You Don't Believe This Is The Title

Moore's Paradox and its relation to the Surprise Exam Paradox, the Knowability Paradox, the Toxin Problem and Newcomb's Problem

MSc Thesis *(Afstudeerscriptie)*

written by

Max van den Broek
(born January 29th, 1995 in Hoorn, the Netherlands)

under the supervision of **Prof. Dr. S.J.L. Smets** and **Dr. F.R. Velázquez Quesada**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:
June 11th 2018

Members of the Thesis Committee:
Prof. Dr. S.J.L. Smets
Prof. Dr. Y. Venema
Dr. F.R. Velázquez Quesada
Dr. A. Özgün
Dr. K. Schulz
Dr. M.D. Aloni

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

"A logical theory may be tested by its capacity for dealing with puzzles, and it is a wholesome plan, in thinking about logic, to stock the mind with as many puzzles as possible, since these serve much the same purpose as is served by experiments in physical science."

Bertrand Russell

UNIVERSITY OF AMSTERDAM

Abstract

Faculty of Science
Institute for Logic, Language and Computation

Master of Logic

You Don't Believe This Is The Title
Moore's Paradox and its relation to the Surprise Exam Paradox, the Knowability Paradox, the Toxin Problem and Newcomb's Problem

by Max VAN DEN BROEK

This thesis concerns Moore's paradox and its relations to other paradoxes and problems. In particular, it concerns the relations between Moore's paradox and the surprise exam paradox, the knowability paradox, the Toxin problem, Newcomb's problem and multiple problems that are formulated for the first time in this thesis. The main claim defended in this thesis is that these problems are similar insofar as they all involve *Moore sentences*.

To defend this claim we develop a formal account of Moore sentences. Briefly, we state that a sentence ϕ is a Moore sentence for an agent A if (i) ϕ contains a modal operator $\Box_A$, (ii) ϕ is satisfiable with respect to a class of models C suitable for modeling $\Box_A$ and (iii) $\Box_A\phi$ is unsatisfiable with respect to C. This definition has multiple upshots compared to previous definitions. Among these are that it adequately captures the intuition that doxastic Moore sentences can be true but cannot be believed. Further, the definition is formulated so that it is not limited to doxastic Moore sentences but applies to Moore sentences concerning other propositional attitudes as well.

Using our formal account of Moore sentences we discuss under what assumptions it can be said that a Moore sentence is involved in the surprise exam paradox, the knowability paradox, the Toxin problem and Newcomb's problem. This discussion produces several interesting results. We resolve a debate about which sentence, precisely, is a Moore sentence in the surprise exam paradox. We also reconsider the consequences of the knowability paradox for the anti-realist verification thesis. Further, we introduce new variations of the Toxin problem and Newcomb's problem, as well as two paradoxes concerning intention and desire that are similar to the knowability paradox, and suggest how all these problems may be solved.

Acknowledgements

I would like to express my gratitude to those who helped to put this thesis together. First and foremost I would like to thank my supervisors Sonja Smets and Fernando Velázques Quesada. It is rare that I meet people who are as enthusiastic as I am about paradoxes, and Sonja and Fernando matched my enthusiasm for paradoxes only when they did not exceed it. Our discussions were always thought-provoking, insightful and enjoyable. I also thank Sonja and Fernando for keeping my writing sharp, especially regarding the technical parts.

Secondly I thanks all the other Master of Logic students, for helping me out with the thesis and for my experiences in the Master of Logic in general. I certainly never came across a group of people with a higher median IQ or more willingness to discuss complicated issues concerning philosophy, mathematics or language. I should thank several students for acquainting me with topics I knew next to nothing about: Haukur and Valentin for machine learning algorithms, Luca, David and Thijs for complexity theory and automata, and Jonathan and Morwenna for computational models of language and cognition. I should also thank Silvan, David and Mina, for providing feedback on draft versions of the thesis.

Finally, I thank my parents and my brother for my upbringing, which is at the basis of any and all my accomplishments. And I thank my girlfriend, who provided emotional support as well as feedback on draft versions, at the times I needed it most.

Contents

Dedicated to those who are puzzled but believe they are not

Chapter 1

Introduction

This thesis concerns Moore's paradox and its relation to other philosophical paradoxes and problems.[1] In particular, it concerns the relation between Moore's paradox and the surprise exam paradox, the knowability paradox, the Toxin problem and Newcomb's problem. Additionally, it concerns four newly formulated problems concerning intention and desire, which we call Potion problem, the omissive Newcomb problem, the intentionability paradox and the desirability paradox. The focus of the discussion in this thesis rests not on the individual problems and paradoxes themselves, but on the similarities and dissimilarities between them. The claim we defend in this thesis, then, is that all the above mentioned paradoxes, problems and puzzles are similar insofar as they all involve *Moore sentences*.

To defend this claim we develop an account of Moore sentences which we can loosely characterize as sentences that function akin to the famous Moore sentence 'It is raining but I do not believe it'.[2] This sentence is famous for appearing to be contradictory, even though it describes a perfectly possible situation.[3] On our account, this sentence is peculiar because it can be true but cannot be believed (what we mean exactly when we say that a sentence 'can be true' or 'cannot be believed' is explained in Chapter 3). Our account continues a tradition started by Hintikka, who Hintikka defined Moore sentences as sentences that can be true but cannot be believed. More specifically, Hintikka defines a Moore sentence to be a sentence ϕ that is satisfiable in a KD4 modal logic, while the sentence that describe that an agent believes ϕ is unsatisfiable in the same logic.[4] The account of Moore sentences as possibly true but unbelievable sentences was further developed largely by Sorensen who also described Moore sentences as *blindspots*: sentences that, despite being true, cannot be believed.[5] Sorensen formulated his definition in natural language and gave several objections to Hintikka's use of modal logic to characterize Moore sentences. After Sorensen's contribution, natural language accounts of why Moore sentences can be true but cannot be believed, became quite popular.[6]

[1] In this thesis we are not interested in the distinctions between paradoxes and problems and simply follow established nomenclature when describing either. Wittgenstein coined the term 'Moore's paradox' (Wittgenstein, Anscombe, and Wittgenstein, 1963, 87), but some authors shy away from calling the problem a paradox and label it *Moore's problem* instead (e.g. Sorensen, 1988. We arbitrarily choose to follow Wittgenstein's terminology.

[2] Green and Williams, 2007.

[3] That the sentence describes a possible situation becomes immediately apparent if we consider the third person version of the sentence: 'It is raining but *he* doesn't believe it is raining'. If 'he' refers to the same person as 'I' did in the first person variation of the sentence, then both sentences describe the same situation, and the third person variant of the sentence clearly describes a possible situation.

[4] Although we discuss some concerns about they way he formalized this claim in Chapter 2.

[5] Sorensen, 1988.

[6] For an overview of such accounts, see Williams, 2015b.

Our account of Moore sentences continues the tradition of Hintikka, Sorensen and their successors, in that we also treat the famous Moore sentence mentioned above as being possibly true yet unbelievable. But, our account differs from most accounts in this tradition in two ways. Firstly, because we bring back the formal approach advocated by Hintikka, while most authors in this tradition since Hintikka develop their accounts in natural language.[7] We favor a formal approach because the precision of formal languages lends itself to a clear and rigorous analysis of Moore sentences. In Chapter 3, we address Sorensen's objections to Hintikka's formal approach, as most of his objections also apply to our approach.

The second way in which our account differs from most other accounts in the tradition of Hintikka, Sorensen and successors, is that we consider *non-doxastic* Moore sentences as well.[8] Although most of the discussion about Moore sentences has historically centered around doxastic Moore sentences, it has recently been argued that there also exist Moore sentences concerning desires.[9] The idea is that, just like there are sentences that can be true but cannot be believed to be true, there are sentences that can be true but that cannot be desired to be true. And, for reasons we explain in Chapters 2 and 5, we believe there are also Moore sentences concerning intentions, that is to say, sentences that can be true but that one cannot intend to make true. To be as general as possible, we would thus like to understand Moore sentences as sentences that can be true, but that one cannot have a specific propositional attitude towards.[10] We make this account clear and precise in Chapter 3.

Equipped with the account of Moore sentences we sketched above, we argue that Moore sentences are involved in the surprise exam paradox, the knowability paradox, the Toxin problem and Newcomb's problem. Some of our arguments will be familiar to those who are acquainted with the literature on philosophical paradoxes. In particular, the role of Moore sentences in the surprise exam paradox and the knowability paradox have been studied to a considerable extent in the literature.[11] Nevertheless, we aim to make some interesting novel contributions to the discussions of these paradoxes. Firstly, we contribute to a long-standing debate about which sentence, precisely, is a Moore sentence in the surprise exam paradox.[12] Further, we reconsider the consequences of the knowability paradox for the anti-realist verification thesis that all truths can be known. We distinguish two interpretations of the thesis corresponding to two phases in the learning process of agents and argue that, under one interpretation of the thesis the knowability paradox is an argument against it, but under another (and more important) interpretation, it is not.

We also discuss the role of Moore sentences in the Toxin problem and Newcomb's

[7] Green and Williams, 2007, Williams, 2015a, Williams, 2015b, Almeida, 2001.

[8] $\delta o \chi \alpha$ being the ancient Greek word for '(common) belief', doxastic sentences are sentences that concern belief, such as 'I believe it is raining'.

[9] Wall, 2012, Williams, 2014, Cholbi, 2009.

[10] Propositional attitudes are attitudes that one may have towards propositions, for example one can doubt a proposition, one can intend to make a proposition true, one can hope for a proposition to be true, and so forth. In this thesis we mainly work with sentences concerning the propositional attitudes belief, knowledge, desire and intention, but the definition of Moore sentences we give in Chapter 3 can be used to discuss sentences concerning any propositional attitude.

[11] See e.g. Sorensen, 1988, in particular Chapters 7, 8 and 9.

