps. Formally, a logical consequence is trivial in the following situation.

Definition 6.1 - Trivial consequence
A proposition p is a trivial consequence of a set P of propositions if and only if p can be inferred from P within n applications of inference rules from R.

Being logically competent then, is defined as being able to infer the trivial consequences of your knowledge, as stated in **Definition 5.1**. After applying these n inference rules, the agent will not miss out on any trivial consequences of his knowledge.

The number n determines ones degree of cognitive resources, but these resources also depend on the rules in R. For instance, if R contains solely Modus Ponens, then $p \rightarrow q$ and p would only trivially imply q, independent from n. In contrast, if R contains all rules in classical propositional logic, the trivial consequences of $p \rightarrow q$ and p could be way more, depending on the number n of steps. Accordingly, someone should use a well-filled proof system R and a high value of n if he wants to model a complex agent who has great reasoning skills and a powerful reasoning system. However, if one wants to model a simple agent, with low reasoning skills and a weak reasoning system, then R should be a poor proof system and n should have a low value (Bjerring and Skipper 2019).

With this prescience in mind, the goal of the rest of the chapter is to demonstrate the target logic, which avoids logical omniscience, but allows logical competence.

6.1 Semantics of the target logic

As mentioned before, the target logic will be based on Fagin's impossible worlds model (1995). The impossible worlds in this model possibly contain false formulas, but the worlds can still be considered doxastically possible by agents. Therefore, someones knowledge often contains impossible propositions and thus this model creates logical incompetence. Recall when someone gained knowledge in this logic.

Definition 4.4 - Knowledge in impossible worlds
An agent knows a formula ϕ if and only if ϕ is true in all worlds that are doxastically accessible for the agent.

In order to create a logic that models logically competent agents, the definition of knowledge should be adjusted. Therefore, the relation between an

agent's doxastic states should be considered dynamically. Accordingly, the corresponding reasoning process allows a transition from some doxastic state, which contains premises of an inference, to a revised doxastic state containing the conclusion of the inference as well (Ditmarsh, Hoek, and Kooi 2008). For the target logic, we will apply this dynamic relation to Fagin's impossible worlds. This way, the model will not only be able to capture an agent's knowledge, but also what the agent can infer from that knowledge.

We have seen that a formula q is a trivial consequence of p if and only if q can be inferred from p in at most n steps of reasoning, applying the trivial rules in R. Formally, if P and P' are sets of propositions and P' is a trivial consequence of P, then $P \vdash^n P'$, meaning that P' can be inferred from P in at most n steps. The relation $\vdash^n$ is monotonic.

Definition 6.2 - n-monotonicity
If $P \subseteq P'$ and $P \vdash^n p$, then $P' \vdash^n p$.

Monotonicity assures that logical inferences cannot be affected by adding new assumptions.

As we have seen, the target logic involves a new modal operator n, which is a whole number from zero to countably infinite. The next definition will explain how n operates in the model.

Definition 6.3 - Modal operators in the target logic
Kp: agent knows p.
$\langle n \rangle p$: p is true after n steps of logical reasoning.
$\langle n \rangle Kp$: agent knows p after n steps of logical reasoning.

Now, the target model can be defined.

Definition 6.4 - Target model
A target model is a tuple $M = \langle W^P, W^I, f, V \rangle$, where
a. W^P is a non-empty set of possible worlds,
b. W^I is a non-empty set of impossible worlds,
c. $W = W^P \cup W^I$,
d. f is a function that maps to each world in W a set of worlds from W,
e. V is a function that maps a set of formulas to each world in W.

The accessibility function f assigns to each world a set of doxastically accessible worlds in W, which can be either possible (W^P) or impossible (W^I).

Now, we can proceed in introducing the formalities of our semantics. **Definition 4.4** remains the definition of knowledge in our model, but still there is a semantics needed for our new modal operator $\langle n \rangle$. The following semantics will show that $\langle n \rangle Kp$ is true in w if and only if p can be inferred from every doxastically accessible world from w, within n applications of trivial rules in R.

Consequently, a formula p is true more often in $\langle n\rangle Kp$, than in Kp, since in Kp, p has to be true in every doxastically accessible world, while in $\langle n\rangle Kp$, p should only be inferred within n steps from the formulas that are true in the doxastically accessible worlds. This gives us a little push towards solving the problem of logical incompetence.

We have not defined yet, what exactly follows within n steps of logical reasoning from true formulas in a specific world. The following definition will help formalizing this concept.

Definition 6.5 - n-radius
The n-radius w^n of a world $w \in W$ is defined as follows:
For $w \in W^P$
$w^n = \{w\}$
For $w \in W^I$
$w^n = \{w' \in W^I : V(w) \subseteq V(w') \text{ and } V(w) \vdash^n V(w')\}$
Each world in w^n is called an *n-extension* of w.

