Permitting Subtle Inconsistencies in Modal Space to Improve the Knowledge Representation of a more Human-like Agent

Sannerien van der Toorn

Supervisor: Dr. Colin Caret
Assessor: Dr. Rick Nouwen

A 7,5 ECTS Bachelor Thesis
in Artificial Intelligence

Utrecht University

Faculty of Humanities
Utrecht University
May 25, 2020

Abstract

The field of Artificial Intelligence endeavors to formally describe human knowledge. Modal logic, with its framework of *possible worlds*, tries to capture the knowledge representation of an agent. However, some properties that follow the definition of knowledge assume an omniscient reasoner. To avoid the omniscience problem, *impossible worlds* can be added to the world idiom.
This work addresses which specification of the definition of impossible worlds should be used to represent human knowledge, and thus to avoid logical omniscience. First, by means of a discussion of the approaches of Jago [1] and Bjerrings [2] in the literature on impossible worlds. Second, by combining the proposal of both [1, 2] in a suggestion for a model that uses an distinguishment between blatant and subtle impossible worlds and by permitting *partial worlds*. Next to this, the model introduces a concept of *inconsistency values*. This value provides a deeper insight into the properties of impossible worlds that are most representative for modeling a human-like agent. This work suggests that an improved knowledge representation for an agent who is neither omniscient, nor too unintelligent lies in line with permitting partial worlds with their subtle inconsistencies. Such knowledge representations could be implemented in the area of Artificial Intelligence which is concerned with the creation of intelligent systems that reason like humans do.

Keywords **– omnicience problem, modal logic, impossible worlds, partial worlds, knowledge representation**

Contents

Chapter 1

Introduction

The field of Artificial Intelligence endeavors to formally describe human knowledge. Such formalizations can be used in robots that have social purposes. For example in healthcare, where one aims to develop robots that are intended to imitate human behavior and contact. Additionally, such formalization can be used to design an individual agent that can reason about his knowledge and make new inferences. In intelligent systems with such an individual agent, the inferences and knowledge development of the agent can then be automated [3].

Humans use their knowledge to reason about the epistemic possible versions of the world. For example, consider whether or not it is currently raining. You do not have knowledge of this fact until the moment that you open the curtains (or check the weather forecast). Modal logic, and in particularly its framework of *possible worlds*, tries to capture the different epistemic possible version of the world for an agent and describes when an agent indeed knows a fact. Possible worlds are points in the model in which formulas are evaluated and in which the rules of classical logic apply. Within this framework, "knowing" a fact α, requires α to be true in all the worlds the agent considers possible [4, p. 15]. Based on this definition, knowledge has multiple properties. One of which is the *Distribution Axiom*:

$$\vDash \mathbf{K}_a\phi \wedge \mathbf{K}_a(\phi \rightarrow \psi)) \rightarrow \mathbf{K}_a\psi$$

This axiom implies that each agent knows all the logical consequences of his knowledge [4, pp. 32–33]. For humans however, this is an impossible task; they are non-ideal reasoners [5]. We do not posses the cognitive capabilities to oversee all such consequences. Contributing to this problem is the *Knowledge Generalization Rule*:

$$\text{For all models } \mathcal{M}, \text{ if } \mathcal{M} \vDash \phi \text{ then } \mathcal{M} \vDash \mathbf{K}_a\phi$$

This rule implies that an agent knows all the formulas that are valid in a model [4, pp. 32–33]. This assumes an agent with unlimited knowledge of, for example,

all logical truths. Both these rules thus describe an agent as 'logically omniscient', which is something that we want to avoid when modeling the knowledge of human-like agents. There exist multiple approaches to solve the problem of logical omniscience. The most expressive one is the addition of *impossible worlds* [4, p. 374]. An impossible world also is a point in the model in which formulas are evaluated, but, other than in a possible world, the rules of classical logic do not apply here [6, p. 100].

For example, a tautology could be false in an impossible world. If an agent a considers such an impossible world as epistemic possible, a fails to know the tautology. The impossible world in this example provides a counterexample for an otherwise, omniscient agent a. The precise definition of an impossible world remains ambiguous because there are many ways in which one can define what 'not following the rules of classical logic' means. In this research we investigate which definition of impossible worlds would best describe a human-like agent. To this end, we consider two interpretations.

The first one is the interpretation of an *open world*, described by Jago [1]. These worlds are not closed under any logical rules and whole sentences are assigned a truth value instead of loose primitives. The fact that these worlds are not closed under any logical rules makes this framework from a logical perspective quite extreme. Adding such open worlds to our framework of knowledge therefore allow us too study the extreme side of eliminating logical rules that are to strong for human reasoners.

The second interpretation is described by Bjerring [2]. He claims that the modal space should only contain possible and non-trivial impossible worlds. Within these non-trivial impossible worlds, only subtle inconsistencies are allowed. Blatant inconsistencies should be left out as they can easily be inferred by a competent agent. Both these interpretations of impossible worlds are further discussed in chapter 3.

This work addresses which specification of the definition of impossible worlds should be used to represent human knowledge and thus to avoid logical omniscience. In chapter 4, we discuss our efforts to improve the definitions by Jago and Bjerring, but before further discussing the theories on impossible worlds, we give an overview of the framework of possible worlds and the resulting problem of omniscience.