[12] Quine, 1953, Binkley, 1968.

problem. The role of Moore sentences in these problems has not been studied as extensively as their roles in the surprise exam paradox and the knowability paradox.[13] In part, this is because the Toxin problem and Newcomb's problem concern intentions and desires respectively, so for a full appreciation of their relations to Moore sentences it is necessary to have an account of Moore sentences that is not limited to doxastic sentences. We introduce new variations of both the Toxin problem and Newcomb's problem which we refer to as the Potion problem and the omissive Newcomb problem. We argue that these new problems pose novel challenges to scholars studying desire and intention. Further, we introduce two new paradoxes concerning desire and intention that are reminiscent of the knowability paradox. We call these paradoxes the intentionability paradox and the desirability paradox. The results of the intentionability paradox is that not every sentence that can be true is such that agents can intend to make it true. The result of the desirability paradox is that not every sentence that can be true can be desired to be true. We also propose a way to understand both of these new paradoxes not as problems, but as constructive results about the role of Moore sentences concerning intention and desire in human cooperation. Finally, we use the previously mentioned distinction between different phases in an agent's learning process to address the temporal aspects of the Toxin problem and Newcomb's problem. These temporal aspects illustrate how the Toxin problem and Newcomb's problem are similar and also how they differ.

We hope the upshots of our project to be manifold. In addition to some of the specific results mentioned above, we hope to contribute to the philosophical literature in a number of different ways. By clarifying the connections between different paradoxes in terms of their similarities and dissimilarities, this project opens up many opportunities for future research. First, if it is discovered that two problems A and B are similar and different variations are known of a problem A but not of problem B, this can lead to the discovery of new variations of problem B. Russell stressed the importance of discovering new problems and puzzles:

> "A logical theory may be tested by its capacity for dealing with puzzles, and it is a wholesome plan, in thinking about logic, to stock the mind with as many puzzles as possible, since these serve much the same purpose as is served by experiments in physical science."[14]

We agree with Russell and take it that discovering new variations of known problems can be as useful for testing our theories as discovering completely new puzzles. Our introduction of the omissive Toxin problem, the omissive Newcomb problem, the intentionability paradox and the desirability paradox are all examples of newly found variations of known problems, and we hope that the connections that we describe in this thesis lead to more discoveries of this kind.

Second, if it is discovered that two problems A and B are similar and a solution has been proposed for problem A, this makes it possible to apply this solution (or a similar solution) to problem B. Our application of van Benthem's solution to the knowability paradox to both the intentionability paradox and the desirability paradox is an example of such a line of research, as well as the distinction between different phases in a learning process that we introduce in our discussion of the knowability paradox and are able to use again to address the temporal aspects of

[13] But, see Goldstein and Cave, 2008, for a recent discussion on this topic.

[14] Russell, 1905, 484-485.

the Toxin problem and Newcomb's problem. And again, we hope that the connections between different problems as described in this thesis prompt more research of this kind.

Third, our discussion of the similarities and dissimilarities between problems concerning different topics can serve as a vehicle for studying the similarities and dissimilarities between these topics themselves. For example, we discuss the similarities between the surprise exam paradox, the Toxin problem and Newcomb's problem, which concern knowledge, intention and desire respectively.[15] This discussion prompts us to discuss the formal properties of knowledge, intention and desire, which renders them comparable.

This thesis thus describes interesting results about the connections between different paradoxes as well as a wealth of possibilities for future research. The thesis is structured as follows. In Chapter 2 we introduce Moore's paradox and elaborate on its origin and rich history. This is to give the reader a preliminary understanding of the wide range of discussions that Moore's paradox, and the underlying Moore sentences, has lead to. In Chapter 3, we formally define Moore sentences, and discuss for various syntactic and semantic variations that have been called Moore sentences in the literature under which assumptions these sentences can indeed be said to be Moore sentences. In Chapter 4 we discuss two epistemic paradoxes, the surprise exam paradox and the knowability paradox, together with the role Moore sentences play in them. In Chapter 5 we outline the role of Moore sentences in two non-epistemic problems: the Toxin problem and Newcomb's problem. In this chapter we also introduce the Potion problem, the omissive Newcomb problem, the intentionability paradox, the desirability paradox, and our proposal to deal with these last two problems. In Chapter 6 we provide a brief conclusion to the thesis.

[15] Goldstein and Cave, 2008.

Chapter 2

Introducing Moore's Paradox: a Historical Analysis

"I went to the pictures last Tuesday, but I do not believe that I did" is the sentence that Moore wrote down to puzzle generations of subsequent philosophers.[1] The puzzle of this sentence, as Moore describes it, is that it is a "perfectly absurd thing to say, although *what* is asserted is something which is perfectly possible; it is perfectly possible that you did go to the pictures and yet you do not believe that you did".[2] The challenge to explain why some sentences that describe possible situations are still absurd to assert, has come to be known as *Moore's paradox*.[3]

Moore's paradox has a rich history, and understanding this history is important for appreciating the main claim of this thesis. This claim is that Moore sentences not only play a central role in Moore's paradox, but also in the surprise exam paradox, the knowability paradox, the Toxin problem and Newcomb's problem. To support this claim we develop an understanding of Moore sentences that is not limited to doxastic sentences, but extends to sentences concerning other propositional attitudes, in particular desire and intention. Although such a general understanding of Moore sentences has not been developed systematically in previous literature, examples of non-doxastic Moore sentences can be found throughout discussions of Moore sentences and can even be traced to the conception of Moore's paradox. To root our account of Moore sentences in the previous literature on Moore's paradox and Moore sentences, we present a historic overview of this literature in this section. We do not discuss all literature concerning Moore sentences, but we sketch how the discussions in the literature developed and we note every mention of a non-doxastic Moore sentence.

The early discussion of Moore's paradox is lead by philosophers of language and pragmatists, such as Moore himself, Martinich, and Levinson.[4] They try to explain, through means such as *conversational implicature*, that assertions like 'I went to the pictures last Tuesday, but I do not believe that I did' are not an *assertion* of a contradiction, but nevertheless *imply* a contradiction (in some informal sense of 'imply' that may not correspond to material implication).

Epistemologists also engage in the discussion about Moore's paradox, because

[1] Moore, 1942, 543.

[2] Moore, 1942, 543.

[3] Wittgenstein, Anscombe, and Wittgenstein, 1963, 87.

[4] Moore, 1942, 543-545, Martinich, 1980, Levinson, 1983.

they believe that it is not merely absurd to *assert* 'I went to the pictures last Tuesday, but I do not believe that I did', it is equally absurd to *believe* this sentence.[5] In fact, Moore's paradox is discussed more in epistemology than in philosophy of language, because most philosophers accept *Shoemaker's principle*.[6] Schoemaker's principle states that the absurdity of *asserting* Moore sentences can be reduced to the absurdity of *believing* it, whereas the converse is not the case. The motivation for this principle is the idea that asserting a sentence implies believing is, but believing a sentence does not imply asserting it.

It is quite remarkable that Moore's paradox turns out to be of prime interest to epistemologists, considering that Moore does not introduce it to make a point about what humans can and cannot believe; he introduces it to advance an argument in ethics.[7] In the section where Moore introduces his famous Moore sentence, Moore is directly replying to Stevenson's *Arguments against ethical naturalism*.[8] Moore's ethical doctrine is naturalist, which is to say that Moore believes ethical sentences such as 'Brutus was right to stab Caesar' do not merely refer to the feelings or attitudes of the person who utters it, but refer to objective moral facts. Stevenson argues against Moore's doctrine by advancing several arguments for the claim that there is no difference in meaning between the sentence 'Brutus was right to stab Caesar' and the sentence 'I approve of Brutus' stabbing of Caesar'. Moore rebuts Stevenson's claim by pointing out that although a man who asserts that 'Brutus was right to stab Caesar' may be said to *imply* (in some informal sense of 'imply' that need not correspond to material implication) the sentence 'I approve of Brutus' stabbing of Caesar', but he cannot be said to *assert* it. To support this argument, Moore draws the following analogy:

> "your saying that you did [go to the pictures last Tuesday] does *imply* [in some informal sense], that you believe that you did; and this is why 'I went, but I do not believe I did' is an absurd thing to say. Similarly, the fact that, if you assert it was right of Brutus to stab Caesar, you *imply* that you approve of or have some such attitude this action of Brutus (...) Hence if we hear a man assert that the action was right, we should all take it that he does approve, although he has *not* asserted that he does."[9]

Moore's argument is thus that because 'I went to the pictures last Tuesday but I do not believe I did' is not a contradiction, the two conjuncts of the sentence must have different meanings. That is, there must be a difference in the meanings of the sentences 'I went to the pictures last Tuesday' and 'I believe that I went to the pictures last Tuesday'. Moore stresses that although the former sentence *implies* the latter (in some informal sense of 'imply'), it is not strictly true that if a man asserts the former sentence, he also asserts the latter. This observation is meant to illustrate that there is also a difference between the ethical sentences 'Brutus' stabbing was right' and 'I approve of Brutus' stabbing', which are the sentences that Moore was arguing about with Stevenson. The difference between 'Brutus' stabbing was right' and 'I approve of Brutus' stabbing' is important to Moore because it supports his

[5] Hintikka, 1962 is credited by Williams, 2015b for being the first to notice this.

[6] Green and Williams, 2007, Schoemaker's principle principle is introduced in Shoemaker, 1996, 76-77.

[7] Moore, 1942, 543.

[8] Stevenson, 1942.

[9] Moore, 1942, 543.

ethical doctrine, that ethical sentences such as 'Brutus was right to stab Caesar' cannot be (fully) reduced to sentences that refer to human attitudes such as 'I approve of Brutus' stabbing of Caesar'. It is an interesting artifact of history that although the sentence 'I went to the pictures last Tuesday but I do not believe I did' initially was only interesting to Moore insofar as it is analogous to the sentence 'Brutus stabbing was right but I do not approve of it', the former sentence made it into the philosophical canon while the latter sentence has received hardly any attention by scholars since Moore.