In other words, the n-radius of w is identical to the set of n-extensions of w. What these n-extensions are, however, depends on whether w is a possible or an impossible world. For $w \in W^P$, w simply is his own n-extension. Thus, the n-radius of $w \in W^P$ is w. For impossible worlds, on the other hand, the n-extension of $w \in W^I$ is more than just w. Generally, w' is an n-extension of $w \in W^I$ if and only if (1) $w' \in W^I$, (2) w' verifies what w verifies and (3) w' does not verify anything that cannot be inferred within n steps from what is verified by w. Take a look at the following example. Suppose that $V(w) = \{p \to q, p\}$, $V(w_1) = \{p \to q, p, q\}$ and $V(w_2) = \{p \to q, p, q \wedge q\}$. In this case, both w_1 and w_2 would be in the 1-radius of w, since they can be one-step inferred from w, assuming that R contains Modus Ponens and Conjunction Introduction. So, impossible worlds could have an n-radius bigger than one, contrary to possible worlds which only contain their own world.

For $\langle n\rangle Kp$ to be true, it is necessary that every doxastically accessible world contains at least one n-extension that verifies p. Therefore, we demand a tool that selects, for every doxastically accessible world from w, exactly one n-extension that verifies some formula. The following function will help us with that.

Definition 6.6 - Choice function
Let C be a choice function that, given a set $\mathcal{W} = \{W_1, ..., W_n\}$ of sets of worlds, returns a set $C(\mathcal{W})$ of sets of worlds, that gives all possibilities in which every set of worlds in $C(\mathcal{W})$ contains exactly one world of each set of worlds in $\mathcal{W}$.

This may sound complicated, so lets clarify what is stated in **Definition 6.6**. Suppose that $\mathcal{W} = \{\{w_1\}, \{w_2, w_3\}\}$. The choice function picks out exactly one world of each set of worlds in $\mathcal{W}$ until all possibilities are included. Then $C(\mathcal{W}) = \{\{w_1, w_2\}, \{w_1, w_3\}\}$.

The definition of the choice function and n-extension will be used to illustrate a relation $\backsim^n$ between two specific models. If this relation is present between two models, i.e. $(M, w) \backsim^n (M', w')$, then (M', w') is called n-accessible from (M, w). According to Fagin, this relation holds if and only if 'the set of doxastically accessible worlds from w in M is replaced in M' by a choice of n-extensions of the accessible worlds from w in M.' (Fagin et al. 1995).

So, there is a tool needed that can replace the set of doxastically accessible worlds with a choice of n-extensions of these worlds. For this purpose, n-alteration will come in handy. First, the formal definition will be given, after which more explanation about the n-alteration will follow.

Definition 6.7 - n-alteration
Let $M = \langle W^P, W^I, f, V\rangle$ be a model. F^n is a function from indicated models to sets of accessibility functions:
For $v = w$
$F^n(M, w) = \{g \mid g(v) = c\}$, where $c \in C(\{w'^n \mid w' \in f(w)\})$
For $v \neq w$
$F^n(M, w) = \{g \mid g(v) = f(v)\}$

If $g \in F^n(M, w)$, then g is an n-alteration of f.

Generally, the n-alteration works as follows. In a certain model M, the device looks at the accessible worlds $w' \in f(w)$ from w and determines the n-radius w'^n of these worlds. Note that w'^n is always a set of worlds. Now, we have $\{w'^n \mid w' \in f(w)\}$, containing all n-radii for worlds w'. Consequently, every element $X \in \{w'^n \mid w' \in f(w)\}$ is a set of worlds. Preceding, a choice set is established by applying a choice function on every set $X \in \{w'^n \mid w' \in f(w)\}$. Afterwards, a member of each choice will be picked and added to the choice set $C(\{w'^n \mid w' \in f(w)\})$, which is a new set of worlds. This alternative set of worlds can be seen as a set of worlds that are accessible from the initial world w. Finally, we define a function $\{g \mid g(v) = c\}$ for $v = w$ and a function $\{g \mid g(v) = f(v)\}$ for $v \neq w$. This means that the accessibility functions for all worlds that are not the initial world w remain the same, while the accessibility relations of world w are replaced with a member c of the choice set $C(\{w'^n \mid w' \in f(w)\})$ of n-radii of the accessible worlds w' from w. The new accessibility function g is placed in the function $F^n(M, w)$ and is now an n-alteration of f.

The above described procedure is applied in every possible way for every world $w \in M$. As a result, the set of all n-alterations $F^n(M, w)$ is obtained. Thus, if M is a model, then M' is n-accessible from M if all its elements are the same as in M, except for the accessibility relations, which are replaced by n-alterations g of f.