Chapter 2

Possible worlds

Epistemology started as a philosophy study before Jaakko Hintikka made the first effort in formalizing its concepts in 1962. In this year, he published the book 'Knowledge and Belief', in which he laid the foundation of epistemic logic. Within this book, he introduced the following central notation [7]:

$\mathbf{K}_a p$ which is read as "Agent a knows that p"

S.A. Kripke further extended Hintikka's ideas with mathematical aspects [8] resulting in a framework named after its creator; the Kripke model. Within this model, epistemic alternatives of a certain agent, which are called *possible worlds*, can be represented. The concepts and notations of the Kripke model will also be used in this paper. In this chapter we first explain the syntax and semantics of standard epistemic logic, and then discuss the unsatisfactory result of this model; the logic omniscience problem.

2.1 Framework of knowledge

Within our framework, we will use the formalization of propositional logic to abstract from sentences of natural language. This logic uses the well-known connectives $\neg, \wedge, \vee, \rightarrow$, and $\leftrightarrow$ to relate propositional variables such as ψ and ϕ to each other. With modal logic, an addition to this syntax is made with the operator $\Box$ for necessity and, the operator $\Diamond$ for possibility. A formula $\Box\psi$ is read as 'It is necessarily the case that ψ', and $\Diamond\psi$ is read as 'It is possibly the case that ψ'. Since these two operators are each others duals [1], only one operator will be used as a primitive symbol. We choose to use the operator for necessity, because the modality K for 'knows that' has similar semantics to the necessity operator quantifying over all epistemically possible worlds [10]. Before we give the definition of this modality K, we will first introduce the Kripke model in which K is evaluated.

[1] $\Box\ \psi$ is equivalent to $\neg\Diamond\neg\ \psi$, and $\Diamond\ \psi$ is equivalent to $\neg\Box\neg\ \psi$ [9, p. 4]

2.2 Kripke model

A Kripke model consist of a Kripke frame together with a valuation, these are defined in the following manner [3]:

Definition 2.2.1 (**Kripke frame**)**.** A **Kripke frame** consists of a tuple $\mathcal{F} = \langle W, R \rangle$, such that:

- W is a non-empty set of possible worlds;
- $R \subseteq (W \times W)$ is a binary relation on W; if wRv, we say that the world v is accessible from w.

Definition 2.2.2 (**Kripke model**)**.** A **Kripke model** consists of a tuple $\mathcal{M} = \langle W, R, V \rangle$, such that:

- $\langle W, R \rangle$ is a Kripke frame underlying $\mathcal{M}$;
- $V : W \to \mathcal{P}(\text{VAR})$ is a valuation of the set of atomic propositions VAR; proposition p is true in world w if $p \in V(w)$, and thus false in w if $p \notin V(w)$.

The addition of the relations between worlds in a model provides a way of evaluating formulas that contain the necessity operator ($\Box$). These valuations assign truth values according to the standard classical truth functions.

Definition 2.2.3 (**Necessity ($\Box$)**)**.**
The truth value of the modal operator of necessity ($\Box$) is defined as follows:

- Within a model $\mathcal{M} = \langle W, R, V \rangle$, world $t \vDash \Box\psi$ if, and only if, for **every** world u such that tRu, $u \vDash \psi$.

  From this it follows that within a model $\mathcal{M} = \langle W, R, V \rangle$, world $t \vDash \neg\Box\neg\psi$ if, and only if, for **at least one** world u such that tRu, $u \vDash \psi$. [2]

2.3 Knowledge representation

In a Kripke structure for standard epistemic logic, the knowledge operator K is treated similar to the necessity operator. A knowledge operator ($\mathbf{K}$) is connected with an agent a by the addition of a subscript ($\mathbf{K}_a$). In this way, the knowledge of a specific agent is represented and we can distinguish the knowledge of different agents[3]. The same principle is used for the accessibility relations between worlds.

[2] This is the truth assignment for the possibility operator $\Diamond$.

[3] The reader should bear in mind that this study is, however, only based on representing the knowledge of one agent.

Definition 2.3.1 (**Knowledge of an agent**).
Within a model $\mathcal{M} = \langle W, R, V \rangle$, $t \vDash \mathbf{K}_a\psi$ if, and only if, for **every** world u such that tR_au, $u \vDash \psi$.

Beside the definition of truth, a definition of validity is necessary to define the modal logic semantically.

Definition 2.3.2 (**Validity on a Model**). $\mathcal{M} \vDash$
ψ is valid on M, if and only if, $\mathcal{M}, w \vDash \psi$, for all possible worlds w in $\mathcal{M}$.[4]

Remark. Therefore, if we have a sentence p that is valid on a model, p is true in all worlds, and by definition 2.3.1, it follows that for every agent a in that model, $\mathcal{M} \vDash \mathbf{K}_a\psi$.

To get a better understanding of the formal definitions in this chapter, and of how knowledge and epistemic possibilities can be represented within a Kripke model, we will now discuss an example.

2.3.1 Example of a Kripke Model

Imagine a person - let us call her Lisa - is walking to her shed to grab her bike to leave for school. However, she does not find her bike in the shed. First, Lisa questions her own memory. She might not have placed her bike in the shed at all, she could have placed it down the street (w_1). However, she also considers the possibility that a thief stole her bike (w_2). Thereupon, she wonders if her roommate has taken her spare key and lent her bike without asking permission (w_3). These three scenarios are all epistemic possibilities for Lisa.