That said, ethical Moore sentences have not escaped attention of philosophers completely. Gombay argues that there are *Socratic Moore sentences*, of the forms "A is bad, but I now will to do A" and "A is good, but I do not now will to do A".[10] Gombay argues that according to the Socratic doctrine these sentences are absurd to assert because this doctrine teaches that "No one is evil knowingly".[11]

The claim that no one does evil knowingly needs some further elaboration, as it seems that it's not even rare for people to do evil knowingly in the real world. Santas explains why Socrates endorses this claim nonetheless in terms of two assumptions underlying his doctrine.[12] First, Socrates assumes that every action that an agent performs is performed not for its own sake, but for the sake of achieving something that the agent consider good for himself. So for every action an agent performs, he performs that action because he believes he will be better off having performed that action. Second, Socrates assumes that performing 'right' of 'just' actions always leads to achieving good outcomes for an agent, while performing evil actions always results in bad outcomes.[13] From these two assumptions it follows that no one does evil knowingly, because evil leads to bad outcomes for the agent and no agent knowingly tries to achieve bad outcomes for himself. Of course, whether Socrates' assumptions are plausible is debatable, but we will not engage in this debate at this moment.

More recently, Cholbi also brings the discussion of Moore sentences back to its ethical roots. Cholbi argues that what he calls traditional (epistemological) Moore sentences, such as 'I went to the pictures last Tuesday, but I do not believe that I did' and 'It's raining but I do not believe it', have moral equivalents that are equally paradoxical, such as 'hurting animals is wrong, but I do not care'.[14] Cholbi writes that these moral equivalents of Moore sentences are similar to epistemological Moore sentences in that they describe morally possible situations that are nevertheless absurd to assert.

Recently, Williams and Wall have independently advanced arguments that sentences concerning desires can be paradoxical in similar ways to epistemological Moore sentences.[15] As examples of Moore sentences concerning desires, Williams gives 'I am drinking stout but do not desire to do so'. But according to Williams,

[10] Gombay, 1988, 194.

[11] Gombay, 1988, 194.

[12] Santas, 1964, 158.

[13] These Platonic doctrines are argued for in the *Gorgias* and *the Republic*, for English translations of these books see Irwin, 1979 and Bloom and Kirsch, 2016.

[14] Cholbi, 2009.

[15] Wall, 2012, Williams, 2014.

the similarity between Moore sentences concerning belief and Moore sentences concerning desire is different from the similarity between Moore sentences concerning beliefs and Moore sentences concerning the moral attitudes that Cholbi describes. Moore sentences concerning beliefs and moral attitudes are similar in that they both describe possible situations but are nevertheless *absurd to believe*. Moore sentences concerning desires are similar to Moore sentences concerning belief in the sense that, just like Moore sentences concerning belief are absurd to believe, Moore sentences concerning desire are *absurd to desire*. Thus, the point is not that it is absurd to believe sentences like 'I am drinking stout but do not desire to do so', but it is absurd to desire that these sentences are true. Since, according to both Wall and Williams, desiring to do something while you do not desire to do it, violates some plausible norms about desire formation.

A historical analysis of Moore's paradox shows the breadth of its scope, with roots in ethics, a solid base and and bark in philosophy of language and epistemology, and branches in the fields studying non-doxastic propositional attitudes such as desire. This validates our attempt to define Moore sentences in a way that allows for Moore sentences involving various kinds of propositional attitudes, not just beliefs. And, hopefully, our attempt is vindicated by its usefulness for the research towards the relations between Moore's paradox and other non-doxastic paradoxes, which we present in this thesis.

Chapter 3

Defining Moore Sentences

In this chapter we define Moore sentences formally. To do so we first introduce our formal language. Then we evaluate a past attempt to define Moore sentences formally by Hintikka. We consider Sorensen's objections to his definition and raise some constructive criticism of our own, to arrive at an improved version of his definition. Then we assess for various sentences whether our definition captures the intuitive judgments about under which assumptions these sentences can be said to be Moore sentences, as these intuitions are formulated in the literature. For this assessment we consider well known Moore sentences as well as syntactic and semantic variations thereof.

3.1 Formal Language and Semantics

We define our formal language $\mathcal{L}$ as follows:

Definition 3.1. (Formal language $\mathcal{L}$)

$$\phi ::= p \mid \neg\phi \mid \phi \vee \psi \mid \Box_A\phi$$

where p are elements of a countable set of propositions P. The other boolean connectives ($\wedge, \rightarrow, \leftrightarrow$) are defined as usual.

The reading of the modal operator $\Box_A$ varies throughout this thesis depending on the problem we are discussing. We will make clear which reading we are referring to by using familiar signs for modal operators. We make use of the following signs and readings in this thesis: $\Box\phi$ is read as it is metaphysically necessary that ϕ (since metaphysical necessity is not relative to a particular agent, we omit the index of the modal operator in these cases); $\mathrm{B}_A\,\phi$ is read as agent A believes that ϕ; $\mathrm{K}_A\,\phi$ is read as agent A knows that ϕ; $\mathrm{D}_A\,\phi$ is read as agent A desires that ϕ is true; $\mathrm{I}_A\,\phi$ is read as agent A intends to make ϕ true. In the single agent setting we will omit agent indexes and simply write statements such as $\mathrm{B}\,\phi$ to indicate that some implicit agent believes ϕ. And also in the multi-agent setting we may omit the index of a belief operator if it is clear from the context with which agent it is associated.

For our semantics we work with Kripke models in which the domain is interpreted as a set of possible worlds, the accessibility relations represents the access of agents and the atomic valuation describes what propositions are true in which possible world. Let W denote a finite non-empty set of possible worlds and P denote a countable set of propositions.

Definition 3.2. (Kripke model M) A Kripke model is a tuple $M = \langle \mathbf{W}, R_A, V\rangle$, where $R_A \subseteq \mathbf{W} \times \mathbf{W}$ is an accessibility relation (wR_Av indicates that world v is accessible from world w for agent A). The valuation function $V : \mathbf{W} \rightarrow \wp(\mathrm{P})$ maps every world w to the subset of propositions in P that are true in w ($p \in V(w)$ indicates that proposition p is true in world w).

The semantic interpretation of $\mathcal{L}$-formulas is given by our definition of the truth relation $\models$.

Definition 3.3. (Truth relation $\models$)
- $M, w \models p$ iff $p \in V(w)$
- $M, w \models \neg\phi$ iff $M, w \not\models \phi$
- $M, w \models \phi \vee \psi$ iff $M, w \models \phi$ or $M, w \models \psi$
- $M, w \models \Box_A\phi$ iff for all worlds v s.t. wR_Av, $M, v \models \phi$

Again, in this thesis we make use of different readings of $\Box_A\phi$. We indicate the intended reading by using the operators $\Box$, B_A, K_A, D_A and I_A, as mentioned above. To be clear, the semantics of $\Box\phi$, $\mathrm{B}_A\,\phi$, $\mathrm{K}_A\,\phi$, $\mathrm{D}_A\,\phi$ and $\mathrm{I}_A\,\phi$ are all identical to that of $\Box_A\phi$ given above.

Finally, we should introduce two semantic notions we use in our definition of Moore sentences: *satisfiability* and *unsatisfiability*. Intuitively, a sentence is satisfiable with respect to a class of models iff the sentence is true in at least one world in one of the models in that class. And a sentence is unsatisfiable with respect to a class of models just in case there is no model that contains a world in that class in which the sentence is true. Formally, satisfiability and unsatisfiability are defined as follows:

Definition 3.4. (Satisfiable) A sentence ϕ is satisfiable with respect to a set of models C iff in the domain W of some model $M \in C$ there is a world $w \in W$ such that $M, w \models \phi$.

Definition 3.5. (Unsatisfiable) A sentence ϕ is satisfiable with respect to a set of models C iff ϕ is not satisfiable with respect to C.

3.2 Hintikka's Definition

Despite many attempts to characterize what exactly is absurd about saying or believing a Moore sentence,[1] few authors bother to give an exact definition of Moore sentences. The majority of the authors are content to describe them as 'sentences like 'I believe it is raining but it is not''. For example, Shoemaker describes Moore sentences as follows:

> "Conjunctive sentences such as Max Black's 'Oysters are edible but I do not believe it,' or the more usual 'It is raining but I do not believe that it is raining,' I will call 'Moore-paradoxical sentences."'[2]

Similarly, in *Moore's paradox: new essays*, Green and Williams describe Moore's paradox and the corresponding Moorean absurdity as:

> "What is absurd [...] is to utter [I went to the pictures last Tuesday but I do not believe that I did] assertively. What is paradoxical is that there should be such an absurdity that cannot be fully explained in terms of a semantic contradiction generated by the words themselves."[3]

[1] See e.g. Heal, 1994, Collins, 1996, DeRose, 1991.

[2] Shoemaker, 1995.

[3] Green and Williams, 2007.

Among the select group of authors that defines Moore sentences,[4] Hintikka is the most interesting for the work in this thesis, as he formulates his definition in formal terms. Hintikka equates Moore sentences with *doxastically indefensible* sets of sentences. A set of sentences $\{\phi_1, ..., \phi_n\}$ is doxastically indefensible for an agent A iff $\mathrm{B}_A(\phi_1 \wedge ... \wedge \phi_n)$ is indefensible, where $\mathrm{B}_A\,\phi$ reads as 'agent A believes that ϕ'. *Indefensibility* in Hintikka's sense is synonymous with *inconsistency* or *unsatisfiability* on a Kripke model that is characterized by seriality and transitivity.[5,6] (Hintikka chooses the term 'indefensibility' because he finds the term more natural that the term 'unsatisfiability' in a context of beliefs.[7])

Hintikka's definition is precise and captures some of our basic intuitions about the nature of Moore sentences. Firstly, our intuition that Moore sentences are absurd to believe is captured by the fact that Moore sentences are doxastically indefensible. Secondly, Hintikka stresses that sentences can be doxastically indefensibility without being indefensible simpliciter.[8] This captures our intuition that Moore sentences are absurd to believe, but are not contradictions themselves.[9]

Nevertheless, several objections have been raised against Hintikka's definition, and in fact against the use of formal languages to define Moore sentences altogether.[10] In the next section we discuss and rebut these objections, before giving some constructive criticism of our own. Our own criticisms are not directed against the use of formal definitions of Moore sentences altogether but rather point to some imperfections in Hintikka's definition, which guide us in the following section, where we propose our own formal definition of Moore sentences.