Briefly explained, **Definition 6.7** states that a function g is an n-alteration of f if and only if $g(w)$ is a choice of n-extensions of the accessible worlds of w. For more clarity, take a look at following example. Suppose that $f(w) = \{a_1, a_2\}$ and $g(w) = \{b_1, b_2\}$, then g is an n-alteration of f if and only if $b_1 \in a_1'^n$ and $b_2 \in a_2'^n$.

The formal definition of n-accessibility $\backsim^n$ follows from our explanation of n-alteration.

Definition 6.8 - n-accessibility
Let $M = \langle W^P, W^I, f, V\rangle$ and $M' = \langle W^{P\prime}, W^{I\prime}, f', V'\rangle$ be two models.
Then $(M, w) \backsim^n (M', w')$ if and only if:
a) $w' = w$,
b) $W' = W$,
c) $V' = V$ and
d) $f' \in F^n(M, w)$.

Definition 6.8 states that (M, w) can access (M', w') if and only if the set of doxastically accessible worlds $w' \in f(w)$ of worlds $w \in M$ is replaced in M' by a choice of n-extensions $\{w'^n \mid w' \in f(w)\}$ of those doxastically accessible worlds. Figure 6.1 provides an illustration of this concept. In the figure, a solid arrow from w to w' illustrates a doxastically accessible relation from w to w'. The dashed arrows from w' to w'', labelled n, represent that worlds w'' are n-extensions of w. We can see that in this figure, $(M, w) \backsim^n (M', w')$, since the set of doxastically accessible worlds $\{\alpha_1, \alpha_i, \alpha_r\}$ from $w \in M$ is replaced in M' by a choice of n-extensions $\{\epsilon_1, \epsilon_i, \epsilon_r\}$ of those doxastically accessible worlds (Bjerring and Skipper 2019).

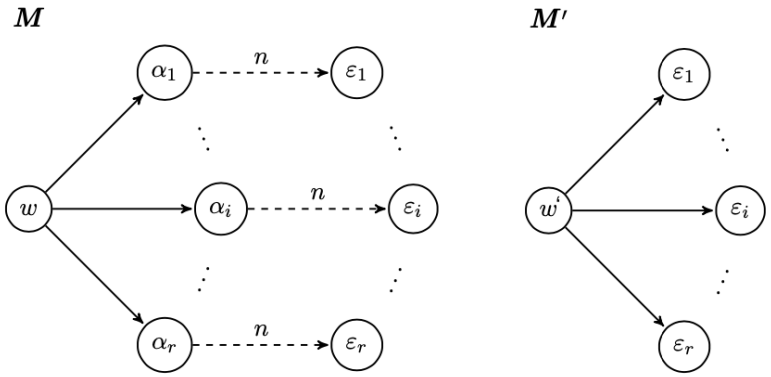

Figure 6.1: n-accessibility (Bjerring and Skipper 2019)

We can see the set of n-accessible models of a model M as all the different possibilities in which a agent can adjust his doxastic state, as a consequence of applying n trivial inference rules from R. This is the dynamic part of the theory, which facilitates the replacement of a model M with a different n-accessible model M' in n steps, representing an agent who can perform n steps of logical reasoning.

Finally, we have come to the last definition by which we will define truth in the target logic.

Definition 6.9 - Truth in the target logic
Let $M = \langle W^P, W^I, f, V \rangle$ be a model, then
For any possible world $w \in W^P$:

a) $w \vDash \phi$ iff $\phi \in V(w)$
b) $w \vDash \neg\phi$ iff $\phi \notin V(w)$
c) $w \vDash \phi \wedge \psi$ iff $w \vDash \phi$ and $w \vDash \psi$
d) $w \vDash K\phi$ iff $w' \vDash \phi$ for all $w' \in f(w)$
e) $w \vDash \langle n \rangle \phi$ iff $w' \vDash \phi$ for some w' such that $(M, w) \backsim^n (M', w')$
f) $w \vDash \langle n \rangle K\phi$ iff $w' \vDash K\phi$ for some w' such that $(M, w) \backsim^n (M', w')$

For any impossible world $w \in W^I$:

g) $w \vDash \phi$ iff $\phi \in V(w)$
h) $w \vDash \neg\phi$ iff $\neg\phi \in V(w)$
i) $w \vDash \phi \wedge \neg\phi$ iff $\phi \in V(w)$ and $\neg\phi \in V(w)$
j) $w \vDash \neg(\phi \vee \neg\phi)$ iff $\phi \notin V(w)$ and $\neg\phi \notin V(w)$
k) $w \vDash \phi \wedge \psi$ iff $w \vDash \phi$ and $w \vDash \psi$.