Does Lisa know she has a bike (we call this p) according to the semantics of modal logic? And does she know that her bike is not in the shed (we call this q)? The question whether Lisa is able to have knowledge of the proposition p and q can be made transparent by representing it in a Kripke model (see figure 2.1):

[4]This validity can be extended to a frame, a class of frames and to general validity [3].

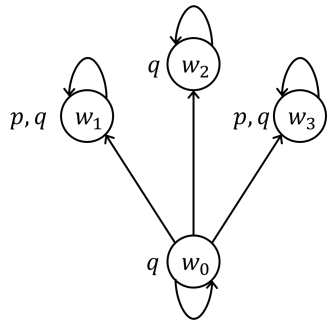

Figure 2.1: The example of Lisa's bike

- $\mathcal{M}_1 = \langle W, R, V \rangle$, where
- $W = \{w_0, w_1, w_2, w_3\}$,
- $R_l = \{(w_0w_0), (w_0w_1), (w_0w_2), (w_0w_3), (w_1w_1), (w_2w_2), (w_3w_3)\}$,
- $V(w_0) = \{q\}, V(w_1) = \{p, q\}, V(w_2) = \{q\}, V(w_3) = \{p, q\}$

In this model, Lisa is located in world w_0 and her three epistemic possibilities of what happened are w_1, w_2 and w_3. The arrows to these worlds represent the idea that Lisa thinks of them as epistemic possibilities. In this model the following things can be concluded:

$\mathcal{M}_1, w_0 \vDash \mathbf{K}_l q$, since w_0, w_1, w_2 and w_3 are the only reachable worlds from w_0, and $w_0 \vDash q$, $w_1 \vDash q$, $w_2 \vDash q$ and $w_3 \vDash q$. It follows that $w_0 \vDash \mathbf{K}_l q$. So q is necessarily true in w_0 and hence Lisa knows that her bike is not in the shed.

$\mathcal{M}_1, w_0 \vDash \neg\mathbf{K}_l \neg p$, because $w_0 R_l w_1$ and $w_1 \vDash p$ together imply that $w_0 \vDash \neg\mathbf{K}_l \neg p$. So, p is possible true in w_0 and hence Lisa does not know that she does not have a bike.

2.4 The logical omniscience problem

Based on the properties we discused earlier, the Kripke model seems valuable for representing the knowledge of an agent. However, some properties that follow throw a spanner in the works when it comes to representing knowledge in a human-like manner. As already remarked at definition 2.3.2 and in the introduction, valid formulas in a model imply the knowledge of these formulas for an agent. More formally, this is the *Knowledge Generalization Rule*.

$$\text{For all models } \mathcal{M}, \text{ if } \mathcal{M} \vDash \phi \text{ then } \mathcal{M} \vDash K_a\phi \text{ [4]}.$$

This shows us that the standard epistemic logics assume an agent with unlimited knowledge of, for example, all logical tautologies, including the ones that are hard to derive and cannot instantly be spotted. Furthermore, knowledge has the following property, named the *Distribution Axiom*:

$$\vDash K_a\phi \wedge K_a(\phi \to \psi)) \to K_a\psi \text{ [4]}$$

This axiom implies that each agent knows all the logical consequences of his knowledge [4, pp. 32–33]. Knowing each logical consequence of one's knowledge is, however, an impossible task for a human, due to their limited cognitive capabilities.

For example, imagine the student Dylan that has just started his study of Logic. He knows that the tautology $\rho : (p \vee \neg p)$ is trivially true. Additionally, there is a tautology $\xi : (p \to (q \to r)) \to (p \wedge q \to r))$ that is non-trivial and more difficult to deduce the consistency from for Dylan. The *Knowledge Generalization Rule* now implies that Dylan has knowledge of both these tautologies. However, Dylan is not familiar with the tautology s, and even if he was presented with it, he would not be able to deduce the truth table. Also, if our agent Dylan would be a well-trained logic professor, the same result could be established with a harder tautology.

The problem that is described here is the problem of logical omniscience. We want to avoid logical omniscience when modeling the knowledge of human-like agents. In the following chapters, a solution to the logical omniscience problem, by means of the addition of *impossible worlds* to the framework of possible worlds is investigated.

Chapter 3

Impossible world approaches

As explained in the previous chapter, possible worlds turned out to be an overly idealized version of knowledge representation which culminated in the logical omniscience problem. Questions have been raised about the use of epistemic logic for human-like knowledge representation [11]. However, several serious approaches to deal with logical omniscience within epistemic logic have been discussed in literature [4]. Given that the tools of modal logic will continue to be used, but avoiding logical omniscience is desired, the framework of possible worlds needs to be extended with the use of *impossible worlds*. The literature gives multiple interpretations of what such an impossible world is exactly. In this chapter, we will explore two of these possible interpretations [1, 2]. First, however, we will explain how impossible worlds help us to avoid the problem of logical omniscience in the following section.