3.2.1 Sorensen's Objections

In his treatise of Moore's paradox, Sorensen argues against using doxastic logic to characterize Moore sentences.[11] Many of his objections are directed specifically against Hintikka's definition of Moore sentences, but some range wider and aim to discredit any attempt to formally define Moore sentences. In this section we discuss and address his objections.

The Strength of Modal Logic

Sorensen's first objection to Hintikka's approach is that most philosophers agree that the modal logic Hintikka uses is too strong to characterize beliefs. Sorensen claims that Hintikka uses the modal logic S4, which is characterized by reflexivity and transitivity. But this objection is factually incorrect. Hintikka does not use the modal logic S4 to characterize beliefs. This is clear because the modal logic S4 is reflexive ($\mathrm{B}\,\phi \rightarrow \phi$) and Hintikka's logic is characterized only by seriality and transitivity.[12]

[4] Hintikka, 1962, Sorensen, 1988, Almeida, 2001.

[5] Hintikka, 1962, 31-32.

[6] Note that Hintikka's definition does not require that Moore sentences can be true; we come back to this point later in this section.

[7] Hintikka, 1962, 32.

[8] Hintikka, 1962, 72.

[9] Although this intuition is not captured completely because all sentences that are indefensible are also doxastically indefensible; more on this later.

[10] Sorensen, 1988, Chapter 1 section IIc.

[11] Sorensen, 1988, Chapter 1 section IIc.

[12] Hintikka, 1962, 47-51.

Hintikka even explicitly mentions that belief is not veridical, and thus is not defined with respect to a class of models that is reflexive:

> "The condition $\mathrm{K}\,\phi \rightarrow \phi$ [where $\mathrm{K}\,\phi$ is read as 'ϕ is known'] does not have a doxastive counterpart. (...) What $\mathrm{K}\,\phi \rightarrow \phi$ amounts to is that whatever is known should be true. There is no reason why what is believed should be true."[13]

Sorensen also claims that Hintikka makes use of some problematic rules concerning beliefs. Sorensen objects that:

> "in his proofs of the unbelievability of Moore sentences, Hintikka appeals to the rule that belief in p implies belief that one believes p, and the rule that belief collects over conjunction."[14]

The objection is thus that Sorensen appeals to positive introspection ($\mathrm{B}\,\phi \rightarrow \mathrm{B}\,\mathrm{B}\,\phi$) and collection of belief over conjunction ($(\mathrm{B}\,\phi \wedge \mathrm{B}\,\psi) \rightarrow \mathrm{B}(\phi \wedge \psi)$) in his proofs of the doxastic indefensibility of Moore sentences.[15]

However, this objection seems to contain a factual inaccuracy as well. Because Hintikka does not use belief collection over conjunction in his proof that Moore sentences are doxastically indefensible. Hintikka formally represents a Moore sentence as $p \wedge \neg\,\mathrm{B}\,p$ and proves that it is doxastically indefensible using only the consistency of beliefs ($\mathrm{B}\,\phi \rightarrow \neg\,\mathrm{B}\,\neg\phi$), distribution of belief over conjunction ($\mathrm{B}(\phi \wedge \psi) \rightarrow (\mathrm{B}\,\phi \wedge \mathrm{B}\,\psi)$) and positive introspection:[16]

(1) $\mathrm{B}(p \wedge \neg\,\mathrm{B}\,p)$ (supposition for reductio)
(2) $\mathrm{B}\,p \wedge \mathrm{B}\,\neg\,\mathrm{B}\,p$ (distribution of belief over conjunction, 1)
(3) $\mathrm{B}\,p$ (elimination of conjunction, 2)
(4) $\mathrm{B}\,\neg\,\mathrm{B}\,p$ (elimination of conjunction, 2)
(5) $\neg\,\mathrm{B}\,\neg\neg\,\mathrm{B}\,p$ (consistency of beliefs, 4)
(6) $\neg\,\mathrm{B}\,\mathrm{B}\,p$ (elimination of negation, 5)
(7) $\mathrm{B}\,\mathrm{B}\,p$ (positive introspection, 3)
(8) $\perp$ (introduction of conjunction, 6 and 7)

From the assumption that $\mathrm{B}(p \wedge \neg\,\mathrm{B}\,p)$ a contradiction is derived. This means that $\mathrm{B}(p \wedge \neg\,\mathrm{B}\,p)$ is unsatisfiable, or as Hintikka would say indefensible. Then, since any set of sentences $\{\phi_1, ..., \phi_n\}$ is doxastically indefensible for an agent A iff $\mathrm{B}_A(\phi_1 \wedge ... \wedge \phi_n)$ is indefensible, this implies that the singleton set containing only the sentence $p \wedge \neg\,\mathrm{B}\,p$ is doxastically indefensible.

Clearly, belief collection over conjunction is not used in this derivation. If Sorensen meant to claim that Hintikka explicitly refers to collection of belief over conjunction in his derivation of the doxastic indefensibility of $p \wedge \neg\,\mathrm{B}\,p$, this objection would thus be inaccurate.

[13] Hintikka, 1962, 48.
[14] Sorensen, 1988, 21.
[15] Van Ditmarsch et al., 2015.
[16] We have rewritten this proof in our own formal language. The original proof can be found in Hintikka, 1962, 62.

We should grant Sorensen however, that although Hintikka does not explicitly refers to collection of belief over conjunction in his derivation, in the semantic model he works with belief does collect over conjunction.[17] In this sense, he is thus committed to such properties of belief. We would like to stress though, that because Hintikka does not use explicitly refer to collection of belief over conjunction in his derivation of the doxastic indefensibility of Moore sentences, he could make the same derivation in a logic that validates all rules that he refer to explicitly and not collection of beliefs over conjunction.

In any case, Sorensen's objection that Hintikka makes use of positive introspection still stands, as Hintikka makes use of this rule in his proof that $p \wedge \neg \mathrm{B}\, p$ is doxastically indefensible. Sorensen mentions that *some philosophers* find Hintikka's use of positive introspection problematic, but unfortunately he does not include references to the philosophers he has in mind.

There are contemporary philosophers who have argued that positive introspection is an idealization of the behavior of real world agents, most notably Williamson.[18] Williamson gives multiple arguments against positive introspection (of knowledge, but they can easily be extended to positive introspection of beliefs). For example, he argues that if I come to know that ϕ, then even if I do also come to know that I know ϕ through introspection, there must be a time delay between my coming to know that ϕ and my coming to know that I know ϕ. This must be the case because in order for me to get to know that I know ϕ through introspection, I must have the knowledge that ϕ before I can detect it with my introspection. Thus, if I come to know ϕ there is a moment in which I know ϕ, but I do not yet know that I know ϕ. Since there are moments in which I know ϕ but I do not know that I know ϕ, positive introspection is not valid as a logical principle.[19]

In this thesis we will step over the concern of time delay in introspection, as well as related concerns such as the problem of logical omniscience.[20] We take it that a formal accounts of knowledge and belief can still be valuable, especially if we understand the doxastic models we work with to be a representation of the implicit set of beliefs that an agent is committed to, rather than the beliefs he explicitly believes or is aware of.[21] The reader that is interested in accounts that attempt to deal with the problem of logical omniscience are referred to the overview article of Sim, and to the work of Solaki, who develops a logical account of bounded rationality that more closely resembles belief and knowledge as observed in real world agents.[22]

Appealing to What is Obvious

The second objection that Sorensen raises is that Hintikka's definition "makes a problematic appeal to what is obvious to speakers and their audiences".[23] This objection is clearly confused, as Hintikka's definition only makes an appeal to properties of formal models and avoids vague natural language notions. We think that this

[17] This can easily be verified from the definitions of his belief operators, given in Hintikka, 1962, 47-51.

[18] Williamson, 1999.

[19] Williamson, 1999, Chapter 6, section 4.

[20] For a description of the problem of logical omniscience, see Stalnaker, 1991.

[21] Meyer, 2003.

[22] Sim, 1997, Solaki, 2017.

[23] Sorensen, 1988, 21.

confusion stems from remarks like:

> “The absurdity of [p but I do not believe that p] is due to the fact that it is doxastically indefensible for the speaker to utter (although it is not itself indefensible) and that this doxastic indefensibility *is demonstrable in so simple a way it is felt by the speakers of the English language*.”[24] (our emphasis)

We suspect that Sorensen reads this passage as if Hintikka *defines* Moore sentences or the associated absurdity as something that is ‘demonstrable in so simple a way it is felt by the speakers of the English language’. But this is not what Hintikka does in this passage. Rather, Hintikka mentions that his definition of Moore sentences as doxastically indefensible sets of sentences applies to those (sets of) sentences that English speakers would find absurd to assert. So Hintikka does not suddenly supplement his formal definition with an appeal to what is obvious to English speakers, but instead states that his definition maps on to our everyday experience of the peculiarity of Moore sentences.