This definition deserves a little more explanation at some points. First of all, it is noteworthy that in a possible world $w \in W^P$, ϕ is either true (iff $\phi \in V(w)$) or false (iff $\phi \notin V(w)$), while in an impossible world $w \in W^I$, ϕ can also be both true and false (iff $\phi \in V(w)$ and $\neg\phi \in V(w)$) or neither true or false (iff $\phi \notin V(w)$ and $\neg\phi \notin V(w)$). Second, d) is a formal description of **Definition 4.4**. It states that someone has knowledge of a formula in world w if and only if this formula is true in every world that is doxastically accessible to the agent from w. Third, it is stated in f) that $\langle n \rangle K\phi$ is true at world w if and only if $K\phi$ is true at some model that is n-accessible from (M, w). Figure 6.2 provides an illustration of the semantics of $\langle n \rangle K\phi$. In this figure, $\langle n \rangle Kp$ is true at w, since p is a choice of n-extensions of the doxastically accessible worlds $\{\alpha_1, \alpha_i, \alpha_r\}$ from w. Therefore, there exists a possible model (M', w'), in which $w' \vDash p$ for all $w' \in f(w)$, such that $(M, w) \backsim^n (M', w')$. So, $\langle n \rangle Kp$ is true at w in the model of figure 6.2.

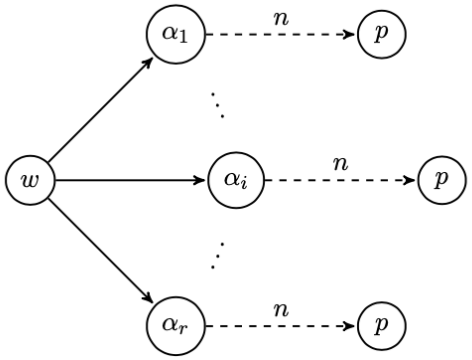

Figure 6.2: $\langle n \rangle Kp$ in the target logic (Bjerring and Skipper 2019)

In other words, namely those of Bjerring and Skipper, $\langle n \rangle Kp$ is true at world w if and only if 'p follows within n steps of reasoning from the truths at each accessible world from w.'(Bjerring and Skipper 2019).

The total semantics for our target logic is clear at this point, so now we can proof that it models logically non-omniscient, yet logically competent agents.

6.2 Solving logical omniscience

First, it will be shown how the target logic avoids logical omniscience. Since the model is a dynamized version of the impossible worlds model, the proofs are quite similar to those in section 4.2.1. For convenience, the invalidity of one rule, Closure Under Known Implication **(2)**, will be proved.

Theorem 6.1 - **(2)** is not valid in the target logic.

Proof. Suppose that $K(\phi \rightarrow \psi)$ and $K\phi$ is true at some possible world $w \in W^P$
By d) ϕ is true at all w' such that $w' \in f(w)$
However, ψ need not be true at all w' such that $w' \in f(w)$, since there could be some world w' in which $\neg\psi \in V(w')$
Therefore, $K\psi$ is not necessarily true at w
So, **(2)** is not valid in the target logic.

The proofs of **(1)**, **(4)** and **(5)** are so similar to the proofs using impossible worlds, that they need not be discussed in this chapter. The target logic invalidates all the principles that allow logical omniscience and therefore solves this problem. The only thing left for us to proof, is that the logic still models agents with some degree of logical competence.

6.3 Allowing logical competence

In chapter 5, logical competence is thoroughly discussed, explaining why this concept is so important in order to simulate how human agents reason. The purpose of this paper still is to model human reasoning and to model agents with different degrees of logical competence and the target logic is capable of achieving this goal. Before proving that the logic models logically competent agents, we need to know how agents gain knowledge of a formula ψ that follows from their current knowledge $\{\phi_1, ..., \phi_i\}$ within n steps of reasoning.

Corollary 6.1 - n-distribution:
If $\{\phi_1, ..., \phi_m\} \vdash^n \psi$, then $\{K\phi_1, ..., K\phi_m\} \vDash \langle n \rangle K\psi$

Theorem 6.2 - **Corollary 6.1** is valid in the target logic.

Proof. Let $M = \langle W^P, W^I, f, V\rangle$ be a model
Suppose that $\{\phi_1, ..., \phi_m\} \vdash^n \psi$ and $w \vDash \{K\phi_1, ..., K\phi_m\}$ with $w \in W^P$
We have to show that $w \vDash \langle n\rangle K\psi$
By f) $w \vDash \langle n\rangle K\psi$ iff $w' \vDash K\phi$ for some w' such that $(M, w) \backsim^n (M', w')$
By **Def. 6.8** $(M, w) \backsim^n (M', w')$ for some $M' = \langle W^P, W^I, f', V\rangle$,
where $f' \in F^n(M, w)$
By **Def. 6.7** $f(w) = c$ for some choice $c \in C(\{v^n \mid v \in f(w)\})$
By d) $w \vDash K\phi$ iff $w' \vDash \phi$ for all $w' \in c$
By **Def. 6.2** there is a choice $c' \in C(\{v^n \mid v \in f(w)\})$ such that,
if $M' = \langle W^P, W^I, f', V\rangle$, where $f'(w) = c'$, then $w' \vDash \phi$ for all $w' \in c'$
Given that $\{\phi_1, ..., \phi_m\} \vdash^n \psi$, there is a choice $c' \in C(\{v^n \mid v \in f(w)\})$ such
that, if $M' = \langle W^P, W^I, f', V\rangle$, where $f'(w) = c'$, then $w' \vDash \psi$ for all $w' \in c'$
By d) $w' \vDash K\psi$
By **Def. 6.7** $f' \in F^n(M, w)$
By **Def. 6.8** $(M, w) \backsim^n (M', w')$
So, for some model (M', w') such that $(M, w) \backsim^n (M', w')$
And $\{\phi_1, ..., \phi_m\} \vdash^n \psi$ and $w \vDash \{K\phi_1, ..., K\phi_m\}$, then $w' \vDash K\psi$
Thus, by f), $w \vDash \langle n\rangle K\psi$
So, **Corollary 6.1** is valid in the target logic.