3.1 Omitting logical omniscience with impossible worlds

In our previous described model of possible worlds in section 2.4, an agent had no limit on its knowledge capacity. With the arrival of impossible worlds, the 'world idiom' is extended with a new kind of world which can introduce such a limit. The set of worlds W now consist of a set of possible worlds P and a set of *impossible worlds* $I = W - P$. Possible worlds can have accessibility relations to these impossible worlds. Impossible world have the important property of being able to 'break the rules of classical logic'. Which specific rules can be broken, depends on the types of impossible worlds. However, in each case, such an impossible world generates a counterexample of an otherwise omniscient agent. To illustrate this, remember Dylan from the example in section 2.4. He was expected to know a difficult tautology (ξ) to be true according to the rules of

classical logic. Now, we include the impossible world q in the model with the valuation $\xi \notin V(q)$, i.e. ξ is false in world q. Additionally, we add an accessibility relation for Dylan to this impossible world q. As a result, it is no longer the case that ξ is true in every world that is accessible to Dylan. According to definition 2.3.1, Dylan no longer has any knowledge of ξ in this model, making him non-omniscient. [1]

This example shows that omniscience can be avoided by adding a place in the model in which the rules of classical logic do not apply. At the same time, this addition raises question regarding the consequences on the framework of worlds. How does one define logical truth in a framework containing both possible and impossible worlds? Do we allow every inconsistency in our impossible worlds? Jago [1] and Bjerring [2] have both thought about these consequences and came up with their own versions of an impossible world. The rest of this chapter describes and analyzes their approaches.

3.2 Jago's approach

In [1], Jago reviews Cresswell's approach [12] of the addition of a new type of world. Although he criticizes this approach, his final proposal is an addition to this model.

Cresswell introduces an idea that does not follow the classical rules of logic, but also does not tolerate the impossible [1]. Therefore, his addition of another type of worlds is not named impossible worlds, as there is not happening the impossible, but rather *non-classical worlds*, as the connectives have a nonstandard meaning. A mathematical function in every world denotes the link between a connective and its denotation, giving classical values at classical worlds [1], and non-classical values in non-standard worlds[2].

Jago criticizes this approach because it does not maintain the idea that agents gain knowledge by eliminating scenarios that are considered as possibilities. He states that an agent, with the simple knowledge of how truth is normally assigned to two formulas connected by a connective, can easily differentiate between classical and non-classical worlds. For example, imagine a reasonable agent observing a world w_0 in which the following valuation is given: $p \in V(w_0)$, $q \notin V(w_0)$ but $(p \wedge q) \notin V(w_0)$. Every agent that has the minimal understanding of a conjunction will notice that this valuation is not classical. An agent then loses the epistemic access to such a world and it is back to square one; it knows all classical tautologies.

Hence, Jago states that non-classical worlds need to be genuinely impossible worlds for achieving their purpose of bypassing omniscience. He proposes to allow contradictions to have a positive truth value. As a result, worlds are paraconsistent, which means that the valuation V of primitives can have the value $true$, $false$, $true$ & $false$ or *neither one of them*. Due to these extra

[1] This idea was first explored by Hintikka in his 1962 book 'Knowledge and Belief'.

[2] A similar method is used in AI by Levesque for his logic of implicit belief [13].

values of validations, $false$ does not logically follows from $\neg true$ anymore. The definitions of truth and falseness are redefined as follows:

Definition 3.2.1 (**Truth validation for non-classical worlds**).

- $w \vDash_t p$, if and only if, $true \in V(p, w)$
- $w \vDash_f p$, if and only if, $false \in V(p, w)$
- $w \vDash_t \neg\phi$, if and only if, $w \vDash_f \phi$
- $w \vDash_f \neg\phi$, if and only if, $w \vDash_t \phi$
- $w \vDash_t \phi \wedge \psi$, if and only if, $w \vDash_t \phi$ and $w \vDash_t \psi$
- $w \vDash_f \phi \vee \psi$, if and only if, $w \vDash_f \phi$ or $w \vDash_f \psi$

The classical worlds have precisely one element ($False$ or $True$) in the valuation set V for every primitive in order to keep them classical. In order to keep the definition of knowledge behave classically, we define it in terms of $\vDash_t$:

Definition 3.2.2 (**Knowledge for non-classical worlds**).

- $w \vDash_t \mathbf{K}_a\phi$ iff $w' \vDash_t \phi$ for all worlds w' such that wR_aw'
- $w \vDash_f \mathbf{K}_a\phi$ iff it is not the case that $w' \vDash_f \mathbf{K}_a\phi$

The aforementioned idea that an impossible (non-classical) world can now form a counterexample for an otherwise omniscient agent is accomplished. For example consider the following model:

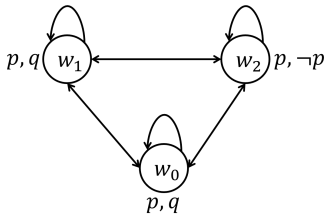

Figure 3.1: Non-classical worlds example

In this model, world w_2 is a non-classical world, which yields the possibility of an empty valuation set for q in world w_2. From this model it can be concluded that $(p \rightarrow q)$, or in other words $(\neg p \vee q)$, is true in all worlds. Therefore, $\mathbf{K}_a(p \rightarrow q)$ is true in all worlds, regardless of the world in which the agent is located. Also $\mathbf{K}_a p$ holds due to the fact that p is true in every world. However, $\mathbf{K}_a q$ does not hold due to the undefined valuation of q in world w_2. The principle of closure thus fails in this model.