Believing that Someone Else Is Inconsistent

Sorensen borrows a third objection from Linsky. Linsky writes that on Hintikka’s account, it is epistemically indefensible for an agent A to believe that another agent B holds inconsistent beliefs.[25] The proof he presents for this claim is rather simple. Take an inconsistent sentence $\perp$, that is to say, a sentence that cannot be true in any world of any model that adheres to the constraints Hintikka proposes (seriality and transitivity). It is easy to see that $\mathrm{B}_A \perp$ is also inconsistent, because $\mathrm{B}_A \perp$ is only true in a world w if $\perp$ is true in all worlds v in the model that are accessible from world w for agent A. But, since $\perp$ is inconsistent, it is true in no worlds in the model. (And because Hintikka requires his models to be serial, there is at least one world u accessible from w for A, so $\mathrm{B}_A \perp$ cannot be vacuously true in w either.) So $\mathrm{B}_A \perp$ is also inconsistent. And then, since we just proved for an arbitrary inconsistent sentence $\perp$ and an arbitrary agent A that $\mathrm{B}_A \perp$ is inconsistent, it follows that since $\mathrm{B}_A \perp$ is inconsistent, $\mathrm{B}_B \, \mathrm{B}_A \perp$ is also inconsistent. Thus, by Hintikka’s definition, $\mathrm{B}_B \, \mathrm{B}_A \perp$ is doxastically indefensible. Linsky finds this result objectionable, because “Hintikka’s doxastic logic requires sets of statements to be declared ‘indefensible’ which are not indefensible in accordance with his informal and intuitive account of this concept (...) nothing in logic can tell [a person A] that it is false that another person B believes an inconsistent statement. Even a logical saint may believe that others are not logical saints.”[26]

We think there is some strength to Linsky’s objection, because intuitively it can indeed be defensible for an agent A to believe that some other agent B holds inconsistent beliefs. But, these situations are also not too hard to represent in doxastic logic. Typically, doxastic logics require the beliefs of a single agent to be consistent. In the modal logic KD45 for example, which is often used to model agents’ belief systems,[27] all relations on the models are required to be serial (for every world w, there is a world v such that vRw). But, as this holds in every world of every model,

[24] Hintikka, 1962, 71.
[25] Linsky, 1968.
[26] Linsky, 1968, 501.
[27] Meyer, 2003.

it is common belief among all agents that every agent is consistent. In this framework it is thus indeed indefensible for an agent to believe that some other agent is inconsistent. But, a simple solution to this problem is to drop the seriality constraint on the relations of the models. If (some) relations are allowed to be non-serial, it is possible to represent agents holding inconsistent beliefs.

Another common solution for dealing with inconsistent agents is to allow for *impossible worlds* in the models that represent agents' beliefs.[28] Different definitions of impossible worlds can be given. Priest defines them as worlds that are inconsistent with respect to the laws of classical logic.[29] So in an impossible world $p \wedge \neg p$ or any other inconsistent sentence $\perp$ might be true. And clearly, in a model that includes impossible worlds, we can represent agents that hold inconsistent beliefs: $\mathrm{B}_A \perp$ is true in w if it happens to be so that in all worlds v accessible from w for A, $\perp$ is true. And thus situations like the one Linsky describes, in which an agent A believes another agent B to hold inconsistent beliefs, can also be represented.

Of course, not everyone likes impossible worlds, primarily because it is hard to wrap one's head around what they are supposed to represent.[30] For the readers that do not wish to resort to impossible worlds to rebut Linsky's objection, and also do not want to drop seriality as a constraint on the relations on their models, we should mention that the definition of Moore sentences we offer in the following section is improved in such a way that it is invulnerable to Linsky's objection. Hintikka's definition is vulnerable to Linsky's objection because of a rather fundamental error in the definition. We explain this error in more detail in the next section, but the key observation is that according to Hintikka's definition, any set of sentences containing an inconsistent sentence is doxatically indefensible, as doxastic indefensibility requires only of a set of sentences that the conjunction of those sentences cannot be believed. Since Hintikka wishes to characterize Moore sentences as being doxastically indefensible, this implies that sets of sentences that contains an inconsistent sentence is a Moore sentence (or more precisely a Moore 'set-of-sentences). This is an error in the definition of Hintikka, because the hallmark of Moore sentences is that they are absurd *despite describing perfectly possible situations*. Inconsistent sentences do not describe possible situations and should therefore not be classified as Moore sentences. In our definition we fix this error by posing a satisfiability condition on Moore sentences. Therefore, we also do not classify sentences like $\mathrm{B}_B\,\mathrm{B}_A \perp$ as Moore sentences.

Gullibilism

The final objection Sorensen raises is from *gullibilism*, the philosophical position that in beliefs 'anything goes' because for every sentence there could be someone loony enough to believe it.[31] We should first mention again that gullibilism is not an objection to modeling beliefs in terms of logics, as one could simply model beliefs in a non-classical logic that puts no restrictions on what can be believed. But more importantly we should reply that anyone who adheres to gullibilism must have a hard time explaining why it should be absurd to believe a Moore sentence - after all, anything goes in beliefs. Gullibilism thus is not an argument specifically against

[28] See Rantala, 1982 for an example of an impossible worlds semantic.

[29] DeRose, 1997.

[30] For a suggestion on how to do this, see Priest, 1992.

[31] Sorensen, 1988, 22.

using formal methods to characterize beliefs, it is an argument against all methods that put constraints on what can or should be believed (and Sorensen's own method is certainly among these).

3.2.2 Limitations of Hintikka's Definition

Despite resisting most of Sorensen's objections, Hintikka's definition has some limitations. Firstly, Hintikka characterizes Moore sentences as doxastically indefensible statements, which he describes as:

> "The definition of doxastic indefensibility shows that this notion constitutes a direct generalization of the puzzling properties of Moore's sentence. Doxastically indefensible statements (utterances) might be true in the sense that the same form of word could be used by some other speaker to make a correct and true statement. (...) Doxastically indefensible statements are nevertheless impossible for the speaker to believe consistently."[32]

It is clear that Hintikka intends to characterize Moore sentences as sentences that (i) could be true and (ii) are impossible for the speaker to believe consistently. But nothing in the definition of a doxastically indefensible sentence guarantees that it can be true. In fact, all sentences that cannot be true are doxastically indefensible. To see this, consider that a set of sentences $\{\phi_1, ..., \phi_n\}$ is doxastically indefensible for an agent A iff $\mathrm{B}_A(\phi_1 \wedge ... \wedge \phi_n)$ is indefensible, i.e. iff $\mathrm{B}_A(\phi_1 \wedge ... \wedge \phi_n)$ is unsatisfiable. Clearly then any set of sentences $\{\phi_1, ..., \phi_n\}$ that is itself unsatisfiable is also doxastically indefensible for any agent A, because if $\phi_1 \wedge ... \wedge \phi_n$ is unsatisfiable there can be no world in which it is true, and thus there can also be no world in which $\mathrm{B}_A(\phi_1 \wedge ... \wedge \phi_n)$ is true. So on Hintikka's definition any unsatisfiable sentence is doxastically indefensible. This leads to counterintuitive results. Under this definition, any inconsistent sentence $\perp$ is a Moore sentence. This is clearly a mistake, as the hallmark of Moore sentences is that they are absurd, despite describing a perfectly possible situation. Intuitively, 'It is raining but I do not believe it' is intriguing exactly because it is absurd even though it describes a situation that can be true. Sentences like 'it is raining and it is not raining' are of a completely different order, as they are only absurd to utter in the sense that they are necessarily false. When we introduce our improved definition in the next section, we impose a satisfiability constraint on Moore sentences, to guarantee that they are absurd *despite* describing possible situations.

Hintikka's definition is also limited in that it explicitly refers to beliefs, which makes it inapt to deal with non-doxastic Moore sentences. For example, Williams argues that 'I drink stout but I do not want to' is a Moore sentence because it cannot be desired to be true, even though it can be believed.[33] In the final section of this chapter we elaborate on Williams' argument for this claim. For now it is important to note that if Williams is right in asserting that 'I drink stout but I do not want to' is a Moore sentence that can be believed, then Hintikka's definition of Moore sentences in terms of doxastic indefensibility is not complete, because under this definition only (sets of) sentences that cannot be believed are Moore sentences.

[32] Hintikka, 1962, 72.

[33] Williams, 2014.

3.3 New Definition of Moore Sentences

With our language and semantics in place and the philosophical considerations of the previous section in mind, we define Moore sentences as follows:

Definition 3.6. (Moore sentence) A sentence ϕ is a Moore sentence for an agent A iff:
- ϕ contains a modal operator $\Box_A$
- ϕ is satisfiable with respect to a class of models C that is appropriate for modeling $\Box_A$
- $\Box_A \phi$ is unsatisfiable with respect to C

This definition captures and extends the intuition behind the informal definition of Moore sentences as sentences that can be true but cannot be believed (or desired, or intended, and so forth). Also, the definition can deal flexibly with different kinds of modal operators, because the definition of (un)satisfiability is not tied to any particular class of models. Because of this the definition applies equally well to for example sentences containing the operator B_A and to sentences containing the operator K_A, even if we want to model these operators with respect to different classes of Kripke models.

Also note that our definition does not have the shortcomings that Hintikka's definition has. Our definition guarantees that Moore sentences can be true, and is suitably extended to account for non-doxastic Moore sentences. Also note that Linsky's objection to Hintikka's definition of Moore sentences, that it mistakenly classifies $\mathrm{B}_B \, \mathrm{B}_A \perp$ as a Moore sentence, does not apply to our definition, because $\mathrm{B}_B \, \mathrm{B}_A \perp$ is not satisfiable and therefore not a Moore sentence.[34]

3.4 Testing the New Definition

In this section we use the definition to classify sentences as (non-)Moorean. The definition succeeds if it can be used to make claims that are in line with our intuitions (as these are formulated in the literature) about which assumptions about propositional attitudes lead some sentences to be Moore sentences. First, we discuss the most well known Moore sentences, which are 'It is raining but I do not believe it' and a slight variation thereof. We discuss under which assumptions about belief these can be said to be Moore sentences. Second, we treat sentences that share their syntactic structure with these well-known Moore sentences, but involve different propositional attitudes. We again consider under which assumptions about the involved propositional attitudes these sentences can be said to be Moore sentences. Third, we discuss some sentences that do not share their basic syntactic structure with the well known Moore sentences, but that have nevertheless been called Moore sentences in the literature. Again, we assess under what assumptions about the involved propositional attitudes these can be said to be Moore sentences. This third section should justify our choice to define Moore sentences semantically rather than

[34] Linsky could reply that although we do not characterize $\mathrm{B}_B \, \mathrm{B}_A \perp$ as a Moore sentence, it is still problematic that $\mathrm{B}_B \, \mathrm{B}_A \perp$ is unsatisfiable on the class of models we use to model belief, because agents should be able to believe that other agents hold inconsistent beliefs. But, as we discussed in the fourth subsection of section 3.2.1, to accommodate the satisfiabilty of $\mathrm{B}_B \, \mathrm{B}_A \perp$ Linsky can choose to model belief for example with respect to a class of models of which the relations are not constrained by seriality.

syntactically, because some sentences can be argued to be Moore sentences despite being syntactically dissimilar to the well known Moore sentences.