Corollary 6.1 states that if some formula ψ follows from premises $\{\phi_1, ..., \phi_m\}$ within n steps of reasoning and if the agent has knowledge of all premises, then the agent will also come to know ψ after n steps of reasoning. We can look at **Corollary 6.1** as a dynamized version of Closure Under Known Implication **(2)**. **(2)** states that $K(\phi \to \psi) \to K\phi \to K\psi$. In this principle, knowledge is closed under entailment, while **Corollary 6.1** assumes no such principle. n-distribution merely states that an agent can gain knowledge about a formula ψ in n steps of reasoning from ϕ if ψ can be inferred from ϕ within n steps of reasoning.

Forthwith, we will continue to proof that the target logic allows logical competence. In chapter 2 and 5 we have seen that there are two important principles that should be retained in the target logic: Conjunction Elimination **(3)** and DeMorgan's Theorem **(6)**. **(3)** states that if an agent knows a conjunction of two formulas, than he also knows the two formulas independently. **(6)** states that logically competent agents know that if a conjunction is false, then at least one of the two formulas in the conjunction is false. First, a general theorem will be discussed, which proofs that the target logic allows logical competence. Thereafter, it will be proved that the model validates **(3)** and **(6)**.

Theorem 6.3 - The target logic allows logical competence.

Proof. Suppose $K\phi$ is true at w for some possible world $w \in W^P$
Consider any ψ, such that $\phi \vdash^n \psi$
Then, by **Theorem 6.2**, $\langle n\rangle K\psi$ is true at w

So, the target logic allows logical competence.

Theorem 6.3 states that agents can gain knowledge in some n steps of reasoning from what they already know. Important to notice here, is that they cannot gain knowledge about all consequences of their knowledge. Suppose that an agent has knowledge of ϕ at a possible world and that ψ follows from ϕ in more than n steps of logical reasoning, then the agent will not be able to know ψ at w, since he can only make n inferences from ϕ. Formally, if $w \vDash K\phi$ and $\phi \vdash^{>n} \psi$, then $w \nvDash \langle n \rangle K\psi$. Thus, there are still a lot of non-trivial consequences the agent will not be able to infer from the propositions he knows. This confirms that the target logic models logically competent agents, without making them logically omniscient. Generally, the model satisfies that the agent never overlooks a trivial consequence of his knowledge, but it also assures that agents cannot know *all* consequences of their knowledge.

With this general proof in mind, we will show that the target logic allows a dynamized version of **(3)**. Remember that **(3)** stated that $K(\phi \wedge \psi) \rightarrow K\phi$. It is possible for a person to know $\phi \wedge \psi$ without knowing ϕ. In this situation, however, it is very easy for him to come to know ϕ, since it is a part of the former. The content of ϕ and ψ is already known to the agent, so knowledge of ϕ can easily be gained. The dynamized version of **(3)** looks as follows:

(3)^d Dynamized Conjunction Elimination: $K(\phi \wedge \psi) \rightarrow \langle n \rangle K\phi$

Following theorem will show that **(3)**^{d} is valid in the target logic.

Theorem 6.4 - **(3)**^{d} is valid in the target logic.

Proof. Suppose $w \vDash K(\phi \wedge \psi)$ where $w \in W^P$
Obviously, $\phi \wedge \psi \vdash^n \phi$ for $n = 1$, assuming that Conjunction Elimination is in the set of inference rules R
Then, by **Theorem 6.2** $w \vDash \langle n \rangle K\phi$
Thus, if $w \vDash K(\phi \wedge \psi)$, then $w \vDash \langle n \rangle K\phi$
So, **(3)**^{d} is valid in the target logic.

In above theorem, ϕ is in the 1-radius of w, since ϕ can be one-step inferred from the premises that are true in w, assuming that R contains Conjunction Elimination. In chapter 5 we have seen why rational agents can easily make this implication. Knowledge of ϕ and ψ consistently implies knowledge of ϕ independently. Therefore, this trivial rule is contained in the set R of all logically competent agents.