It appears that Jago found what he was looking for. However, he discovered another type of omniscience in these non-classical worlds. The closure under

addition and deletion of double negations, for example, ensures that part of the omniscience survives when the chain of negation becomes too long. One double negated primitive is easily deducted to be true by a rational agent. However, if the chain of negations becomes so long that it exceeds the capability of a rational agent to count them all and deduce the valuation of the primitive, it cannot be known to be true. This problem exists in every formula that has too many connectives and, although this omniscience is not as inadmissible as the ones earlier described in section 2.4, Jago states that, to reach a total omniscience-free-space, it also should be disposed of. Thereupon, Jago introduces his final and completely omniscience-free type of worlds: *open worlds*.

As the name might suggest, open worlds are completely open in the sense that they are not closed under any rule of inference [1], except that p is inferred from p. To still be able to use the connectives somewhat meaningfully, all sentences, not just the atomic sentences, are now assigned truth values arbitrarily and therefore may behave arbitrarily [14]. In a model with open worlds, endlessly long sentences with double negations are not a problem anymore because the valuation of such sentences is given at once.

Jago points out that although open worlds finally reach the goal of abandoning omniscience, a model including all of them loses any interesting logical properties [1]: even the most obvious contradictions could be true in an open world. This brings us to the other side of the problem: modeling obvious contradictions. Jago states that, in order to model rational agents, such obvious contradictions should not be included in the epistemic accessible space. Only open worlds that do not consist of sentences from which explicit contradictions can be inferred by a rational agent, should be accessible for such a rational agent. Bjerring's idea of impossible worlds continues this line of thought and separates obvious contradictions from subtle ones.

3.3 Bjerring's approach

Bjerring's aim is to construct a modal space in which the knowledge of an agent that has bounded but non-trivial cognitive and computational resources can be modelled [2]. Such an agent could have subtle inconsistencies in his set of knowledge, but not blatant ones. Therefore, such an agent should only have access to subtly impossible worlds but no access to blatantly impossible worlds. This ensures that the agent does not see blatantly impossible worlds and their trivial inconsistent content as epistemic possibilities.

Bjerring reinforces this idea by introducing a precise notion of the difference between the subtle and blatant inconsistencies. Bjerring highlights that this distinction depends on the kind of agents that should be modeled. A well-trained logic professor can deduce the falseness of far more difficult, negated tautologies than a first-year logic student can. For his model, Bjerring wants to model a moderately ideal agent, which is more closely resembled by the student than the professor in our example. This agent possesses the inference rules of logic and their effects. Consequently, it is able to understand and have

knowledge of the result of applying a logical inference rule once to a logical sentence. Therefore, Bjerring chooses to categorize inconsistencies in terms of the amount of inference rules that are required to reach an explicit contradiction of the form $\psi \wedge \neg\psi$. Bjerring bounds the blatant inconsistencies at one inference step. He makes the following three small adjustments in order to make this restriction work [2]:

- In one inference, both ψ and ϕ are inferred form a conjunction ($\psi \wedge \phi$);
- In one inference, both ψ and $\neg\phi$ are inferred from a negated conditional $\neg(\psi \rightarrow \phi)$;
- Agents can exclude the set $\{\psi, \neg\psi\}$ without making an inference.

With these three stipulations, Bjerring reaches his formal definition [2] of blatancy.

Definition 3.3.1 (**Blatant**)**.** A sentence ψ or a set Γ of sentences is blatantly inconsistent, if and only if, a contradiction $\{A, \neg A\}$ can be inferred from ψ or Γ by use of at most one application of the inference rules; otherwise, if inconsistent, ψ or Γ is subtly inconsistent.

With this definition, impossible worlds that contain a blatant inconsistency can now formally be distinguished from the other worlds and can be made non-accessible for our agent. This seems to be a solution. However, Bjerring then defines the following theorem [3]:

Theorem 3.3.1 (**Bjerring's trident**)**.**
(Result) There is no modal space such that:

- (R1) there are impossible worlds;
- (R2) there are no partial worlds;
- (R3) there are no blatantly impossible worlds.

Although the conclusion that it is not possible to construct such a modal space is unfortunate, it is not the terminus. Bjerring concludes that to still be able to make a modal space for a non-omniscient but non-trivial agent, either R1, R2, or R3 should be excluded. Because we chose to evade logical omniscience by using impossible worlds, R1 cannot be excluded. R3 is also retained because the modal space otherwise collapses to the aforementioned result of an unintelligent agent that does not represent the idea of a 'moderately non-ideal agent'. According to Bjerring, the most appealing solution that remains is letting go of R2; investigating a modal space that includes partial worlds. Partial worlds are worlds that are not maximal, which means that for the sentences in a partial world, the truth value does not have to be either true or false, it could also be

[3] A discussion of this theorem lies beyond the scope of this study but can be read in Bjerring's paper[2].

undefined [15]. In the next chapter we further discuss the use *partial worlds* in combination with the application of Bjerring's distinction between blatant and subtle inconsistencies and suggest an improved model for human-like knowledge representation.

Chapter 4

Modal space with inconsistency values

As Bjerring already argued in his approach [2] discussed in the previous chapter, there is a demand for a distinction between blatantly and non-trivially impossible worlds. However, while adding these non-trivially impossible worlds, Bjerring concludes that partial worlds should be accepted in order to construct a working modal space. Partial worlds are worlds that are not maximal. This means that for sentences in a partial world, the truth value does neither have to be true nor false; it could also be undefined [15].