3.4.1 Canonical Moore sentences

The original, and also the most famous, Moore sentence was introduced by Moore himself:

(1) "I went to the pictures last Tuesday but I don't believe that I did."[35]

Moore also introduced a slight variation on (1) (although Green and Williams suggest Moore may have been unaware that this sentence was structurally different from (1)[36]) when he wrote the sentence

(2) "I believe that he has gone out, but he has not."[37]

These sentences differ in that (1) reports the omission of a belief in a fact, whereas (2) reports the commitment to a false belief. These sentences have therefore been called the *omissive* Moore sentence and the *commissive* Moore sentence, respectively. We refer to these two sentences as the *canonical Moore sentences*.

The sentences (1) and (2) can straightforwardly be formalized as $p \wedge \neg \mathrm{B}\, p$ and $p \wedge \mathrm{B} \neg p$. To asses whether these are Moore sentences, we need to check whether $p \wedge \neg \mathrm{B}\, p$ and $p \wedge \mathrm{B} \neg p$ are satisfiable with respect to a class of Kripke models C suitable for modeling belief, and whether $\mathrm{B}(p \wedge \neg \mathrm{B}\, p)$ and $\mathrm{B}(p \wedge \mathrm{B} \neg p)$ are unsatisfiable with respect to C. In the discussion below, we work with KD45 Kripke models, which are often used to model beliefs.[38]. KD45 Kripke models are characterized by seriality (for all words w, there exists a world v s.t. wRv), transitivity (for all worlds w, v and u such that wRv and vRu, wRu) and being Euclidean (for all worlds w, v and u such that wRv and wRu, vRu). We assume that belief distributes over conjunction ($\mathrm{B}(\phi \wedge \psi) \rightarrow (\mathrm{B}\, \phi \wedge \mathrm{B}\, \psi)$) and is consistent ($\mathrm{B}\, \phi \rightarrow \neg \mathrm{B} \neg \phi$), and that positive introspection ($\mathrm{B}\, \phi \rightarrow \mathrm{B}\, \mathrm{B}\, \phi$) and negative introspection ($\neg \mathrm{B}\, \phi \rightarrow \mathrm{B} \neg \mathrm{B}\, \phi$) hold.

With these assumptions in place, we are in a position to prove that $p \wedge \neg \mathrm{B}\, p$ and $p \wedge \mathrm{B} \neg p$ are Moore sentences. We start by proving that $p \wedge \neg \mathrm{B}\, p$ is a Moore sentence. We do this by (i) proving that $p \wedge \neg \mathrm{B}\, p$ is satisfiable with respect to a class of Kripke frames that is characterized by seriality, transitivity and being Euclidean, and (ii) proving that $\mathrm{B}(p \wedge \neg \mathrm{B}\, p)$ is unsatisfiable given our assumptions that belief distributes over conjunction, is consistent and that positive and negative introspection hold.

Proof

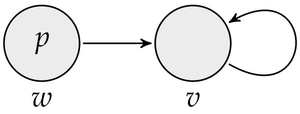
(i) Take a model M with two possible worlds w and v and the accessibility relation R such that wRv and vRv, and let p be true in w but false in v. Then M is serial and transitive, and $p \wedge \neg \mathrm{B}\, p$ is true in w.

[35] Moore, 1942, 543.

[36] Green and Williams, 2007, 5.

[37] Moore, 1944.

[38] Meyer, 2003



(ii)
(1) $\mathrm{B}(p \wedge \neg\,\mathrm{B}\,p)$ (supposition for reductio)
(2) $\mathrm{B}\,p \wedge \mathrm{B}\neg\,\mathrm{B}\,p$ (distribution of belief over conjunction, 1)
(3) $\mathrm{B}\,p$ (elimination of conjunction, 2)
(4) $\mathrm{B}\neg\,\mathrm{B}\,p$ (elimination of conjunction, 2)
(5) $\neg\,\mathrm{B}\neg\neg\,\mathrm{B}\,p$ (consistency of beliefs, 4)
(6) $\neg\,\mathrm{B}\,\mathrm{B}\,p$ (elimination of negation, 5)
(7) $\mathrm{B}\,\mathrm{B}\,p$ (positive introspection, 3)
(8) $\perp$ (introduction of conjunction, 6 and 7)

A similar proof can be given for $p \wedge \mathrm{B}\neg p$. Again, we prove (i) that $p \wedge \mathrm{B}\neg p$ is satisfiable with respect to a class of Kripke frames that is characterized by seriality, transitivity and being Euclidean, and (ii) proving that $\mathrm{B}(p \wedge \mathrm{B}\neg p)$ is unsatisfiable given our assumptions that belief distributes over conjunction and is consistent and that positive and negative introspection hold.

Proof
(i) Notice that in the model M given in the proof that $p \wedge \neg\,\mathrm{B}\,p$ is satisfiable above, $p \wedge \mathrm{B}\neg p$ is also true in w in M.

(ii)
(1) $\mathrm{B}(p \wedge \mathrm{B}\neg p)$ (supposition for reductio)
(2) $\mathrm{B}\,p \wedge \mathrm{B}\,\mathrm{B}\neg p$ (distribution of belief over conjunction, 1)
(3) $\mathrm{B}\,p$ (elimination of conjunction, 2)
(4) $\mathrm{B}\,\mathrm{B}\neg p$ (elimination of conjunction, 2)
(5) $\neg\,\mathrm{B}\neg p$ (consistency of beliefs, 3)
(6) $\mathrm{B}\neg\,\mathrm{B}\neg p$ (negative introspection, 5)
(7) $\neg\,\mathrm{B}\neg\,\mathrm{B}\neg p$ (consistency of beliefs, 4)
(8) $\perp$ (introduction of conjunction, 6 and 7)

In accordance with the intuitions as formulated in the literature, our definition thus correctly classifies (1) and (2) as Moore sentences, assuming that belief is characterized by distribution over conjunction, consistency, and positive and negative introspection.

3.4.2 Varying Semantic Content

The literature has proposed several sentences that have the same syntactic structure as the canonical Moore sentences, $p \wedge \neg\Box_A p$ and $p \wedge \Box_A \neg p$, but that concern some other propositional attitudes than belief. If this new propositional attitude is modeled with respect to a different class of models than belief is, these sentences differ in semantic content from the canonical Moore sentences. If these sentences can nevertheless be said to be Moore sentences this is interesting because this would promise to extend the results of Moore's paradox beyond the study of beliefs, towards the study of propositional attitudes such as knowledge, desire and moral motivation.

Firstly, Williams proposes the following sentence to be an epistemic Moore sentence:[39]

(3) It is raining but I do not know it is.

This sentence can be formalized as $p \wedge \neg \mathrm{K}\, p$. Taking into consideration the distinction between omissive and commissive Moore sentences we introduced in the previous section, one may wonder why Williams does not also propose the commissive variant of (3), which would be formalized as $p \wedge \mathrm{K} \neg p$:

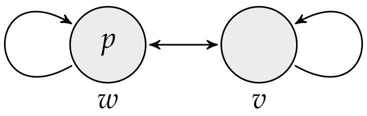
(4) It is raining but I know it is not.

However, it is immediately clear that $p \wedge \mathrm{K} \neg p$ is not a Moore sentence, if we adopt the common assumption that knowledge is *veridical* ($\mathrm{K}\, \phi \rightarrow \phi$).[40] Under this assumption, the class of models that is appropriate for modeling knowledge is characterized at least by reflexivity (for all worlds w, wRw). $p \wedge \mathrm{K} \neg p$ is not a Moore sentence because it is unsatisfiable with respect to such a class of models.

Proof:
(1) $p \wedge \mathrm{K} \neg p$ (supposition for reductio)
(2) ϕ (elimination of conjunction, 1)
(3) $\mathrm{K} \neg p$ (elimination of conjunction, 1)
(4) $\neg p$ (veridicality of knowledge, 3)
(5) $\bot$ (introduction of conjunction, 1 and 3)

Thus, given only the assumption that knowledge is veridical, $p \wedge \mathrm{K} \neg p$ is not a Moore sentence. But, $p \wedge \neg \mathrm{K}\, p$ on the other hand, can be a Moore sentence under this assumption. To prove this, we prove that (i) $p \wedge \neg \mathrm{K}\, p$ is satisfiable with respect to a class of Kripke frames that is characterized at least by reflexivity, transitivity and being Euclidean, and (ii) that $\mathrm{K}(p \wedge \neg \mathrm{K}\, p)$ is unsatisfiable if we assume that knowledge is at least consistent, distributes over conjunction, allows for positive and negative introspection and is veridical.

Proof
(i) Take a model M with an accessibility relation R and worlds w and v s.t. wRw, wRv, vRw and vRv. Let p be true in w and false in v. Then M is serial, transitive, Euclidean and reflexive, and $p \wedge \neg \mathrm{K}\, p$ is true in w.



(ii)
(1) $\mathrm{K}(p \wedge \neg \mathrm{K}\, p)$ (supposition for reductio)
(2) $\mathrm{K}\, p \wedge \mathrm{K} \neg \mathrm{K}\, p$ (distribution of knowledge over conjunction, 1)
(3) $\mathrm{K}\, p$ (elimination of conjunction, 2)
(4) $\mathrm{K} \neg \mathrm{K}\, p$ (elimination of conjunction, 2)
(5) $\neg \mathrm{K}\, p$ (veridicality of knowledge, 4)
(6) $\bot$ (introduction of conjunction, 3 and 5)

[39] Williams, 2015a.
[40] Meyer, 2003

Related to this epistemic Moore sentence proposed by Williams, Yalcin discusses whether the following are Moore sentences:[41]

(5) It is raining and it might not be raining.