Next to **(3)**^{d}, the dynamized version of **(6)** must be proven to be valid in the target logic. In the situation of **(6)** it is also really easy for an agent to come to know the right part of the implication. The content of not ϕ *and* ψ is again in the knowledge of the agent. Thus, he can easily make the implication

that either not ϕ or not ψ. Following is the dynamized version of **(6)**.

(6)^d Dynamized DeMorgan's Theorem: $K\neg(\phi \wedge \psi) \rightarrow \langle n \rangle K(\neg\phi \vee \neg\psi)$

To complete our proof, it will be proved that the target logic validates **(6)**^{d}.

Theorem 6.5 - **(6)**^{d} is valid in the target logic.

Proof. Suppose $w \vDash K\neg(\phi \wedge \psi)$, where $w \in W^P$
Obviously, $\neg(\phi \wedge \psi) \vdash^n \neg\phi \vee \neg\psi$ for $n = 1$, assuming that DeMorgan's Theorem is in the set of inference rules R
Then, by **Theorem 6.2** $w \vDash \langle n \rangle K(\neg\phi \vee \neg\psi)$
Thus, if $w \vDash K(\phi \wedge \psi)$, then $w \vDash \langle n \rangle K(\neg\phi \vee \neg\psi)$
So, **(6)**^{d} is valid in the target logic.

Again, the right part of the implication is in the 1-radius of w, assuming that DeMorgan's Theorem is contained in the set of trivial inference rules R. Chapter 5 explained why this rule is in R for ordinary human agents. The premises automatically imply the conclusion.

At this point it is fair to ask yourself why the Omniscience Rule, Closure Under Known Implication, Disjunction Introduction and the Equivalence Rule are not contained in the set R of trivial inference rules. If that would have been the case, namely, the consequences of those principles could also be 1-step inferred. The difference between these rules and **(3)** and **(6)** is that the former assume that agents create new knowledge of concepts they did not have any knowledge of before. **(1)** states that agents should have knowledge of every logical truth, which is obviously not possible since our knowledge space is not infinite. Therefore, this rule is not trivial and not in R. **(2)** claims that if an agent has knowledge of some implication and if he knows the first part of this implication, then he should also have knowledge of the second part. However, as we have seen in the example about Don, Joe and Jeff, one could fail to make the obvious implication and therefore could fail to know the second part. Thus, Closure Under Known Implication is not a trivial inference rule in R. **(4)** assumes that agents know the negation of the conjunction of a formula they know is false, with a random other formula. This implies some kind of closure under parthood which entails that everything one knows, is partly about every topic. For example, Francesca knows that Fred is not a football player. Disjunction Introduction states that now, Francesca also knows that Fred is not a football player *and* a fan of Bill Withers. Therefore, this rule states that Francesca's knowledge is partly about Bill Withers (which can be replaced with any other concept). To say, in this instance, that she knows nothing about Bill Withers, would be incorrect, since everything she knows would be partly about Bill Withers. Clearly, this rule implies that humans have infinite knowledge, assuming that they know the conjunction of something they know with any other information. Obviously, this is a non-trivial rule which is therefore not contained in

R. Finally, **(5)** states that if a double implication is true, and if an agent knows one of the formulas in this implication, then he also knows the other formula. In this situation a person should not necessarily have knowledge about the double implication and therefore is not committed to knowing the implied formula. It is clear to see that this principle is, again, not trivial. So, **(5)** is not included in the set of trivial inference rules R.

This chapter showed the semantics of the target logic, which is a dynamized version of Fagin's impossible worlds semantics. Different definitions demonstrated how the model works and showed when a formula is true in a world in the model. In sections 6.2 and 6.3, it was proved how the target logic solves logical omniscience and how it allows logical competence.

The model is capable of modelling agents with different reasoning capacities. We can change the value of n and the amount and variety of rules in R to determine someone's cognitive resources. With these means, we can model an agent with no cognitive resources (i.e. $n = 0$) and no inference rules (i.e. $R = \{\emptyset\}$). In this case $\langle n \rangle K\psi$ will be false for any inference, since nothing can be inferred in less than one step. Thus, this version of the target logic will model logically incompetent agents. On the contrary, we can also model agents with infinite cognitive resources (i.e. $n = \infty$) and a complete logical proof system for R. This will result in a logically omniscient agent and $\langle n \rangle K\psi$ will always be true if ψ follows from the agents knowledge in any number n of steps. Therefore, the target logic can model logically incompetent as well as logically omniscient agents. In between these extreme (and highly improbable) situations there is a wide spectrum of agents, just like you and me, with different reasoning powers and cognitive resources.

For all the reasons above, our goal to create a logic that models agents who are logically non-omniscient and logically competent, and that models agents of varying degrees of reasoning ability, is achieved by this target logic.