In this chapter, we propose a new distinction between impossible worlds that builds upon Bjerring's ideas for a formal division of impossible worlds. We introduce the concept of an *inconsistency value* ($\mathcal{I}$) that is given to every world. In the first section (4.1), we present the formal definition of this inconsistency value, after which we explain the interpretation of the value $\mathcal{I}$ in section 4.2. In section 4.3 the new modal space is described together with a discussion of the addition of the partial worlds. Turning now to the formal definition of the inconsistency value.

4.1 Measuring inconsistencies numerically

Recall the example of the beginning logic student Dylan in section 2.4, in which the tautologies ρ and ξ were discussed. There was only one inference step needed for ρ, but considerable more inference steps for ξ. According to Bjerring's approach, ρ therefore is classified as a blatant inconsistency and ξ as a subtle inconsistency. We adopt Bjerring's approach of differentiating between subtle and blatant impossible worlds based on the number of inference steps that is required to infer an inconsistency. We use this number to provide each impossible world with an inconsistency value $\mathcal{I}$. An inconsistency value $\mathcal{I}$ is a real number $0 \leq \mathcal{I} \leq 1$. The intuition behind this value $\mathcal{I}$ is that a zero value reflects a world that is possible (consistent), an one reflects a world that is blatantly

impossible, and every value between zero and one represents a world with a subtle inconsistency. The closer the value is to one, the simpler it is to deduce the inconsistency from the sentence, or set on sentences, in that world. The formal definition is as follows:

Definition 4.1.1 (**A model with inconsistency values**)**.** Given a world w with a sentence ψ or a set Γ of sentences and a threshold $\frac{1}{\beta}$ [1], the *inconsistency value* $\mathcal{I}$ of w is given by the number of inference rules $\boldsymbol{n}$ that need to be applies to ψ or Γ to obtain a contradiction of the form $\{A, \neg A\}$:

$$\textit{inconsistency value } \mathcal{I}(w) = \begin{cases} \frac{1}{\beta}, & \text{if } n = 0 \\ \lim_{n\to\infty} \frac{1}{n} = 0, & \text{as } n \to \infty \text{ (possible worlds)} \\ \frac{1}{n}, & \text{otherwise} \end{cases}$$

The definition contains an exception for the case of $n = 0$ because $\frac{1}{0}$ is not defined. To still be able to work with inconsistencies in which $n = 0$, their inconsistency is set to $\frac{1}{\beta}$ [1], which corresponds to a blatantly impossible world. In possible worlds no inconsistencies can be derived, as they are consistent. In the definition of $\mathcal{I}$, this is reflected by n tends to infinity. Because $\lim_{n\to\infty} \frac{1}{n} = 0$, possible worlds get an inconsistency value of 0. [2]

4.2 Interpretation of inconsistency values with a threshold

Let us now consider the value $\frac{1}{\beta}$, used in the definition of $\mathcal{I}$: a threshold, such that worlds can be evaluated for acceptance for an agent. This results in a modal space where the possible worlds and the subtle impossible worlds are below the threshold and blatantly impossible worlds are on or above the threshold. This allows us to distinguish the unwanted impossible worlds for an agent from the rest. The inconsistency value for an agent is an epistemic value that indicates the importance of an epistemic possibility. If this value is below the threshold, the credibility that an agent has of this possibility is not affected. However, if the inconsistency value is equal or above the threshold, the agent treats this possibility as a worthless one, as it is not accessible to him.

The interpretation of an inconsistency value can be compared with a degree of belief for inconsistencies. The degree of belief is a formal approach in epistemology that takes a probabilistic view on the epistemic possibilities for an agent. According to the Lockean thesis, this quantitative notion of belief is linked to the qualitative notion by means of the comparison with a certain threshold [17]. Hence, an agent should believe a proposition if and only if his degree of belief for this proposition is higher than a certain threshold [17]. The interpretation of the inconsistency value is related to this idea, but it is not the same. For an

[1] An explanation of this value is given in section 4.2.

[2] A prove that the limit of $\frac{1}{n}$ converges to 0 is given in [16].

agent, the threshold for what the agent takes as a blatant inconsistency should be established. As mentioned in the example of the student and the professor in section 3.3, this depends on the number of inference rules, which we denote as β, an agent can apply while still maintaining the understanding and knowledge of the result. This value depends on the logical and computational skills of the agent. Bjerring [2] took the initial threshold to be 1, assuming a fairly incapable agent. A human-like agent reasons non-ideal, but the discussion to what extent this non-ideality should reach, can be conducted using the inconsistency values. For simplicity, the value of 1 is chosen as the default, but other higher values for β are also possible. The threshold becomes $\frac{1}{\beta}$, so $\frac{1}{1} = 1$. A new stipulation to the accessibility relation for each agent is suggested with the following rule:

Definition 4.2.1 (**Accessibility relation with a threshold**)**.** For a model $\mathcal{M} = \langle W, R, V \rangle$ with a threshold β_a for an agent a; $R_a \subseteq (W \times W)$ is a binary relation on W, where for every world $w \in W$ it must follow that $\mathcal{I}(w) < \frac{1}{\beta}$. Otherwise $w \notin R_a$.

As a result, we can have different epistemic logics for each threshold, but they all follow the same definition of knowledge as given in definition 2.3.1.