(6) It is raining and possibly it is not raining.

Yalcin suggests 'it might not be raining' and 'possibly it is not raining' might be interpreted epistemically, as 'I do not know that it is not raining'. Under this interpretation, (5) and (6) have similar meanings to (3) and can thus also be classified as Moore sentences under the assumptions we mentioned in our discussion of (3), baring small semantic difference between asserting that something might be the case, that something possibly is the case and that you do not know that something is the case.

However, this is not the only possible interpretation of the modalities in (5) and (6). Particularly the phrase 'possibly' in (6) can be interpreted *metaphysically*, rather than epistemically. Let $\Diamond$ be the dual of $\Box$ and let the reading of $\Diamond\phi$ be 'it is metaphysically possible that ϕ' (whatever that may mean exactly). Then, under a metaphysical interpretation of 'possibly' (6) means roughly:

(7) It is raining but it is metaphysically possible that it is not raining.

We can formalize (7) as $p \wedge \Diamond\neg p$. And we can easily imagine a commissive counterpart of (7) of the form $p \wedge \neg\Diamond p$, which could be a formalization of a sentence such as:

(8) It is raining but it is metaphysically possible that it is not raining.

There is a long standing debate on the properties of metaphysical possibility that has not yet reached a consensus.[42] But, one assumption that seems to be universally agreed upon is that whatever is the case is possible ($\phi \rightarrow \Diamond\phi$).[43] From this it is clear that $p \wedge \neg\Diamond p$, is not a Moore sentence, as it is not satisfiable in Kripke models that satisfy the assumption that whatever is true is possible (these are reflexive Kripke models).

Proof:
(1) $p \wedge \neg\Diamond p$ (supposition for reductio)
(2) p (elimination of conjunction, 1)
(3) $\neg\Diamond p$ (elimination of conjunction, 1)
(4) $\Diamond p$ (possibility of the actual, 2)
(5) $\perp$ (introduction of conjunction, 3 and 4)

From the assumption that whatever is the case is possible, it thus follows that $p \wedge$

[41] Yalcin, 2007, 1.

[42] See Chapter 10 of Berto and Plebani, 2015 for an excellent overview of the debate on the nature of possible worlds, which touches also on the formal properties of metaphysical possibility.

[43] Menzel, 2017, 1.

$\neg\Diamond p$ is not a Moore sentence, because it is not satisfiable on a class of models appropriate for modeling $\Diamond$.

From the assumption that whatever is the case is possible it also follows that $p \wedge \Diamond\neg p$ is not a Moore sentence. To see this, assume again that whatever is the case is possible ($\phi \rightarrow \Diamond\phi$) and consider that if $p \wedge \Diamond\neg p$ is a Moore sentence, two things are true: (i) $p \wedge \Diamond\neg p$ is satisfiable and (ii) $\Diamond(p \wedge \Diamond\neg p)$ is unsatisfiable, with respect to a class of Kripke models characterized at least by reflexivity. But (i) and (ii) cannot both be true, because (i) implies the negation of (ii). To see this, consider that by the definition of satisfiability if $\phi \wedge \Diamond\neg\phi$ is satisfiable, then there is a world w in a model M in which $p \wedge \Diamond\neg p$ is true. But then, under our assumption that $\phi \rightarrow \Diamond\phi$, this implies that $\Diamond(p \wedge \Diamond\neg p)$ is also true in w in M. So if $p \wedge \Diamond\neg p$ is satisfiable, then so is $\Diamond(p \wedge \Diamond\neg p)$. Therefore if (i) is true, (ii) is false, and thus $p \wedge \Diamond\neg p$ is not a Moore sentence if we assume that whatever is the case is possible.

More recently, authors have proposed Moore sentences with non-epistemic propositional attitudes. Wall introduces Moore sentences concerning desire:[44]

(9) I have cheesecake, and I desire that I have no cheesecake.

(10) I have cheesecake, and no desire that I have cheesecake.

Wall argues that these sentences are similar to doxastic Moore sentences because just like doxastic Moore sentences can be true but are absurd to believe, these sentences can be true but are absurd to desire. Clearly, it is possible that I have cheesecake without desiring to have cheesecake or even whilst desiring not to have cheesecake, but it would be absurd for me to desire such situations to occur. As Williams puts it, it would be absurd for me to desire my own desire to be frustrated, or to desire that I do not desire that what obtains.[45]

Williams' account of the absurdity of (9) and (10) poses an interesting technical challenge. If we formalize (9) and (10) as $p \wedge \mathrm{D}\,\neg p$ and $p \wedge \neg\,\mathrm{D}\,p$ respectively, Williams claims that $\mathrm{D}(p \wedge \mathrm{D}\,\neg p)$ and $\mathrm{D}(p \wedge \neg\,\mathrm{D}\,p)$ describe situations that are absurd (in a way that is similar to the absurdity of believing $p \wedge \neg\,\mathrm{B}\,p$ or $p \wedge \mathrm{B}\,\neg p$). However, because we define Moore sentences in terms of satisfiability with respect to logical frameworks, we need to know the formal properties of desire if we want to capture these absurdities. We discuss the formal properties of desire at length in Chapter 5, where we also consider under what assumptions (9) and (10) can be said to be Moore sentences.

Lastly, Cholbi proposes that there are Moore sentences involving moral motivation, such as:[46]

(11) Hurting animals is wrong, but I do not care.

According to Cholbi, this sentence is a Moore sentence because, just like the canonical Moore sentences, it cannot be coherently *asserted*.[47] Assuming that Schoemaker's

[44] Wall, 2012, 2.

[45] Williams, 2014, 3.

[46] Cholbi, 2009.

[47] Cholbi, 2009, 497.

principle, which we introduced in Chapter 2, is correct, this means that (11) is a Moore sentence because just like canonical Moore sentence, it is absurd to *believe*.

Cholbi's explanation of why (11) is a Moore sentence is interesting, as it seems to defy our definition. For a sentence ϕ to be a Moore sentence for agent A, we require ϕ to contain a modal operator $\Box_A$, such that $\Box_A\phi$ is unsatisfiable (with respect to a class of Kripke models suitable for modeling $\Box_A$). Cholbi suggests that a sentence ϕ that contains one modal operator $\Box^1$ is a Moore sentence because if we take a different modal operator $\Box^2$, $\Box^2\phi$ is unsatisfiable (with respect to a class of Kripke models suitable for modeling both $\Box^1$ and $\Box^2$).

There are different ways we can make sense of Cholbi's explanation. The first is to open up our definition of Moore sentences to allow for sentences that contain one modal operator and become absurd when prescripted with a different modal operator. However, we wish to resist this easy solution as long as we can, as such a widening of our definition may lead to unexpected false positives. A more elegant solution acknowledges that in sentences with multiple modalities, it is not only important to understand the properties of the single modalities, but rather the way the two modalities interact with each other. To determine whether it is absurd to believe Cholbi's (11) for example, we need to analyze in depth the way care and belief interact.

An extensive study towards the interactions between care and belief is beyond the scope of this thesis, but we note that in (11), 'I do not care' seems to mean that the speaker does not *believe* that hurting animals is wrong. It may not generally be the case that not caring for something implies not believing it, but it may be true for moral statements, like that in the first conjunct of (11). If we can say that if I do not care about a moral statement, I do not believe it is true, then (11) implies that 'hurting animals is wrong, but I do not believe so'. And if this reduction is correct, we can thus say that (11) does contain the propositional attitude belief, or rather that (11) implies a sentence that contains belief. And then Cholbi's suggestion that (11) is a Moore sentence because it cannot coherently be believed is in line with our definition of Moore sentences. It is absurd to believe (11), insofar as belief in (11) implies belief in 'hurting animals is wrong, but I do not believe so', which is absurd to believe.

Although no other non-doxastic Moore sentences have been discussed substantially in the literature, we take it that the previously mentioned examples show that Moore sentences can occur in a much wider range of settings than is commonly thought. And we believe that many Moore sentences still have to be discovered. For example, although they are not discussed in the literature, we see no prima facie reason why the following sentences could not be considered Moore sentences:

(12) I intend to go to the gym but I will not.

(13) I should give to charities but I do not.

Clearly, it is possible that I do not go to the gym even though I intend to go. But, prima facie, it would be absurd for me to intend to have the intention to go the gym while not actually going to the gym; we discuss this matter at length in Chapter 5. Similarly, it can be the case that I do not give to charities even though I should.

But, it seems absurd if it should be the case that I don't give to charities while I should (because then I would be morally obliged to fail to meet my moral obligations). Of course, whether these intuitions can be substantiated depends on what formal properties 'intend' and 'should' are assumed to have, but there seems to be enough intuitive ground to warrant an investigation towards these properties.

The question which propositional attitudes can be used to form Moore sentences is as interesting as it is vast. There are many propositional attitudes to consider, and when one forms sentences with them of the syntactic structure of the canonical Moore sentences, many of them sound at least slightly odd. 'It is raining but I fear it is not', 'it is raining but I doubt it is', 'It is raining but I hope it is not', and so forth. Are these sentences absurd to fear, doubt, or hope, respectively? Are they absurd to believe, know or desire? Taking into account the relations between all the different propositional attitudes, the potential for finding new Moore sentences is grand. Exploring which propositional attitudes can be swapped into the syntactic structure $p \wedge \neg\Box_A p$ or $p \wedge \Box_A \neg p$ to create absurd sentences that can be true, and mapping out why some propositional attitudes create absurd statements while others do not, is a worthwhile project for future research.

3.4.3 Varying Syntactic Structures

Although historically the debate on Moore sentences has mostly been about sentences of the syntactic structures $p \wedge \neg\Box_A p$ and $p \wedge \Box_A \neg p$, Moore sentences of deviant syntactic structures have been proposed. Definitions that refer solely to the syntactic structure of sentences are unable to classify these sentences as Moore sentences. Using our semantic definition, we are able to consider of sentences with a non-canonical syntactic structure, whether or not they are Moore sentences.