Chapter 7

Conclusion and discussion

The motivation behind this paper was to solve logical omniscience in modern modal logic. Logical omniscience is the problem that assumes agents to have knowledge about all logical truths and about all consequences of their knowledge (Parikh 1987).

This problem occurs in the standard Kripke semantics, a method that can usefully be applied to a lot of situations. Still, the model allows logical omniscience (Fagin et al. 1995), validating all principles that cause omniscience.

Minimal models (Chellas 1980) solve a couple of these principles, but still allow some degree of omniscience. It validates the Equivalence Rule, which should not be retained in the target logic that will represent human reasoning.

In chapter 4, two hyperintensional models are introduced (Sedlár 2019). First, awareness logic is proposed, which adds the notion of awareness to the possible world semantics (Fagin et al. 1995). This logic assumes that an agent should be aware of a formula and should have implicit knowledge of it in order to explicitly know the formula (Schipper 2014). The impossible worlds model adds the concept of impossible worlds to the already existing possible worlds model (Fagin et al. 1995). This approach states that agents can consider impossible worlds to be doxastically possible. Therefore, this method causes agents to make false assumptions and disables them to perceive all logical truths.

Both hyperintensional models succeed in solving logical omniscience, invalidating all principles that cause the problem. Nevertheless, these models consider agents to be logically incompetent as well. Both models assume that agents cannot apprehend any logical inferences of from their knowledge.

Still, people are able to grasp some logical consequences of their knowledge, which makes them logically competent agents (Cherniak 1981). The target logic, which will attempt to simulate human reasoning, should therefore allow logical competence, while avoiding logical omniscience.

For the target logic, a dynamic version of the impossible worlds model is proposed (Bjerring and Skipper 2019). The model succeeds in finding a balance between too much logical omniscience and too little logical competence. Furthermore, the logic is able to model agents with different cognitive resources.

These different levels of reasoning powers are captured by the step-based element in the model (Elgot-Drapkin et al. 1999). This feature enables different agents to be capable of making a different number of steps of reasoning. Therefore, the goal of the paper is achieved by this target logic.

The dynamic impossible worlds model could be an important model in Artificial Intelligence. A major ambition in this field is to create agents who reason like humans do. Therefore, an epistemic model representing human knowledge and reasoning is necessary. Since human agents are logically non-omniscient yet logically competent, the dynamic target logic is a suitable approach to model human reasoning.

Applying this model could lead to improved human-robot interaction. For example, if humans talk to artificial agents in natural language instead of code, this will narrow the semantic gap between the person and the agent, thus making the interaction between the two more efficient. However, for that to happen, the agent needs to know how to conduct an ordinary human conversation while being logically competent, without being logically omniscient. The dynamized model facilitates this notion, modelling agents who can reason like humans do.

Furthermore, agents should understand how other people reason. Suppose a situation in which you are unable to determine someone else's knowledge. For instance, you know that when a traffic light is green, you can drive and when it is red, you have to stop. Does this mean that you will feel safe while driving? Probably not, since you do not know whether other people also know this rule and, in your experience, they could possibly jump the lights. Thus, without knowing how other agents reason about knowledge, ordinary situations could become impractical. The target logic enables agents to reason about other agents' knowledge, since it allows agents to make a number of trivial inferences from their knowledge.

Concluding, the described dynamic impossible worlds model is a convenient option to model human reasoning. It allows logically competent agents, but avoids logically omniscient agents. Accordingly, agents can reason about their own knowledge as well as other agents' knowledge.

In further research, other techniques can be investigated, since the target logic is based only on the impossible worlds model. In the future, for example, there could be explored whether there is a version of awareness logic that does not eliminate logical competence. Furthermore, other studies could examine the notion of time in reasoning about knowledge, an issue that is not discussed in this paper. Additionally, there can be investigated whether different approaches can be combined, resulting in hybrid models.

Reasoning about knowledge is a challenging subject in modern logic, since the human brain is one of the most complicated things to understand. Therefore, there is more to developing an epistemic model, perfectly simulating human reasoning, than discussed in this paper. Further research is needed to explore other aspects of human reasoning such as intuition and reflection (Nagel 2014).

Conclusively, the proposed target logic is still assumed to be applicable in

a significant number of situations. In order to be certain of this assumption, however, more research should be done in this field.