4.3 The new modal space

The modal space with the filter on the accessibility relations causes the model to contain partial worlds. The reason for the appearance of partial worlds can be best explained in an example.

Imagine a subtle impossible world w in which the valuation of the set of sentences $\{X, Y, Z\} \in V(w)$ and with a threshold of $\beta = 1$. Because it is subtle inconsistent and the threshold $\beta = 1$, there is some contradiction of the form $\{A, \neg A\}$ that can be inferred from the set of sentences, but it takes more than one step to do so. Therefore, it follows that the contradiction of the form $\{A, \neg A\}$ is not already in the valuation of world w. It turns out that in order to reach $\{A, \neg A\}$ $n = 2$, which implies that there is an intermediate proof step. T denotes the step that lies between the current set of sentences and the contradiction. Since we are assuming this world is not blatant, we know that the agent cannot reach the contradiction $\{A, \neg A\}$ in one step. That means that the intermediate formula T is not already represented as being true in this world. On the other hand, since we can reach T in one step, it also means that $\neg T$ is not represented as being true in w, otherwise there already would be a contradiction in one step.

This example could be extended for any subtle inconsistency that takes any lager number of steps to reach a contradiction [3]. Independent of the number of undefined truth values in a model, the question arises what the consequence

[3]If, for example, the threshold for an agent is $\beta = \frac{1}{2}$, there would be two intermediate proof steps that have undefined truth values to make a subtle impossible worlds able to exist for that agent. For an threshold of $\frac{1}{x}$, there are x intermediate proof steps with an undefined truth value.

of such undefined values of validations is on the logical framework and on the epistemic interpretation? The following subsection is a brief discussion on this matter. After that, in section 4.3.2, an example of the modal space is given to get a better idea of the addition of the inconsistency values.

4.3.1 Consequences of partial worlds

First, how can a sentence with an undefined truth value be epistemically interpreted? In a real life example, it is presumably easy to imagine a partial detail of an epistemic possibility. If you hear a story from somebody, many facts are partial, as it neither known that they are true nor false [6, p. 225]. Remember the example in section 2.1 of Lisa, who lost her bike and thought about the scenarios that could have preceeded the incident. In this example, it was not told if there was a security camera hanging in the shed or not. Besides, if the valuation of this useful fact had been part of the given information, the story would certainly not tell, for example, what Lisa ate for breakfast that morning; every scenario is partial in what it explicitly represents [6, p. 225]. Hence, a framework with partial worlds is perfectly acceptable for a representation of human-like knowledge.

However, for the consequences on the framework of logic, it is different. The issue of partial worlds is totally different from the issue of inconsistency. Combining these two concepts is a discussion worthy of a whole new study. Questions that are fundamental to explore are, for example, if a certain sentence A is neither true nor false at a world, which truth value is then provided? Is there a third truth-value introduced or is it more like an absence of a truth value? A reasonable approach to tackle this issue could be the option of the absence of a truth value because this is most in line with the epistemic interpretation discussed above [4].

4.3.2 Example of a modal space with inconsistency values

Imagine a model $\mathcal{M}_2$ (see figure 4.1) in which we have possible worlds w_0 and w_1 and impossible worlds w_2 and w_3. In this model, we want to represent an agent d that is able to know the blatant tautology ρ but not the subtle tautology ξ, just as in the example of Dylan in section 2.4. ρ' is a sentence from which in 1 step $\neg\rho$, a false tautology, can be inferred. ξ' is a sentence from which in 6 steps $\neg\xi$, a false tautology, can be inferred. In world w_0 and w_2 both tautologies (ρ and ξ) are, by definition, true. In impossible world w_2, the subtle tautology ξ does not have a truth valuation, but ξ' and ρ are true. In impossible world w_3, the blatant tautology ρ does not have a truth valuation but ρ' and ξ are true. The inconsistency values of the worlds w_0, w_1, w_2,and w_3 are respectively $0, 0, \frac{1}{6}$, and 1. For agent d the threshold $\beta_d = 1$. Consider the following model $\mathcal{M}_2$ for a formal representation:

[4] For a more elaborated discussion, [15] could be consulted.

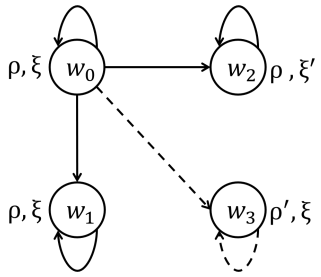

Figure 4.1: Model $\mathcal{M}_2$ with inconsistency values where only the straight lines represent the accessibility relations of agent d.

- $\mathcal{M}_2 = \langle W, R, V, I \rangle$, where
- $W = \{I, P\}$, with $P = \{w_0, w_1\}$ and $I = \{w_2, w_3\}$,
- $R_d = \{(w_0w_0), (w_0w_1), (w_0w_2), (w_1w_1), (w_2w_2)\}$,
- $V(w_0) = \{\rho, \xi\}$, $V(w_1) = \{\rho, \xi\}$, $V(w_2) = \{\rho, \xi'\}$, $V(w_3) = \{\rho', \xi\}$
- $\mathcal{I}_d(w_0) = 0$, $\mathcal{I}_d(w_1) = 0$, $\mathcal{I}_d(w_2) = \frac{1}{6}$, $\mathcal{I}_d(w_3) = 1$

In this model, the knowledge of agent d is assessed from w_0. The blatant tautology ρ is true in every world, except in world w_3 where the inconsistency value is 1. However, according to the definition of knowledge this world does not affect the knowledge of agent d in world w_0 as w_3 is not accessible from w_0, so $w_0 \vDash K_d\rho$.

In w_0 there is no knowledge of the subtle tautology ξ, because not for every accessible world u with an inconsistency value below the threshold it is the case that $u \vDash \xi$. Namely, in w_2 the valuation of this tautology ξ is undefined, so $w_0 \vDash \neg K_d \neg \xi$.

In other words, agent d knows that ρ but does not know that ξ, the desired outcome.

Chapter 5

Conclusion

This thesis has described the problem of logical omniscience in the framework of *possible worlds*. The aim was to examine the addition of different specification of *impossible worlds* to decide which should be used to represent human knowledge and to avoid logical omniscience. The investigation of Jago's approach of open worlds has shown the effects of not limiting impossible worlds to any rules. Although logical omniscience was completely avoided, the result was an modal space which lacked any interesting logical properties. Nonetheless, this approach contributed to our understanding of a modal space without logical rules. Thereby, Jago concluded that a more promising approach for impossible worlds is one in which obvious inconsistencies are not epistemically accessible.

In line with this thought, Bjerring's approach was discussed. He made a formerly defined distinction between blatant impossible and subtle impossible worlds. This idea refined the notion of which impossibilities should be included in the modal space. His conclusion was, however, that such a modal space could not be established without the allowance of partial worlds. Hence, we proposed such a framework in chapter four, with the allowance of partial worlds. The resulting model combines the proposal of both Jago and Bjerring and uses the formerly defined distinguishment between blatant and subtle, without collapsing in modeling either an omniscience agent or an agent that is a too unintelligent.

This model, which includes *inconsistency values*, has provided a deeper insight into the properties of impossible worlds that are most representative for modeling a human-like agent. This suggests that an improved knowledge representation for a human-like agent lies in the line of permitting partial worlds with their subtle inconsistencies.

Although the analysis of this model is not complete, it has extended to an idea for a new approach to representing human-like agents. A natural progression of this work is to further analyze the epistemic interpretation of the inconsistency values and the allowance of partial worlds in the framework. The relation between the inconsistency value and the degree of belief was briefly touched upon in chapter 4. However, further research is required to examine this link more closely.

Taken together, these findings contribute to research that seeks to optimize modal space for the representation of human-like agents. Such a human-like modal space could be implemented in the area of Artificial Intelligence which is concerned with the creation of intelligent systems that reason like humans do, for example in healthcare, where one wants to develop robots that are intended to imitate human contact and behavior.

Bibliography

[1] Mark Jago. "Hintikka and Cresswell on logical omniscience". In: *Logic and Logical Philosophy* 15.4 (2006), pp. 325–354.

[2] Jens Christian Bjerring. "Impossible worlds and logical omniscience: an impossibility result". In: *Synthese* 190.13 (2013), pp. 2505–2524.

[3] Rosja Mastop. *Modal Logic for Artificial Intelligence*. 2012.

[4] Ronald Fagin et al. *Reasoning about knowledge*. MIT press, 2003.

[5] Fausto Giunchiglia et al. "Non-omniscient belief as context-based reasoning". In: *IJCAI*. Vol. 93. 1993, pp. 9206–03.

[6] Francesco Berto and Mark Jago. *Impossible worlds*. Oxford University Press, 2019.

[7] Hector-Neri Castañeda. "Jaakko Hintikka. Knowledge and belief. An introduction to the logic of the two notions. Cornell University Press, Ithaca, N.Y., 1962, x 179 pp." In: *Journal of Symbolic Logic* 29.3 (1964), pp. 132–134. DOI: 10.2307/2271621.

[8] Robert Bull and Krister Segerberg. "Basic Modal Logic". In: *Handbook of Philosophical Logic*. Ed. by D. M. Gabbay and F. Guenthner. Dordrecht: Springer Netherlands, 2001, pp. 1–81. ISBN: 978-94-017-0454-0. DOI: 10.1007/978-94-017-0454-0_1. URL: https://doi.org/10.1007/978-94-017-0454-0_1.

[9] Patrick Blackburn and Johan Van Benthem. "1 Modal logic: a semantic perspective". In: *Studies in Logic and Practical Reasoning*. Vol. 3. Elsevier, 2007, pp. 1–84.

[10] Francesco Berto and Mark Jago. "Impossible Worlds". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2018. Metaphysics Research Lab, Stanford University, 2018.

[11] Rasmus Rendsvig and John Symons. "Epistemic Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2019. Metaphysics Research Lab, Stanford University, 2019.

[12] Max J Cresswell. "Logics and Languages, Methuen and Co". In: *Ltd., London* (1973).

[13] Hector J Levesque. "A logic of implicit and explicit belief". In: *AAAI*. 1984, pp. 198–202.

[14] Graham Priest. *Towards non-being: The logic and metaphysics of intentionality*. Oxford University Press, 2016.

[15] Jens Christian Bjerring et al. "Non-ideal epistemic spaces". In: (2010).

[16] Becky Lytle. "Introduction to the convergence of sequences". In: *University of Chicago* (2015).

[17] Franz Huber. "Formal Representations of Belief". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2016. Metaphysics Research Lab, Stanford University, 2016.