Chapter 8

Appendix

Some logicians take a syntactic point of view, which replaces Kripke's truth assignment by a syntactic assignment. This syntactic assignment assigns to all formulas in all states a truth value. For instance, the syntactic assignment can assign both ϕ and $\neg\phi$ to be true in a state (Fagin et al. 1995). Another reply is of a semantic nature, representing an agent's knowledge by a set of sets of possible worlds (resembling minimal models). Comparatively, the semantic approach represents semantic knowledge by listing propositions one knows, instead of listing the formulas one knows in a syntactical manner (Fagin et al. 1995). A third reply is the nonstandard logic, that changes the notion of truth. The idea of this logic is that formulas ϕ and $\neg\phi$ are assigned truth values independently of each other. Consequently, ϕ can be true, disregarding the truth value of $\neg\phi$. To illustrate this concept, we can think of knowledge in nonstandard logic as consisting of databases of formulas. There is a database of true and a database of false formulas, such that ϕ is true when it is in the database of true formulas and $\neg\phi$ is true when ϕ is in the database of false formulas. Accordingly, since ϕ can be in both databases, ϕ and $\neg\phi$ can possibly both be true. For the same reason, neither ϕ or $\neg\phi$ could be true if no database contains ϕ (Fagin et al. 1995).

9 Bibliography

Berto, Francesco and Mark Jago (2018). "Impossible Worlds". In: *The Stanford Encyclopedia of Philosophy*. DOI: https://plato.stanford.edu/archives/fall2018/entries/impossible-worlds/.

Bjerring, Jens Christian and Mattias Skipper (2019). "A dynamic solution to the problem of logical omniscience". In: *The Journal of Philosophical Logic* 48, pp. 501–521. DOI: https://doi.org/10.1007/s10992-018-9473-2.

Blackburn, Patrick, Maarten de Rijke, and Yde Venema (2014). *Modal Logic*. Cambridge University Press.

Burgess, John (2011). "Kripke Models". In: *Cambridge University Press*, pp. 119–140. DOI: https://doi.org/10.1017/CBO9780511780622.006.

Cerro, Luis Fariñas del, Andreas Herzig, and Jérôme Mengin (2012). *Logics in Artificial Intelligence*. Springer, Berlin.

Chellas, Brian F. (1980). *Modal Logic: An Introduction*. Cambridge University Press. Chap. Minimal Models For Modal Logic.

Cherniak, Christopher (1981). "Minimal Rationality". In: *Mind* 358, pp. 161–183. DOI: http://www.jstor.org/stable/2253336.

Cresswell, M. J. (1975). "Hyperintensional Logic". In: *Studia Logica: An International Journal for Symbolic Logic* 1, pp. 25–38. DOI: https://doi.org/10.1007/bf02314421.

Ditmarsh, Hans van, Wiebe van der Hoek, and Barteld Kooi (2008). *Dynamic Epistemic Logic*. Springer Netherlands.

Elgot-Drapkin, Jennifer et al. (1999). "Active Logics: A Unified Formal Approach to Episodic Reasoning." In: *Institute for Advanced Computer Studies, University of Maryland*. DOI: http://hdl.handle.net/1903/1039.

Fagin, Ronald et al. (1995). *Reasoning About Knowledge*. The MIT Press; 2nd Printing edition.

Hawke, Peter, Aybuke Ozgun, and Francesco Berto (2019). "The Fundamental Problem of Logical Omniscience". In: *Journal of Philosophical Logic*. DOI: https://doi.org/10.1007/s10992-019-09536-6.

Hintikka, Jaakko (1962). *Knowledge and Belief*. Cornell University Press.

— (1975). "Impossible Possible Worlds Vindicated". In: *Journal of Philosphical Logic* 4, pp. 475–484. DOI: https://www.jstor.org/stable/30226996.

Jago, Mark (2006). "Hintikka and Cresswell on Logical Omniscience". In: *Logic and Logical Philosophy*, pp. 325–354. DOI: https://doi.org/10.12775/LLP.2006.019.

Nagel, Jennifer (2014). "Intuition, reflection, and the command of knowledge". In: *Proceedings of the Aristotelian Society,* 1, pp. 219–241. DOI: https://doi.org/10.1111/j.1467-8349.2014.00240.x.

Orlowska, Ewa (1990). "Kripke Semantics for Knowledge Representation Logics". In: *Studia Logica* 49, pp. 255–272. DOI: https://doi.org/10.1007/BF00935602.

Pacuit, Eric (2017). *Neighborhood Semantics for Modal Logic*. Springer.

Parikh, Rohit (1987). "Knowledge and the problem of Logical Omniscience". In: *Methodologies for Intelligent Systems, Proceedings of the Second International Symposium*, pp. 432–439. DOI: http://refhub.elsevier.com/S0168-0072(13)00102-4/bib506172383749534D4953s1.

Rendsvig, Rasmus and John Symons (2021). *Epistemic Logic*. Metaphysics Research Lab, Stanford University.

Sales, Dora and María Pinto (2016). *Pathways into Information Literacy and Communities of Practice*. Elsevier.

Schipper, Burkhard C. (2014). "Awareness". In: *SSRN*. DOI: http://dx.doi.org/10.2139/ssrn.2401352.

Sedlár, Igor (2019). "Hyperintensional logics for everyone". In: *Synthese* 198, pp. 933–956. DOI: https://doi.org/10.1007/s11229-018-02076-7.