Sorensen proposes the following iterated Moore sentences as Moore sentences:[48]

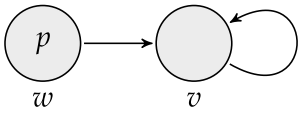
(14) It is raining but I do not believe that I believe it.

(15) It is raining but I believe that I believe it is not.

which could respectively be formalized as $p \wedge \neg \mathrm{B}\,\mathrm{B}\,p$ and $p \wedge \mathrm{B}\,\mathrm{B}\,\neg p$. These sentences can both be said to be Moore sentences under the assumptions that belief is at least consistent, distributes over conjunction and that positive and negative introspection hold. We first prove that $p \wedge \neg \mathrm{B}\,\mathrm{B}\,p$ is a Moore sentence by proving (i) that $p \wedge \neg \mathrm{B}\,\mathrm{B}\,p$ is satisfiable with respect to a class of Kripke models that is characterized at least by seriality, transitivity and being Euclidean and (ii) that $\mathrm{B}(p \wedge \neg \mathrm{B}\,\mathrm{B}\,p)$ is unsatisfiable, assuming that belief is at least consistent, distributes over conjunction and that positive and negative introspection hold.

Proof
(i) Take a model M with worlds w and v and the accessibility relation R s.t. wRv and vRv and let p be true in w but false in v. Then M is serial, transitive and Euclidean, and in w $p \wedge \neg \mathrm{B}\,\mathrm{B}\,p$ is true.

[48] Sorensen, 2000.



(ii)
(1) $\mathrm{B}(p \wedge \neg\,\mathrm{B}\,\mathrm{B}\,p)$ (supposition for reductio)
(2) $\mathrm{B}\,p \wedge \mathrm{B}\,\neg\,\mathrm{B}\,\mathrm{B}\,p$ (distribution of belief over conjunction, 1)
(3) $\mathrm{B}\,p$ (elimination of conjunction, 2)
(4) $\mathrm{B}\,\neg\,\mathrm{B}\,\mathrm{B}\,p$ (elimination of conjunction, 2)
(5) $\mathrm{B}\,\mathrm{B}\,p$ (positive introspection, 3)
(6) $\mathrm{B}\,\mathrm{B}\,\mathrm{B}\,p$ (positive introspection, 5)
(7) $\neg\,\mathrm{B}\,\neg\,\mathrm{B}\,\mathrm{B}\,p$ (consistency of beliefs, 6)
(8) $\perp$ (introduction of conjunction, 4 and 7)

A similar proof can be given for the claim that $p \wedge \mathrm{B}\,\mathrm{B}\,\neg p$ is a Moore sentence. We prove that (i) $p \wedge \mathrm{B}\,\mathrm{B}\,\neg p$ is satisfiable with respect to a class of Kripke models that is characterized at least by seriality, transitivity and being Euclidean and (ii) $\mathrm{B}(p \wedge \mathrm{B}\,\mathrm{B}\,\neg p)$ is unsatisfiable, assuming that belief is at least consistent, distributes over conjunction and that positive and negative introspection hold.

Proof
(i) Note that in the model M supplied in the proof that $p \wedge \neg\,\mathrm{B}\,\mathrm{B}\,p$ is a Moore sentence, $p \wedge \mathrm{B}\,\mathrm{B}\,\neg p$ is true in world w in M.

(ii)
(1) $\mathrm{B}(p \wedge \mathrm{B}\,\mathrm{B}\,\neg p)$ (supposition for reductio)
(2) $\mathrm{B}\,p \wedge \mathrm{B}\,\mathrm{B}\,\mathrm{B}\,\neg p$ (distribution of belief over conjunction, 1)
(3) $\mathrm{B}\,p$ (elimination of conjunction, 2)
(4) $\mathrm{B}\,\mathrm{B}\,\mathrm{B}\,\neg p$ (elimination of conjunction, 2)
(5) $\neg\,\mathrm{B}\,\neg p$ (consistency of beliefs, 3)
(6) $\mathrm{B}\,\neg\,\mathrm{B}\,\neg p$ (negative introspection, 5)
(7) $\neg\,\mathrm{B}\,\neg\neg\,\mathrm{B}\,\neg p$ (consistency of beliefs, 6)
(8) $\neg\,\mathrm{B}\,\mathrm{B}\,\neg p$ (elimination of double negation, 7)
(9) $\mathrm{B}\,\neg\,\mathrm{B}\,\mathrm{B}\,\neg p$ (negative introspection, 8)
(10) $\neg\,\mathrm{B}\,\neg\neg\,\mathrm{B}\,\mathrm{B}\,\neg p$ (consistency of beliefs, 9)
(11) $\neg\,\mathrm{B}\,\mathrm{B}\,\mathrm{B}\,\neg p$ (elimination of double negation, 10)
(12) $\perp$ (introduction of conjunction, 4 and 11)

Another interesting sentence of non-standard syntactic structure was proposed by the famous puzzle logician Raymond Smullyan:

(16) You believe that God exists if and only if God does not exist.[49]

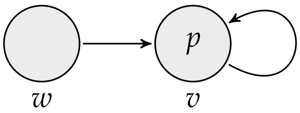
We can formalize this sentence as $\mathrm{B}(p) \leftrightarrow \neg p$. This sentence is interesting because to our knowledge it has not been mentioned in discussions about Moore sentences (it is for example not included in the two articles of Williams about Moore's

[49] Smullyan, 2012, 73.

paradox in speech[50] and Moore's paradox in thought[51], nor in the introductory article Green and Williams wrote for the *New essays* volume on Moore's paradox[52], even though it is a Moore sentence if we assume that belief is characterized at least by consistency, distribution over conjunction, closure under believed implication $((\mathrm{B}\,\phi \wedge \mathrm{B}(\phi \rightarrow \psi)) \rightarrow \mathrm{B}\,\psi)$ positive introspection and negative introspection. To prove this we (i) prove that $\mathrm{B}(p) \leftrightarrow \neg p$ is satisfiable with respect to a class of Kripke models that is characterized at least by seriality, transitivity and being Euclidean and (ii) that $\mathrm{B}(\mathrm{B}(p) \leftrightarrow \neg p)$ is unsatisfiable, assuming that belief is at least consistent, distributes over conjunction and that positive and negative introspection hold.

Proof

(i) Take a model M with two possible worlds w and v and the accessibility relation R such that wRv and vRv, p is not true in w but p is true in v. Then M is serial, transitive and Euclidean, and $\mathrm{B}(p) \leftrightarrow \neg p$ is true in w.



(ii)
(1) $\mathrm{B}(\mathrm{B}(p) \leftrightarrow \neg p)$ (supposition for reductio)
(2) $\mathrm{B}\,p$ (assumption)
(3) $\mathrm{B}\,\mathrm{B}\,p$ (positive introspection, 2)
(4) $\mathrm{B}\,\neg p$ (closure of belief under believed implication, 1 and 3)
(5) $\neg\,\mathrm{B}\,\neg\neg p$ (consistency of beliefs, 4)
(6) $\neg\,\mathrm{B}\,p$ (elimination of double negation, 5)
(7) $\bot$ (introduction of conjunction, 2 and 6)
(8) $\neg\,\mathrm{B}\,p$ (introduction of negation, 2 and 7)
(9) $\mathrm{B}\,\neg\,\mathrm{B}\,p$ (negative introspection, 8)
(10) $\mathrm{B}\,\neg\neg p$ (closure of belief under believed implication (by modus tollens), 1 and 9)
(11) $\mathrm{B}\,p$ (elimination of double negation, 10)
(12) $\bot$ (introduction of conjunction, 8 and 11)

3.5 Conclusion

In this chapter we discussed a previous attempt to formally define Moore sentences, and objections to this definition. Based on this discussion we proposed our own formal definition, which is an improvement on the previous definition in that it guarantees that Moore sentences can be true, and in that it applies to a wider range of Moore sentence than just the doxastic ones. We verified that our definition can be used to make claims that are in line with our intuitions about under which assumptions about propositional attitudes different sentences can be said to be Moore sentences. In the following chapters we use our definition to assess under which assumptions it can be said that Moore sentences are involved in the surprise exam paradox, the knowability paradox, the Toxin problem and Newcomb's problem.

[50] Williams, 2015a.
[51] Williams, 2015b.
[52] Green and Williams, 2007.

Chapter 4

The Surprise Exam Paradox and the Knowability Paradox

The surprise exam paradox and the knowability paradox are two epistemic paradoxes that apparently involve a Moore sentence. In this chapter we discuss under which assumptions about knowledge we can say that this is the case. We also discuss some dissimilarities between the Moore sentences in the surprise exam paradox and the knowability paradox, and the canonical Moore sentences introduced in the previous chapter. By illuminating these similarities and dissimilarities we are able to provide a new perspective on some established results and open questions concerning the surprise exam paradox and the knowability paradox.

4.1 The Surprise Exam Paradox

Many authors have recognized that a Moore sentence is involved in the surprise exam paradox.[1] It is debated however, what sentence in the scenario exactly fulfills this role. In this chapter we use our formal account of Moore sentences to settle this debate.

4.1.1 Introducing the Surprise Exam Paradox

The surprise exam paradox was first presented by O'Connor (although in his version the story concerns an unexpected military drill rather than an unexpected exam).[2] Since then, many variations of it have been proposed. We use a variation of the most common presentation, introduced by Scriven, in which a surprise exam is announced and also successfully given:[3],[4]

> The teacher announces to his students that next week there will be a surprise exam. This is to say that there will be an exam on a day such that prior to that day the students did not believe the exam would be on that day.

[1] Binkley, 1968, Clark, 1994, Schick, 2000, Williams, 2007, Levy, 2009, Gerbrandy, 2007.

[2] O'Connor, 1948.

[3] Scriven, 1951.

[4] Our presentation differs from Scriven's in that we define a surprise exam as an exam of which the students do not believe in advance that it will be given on the day that it is given, rather than as an exam of which the students cannot know in advance on which day it will be given. Also, Scriven spoke about the surprise military drill that O'Connor introduced rather than a surprise exam, but this difference is only superficial.

> A clever student objects to the announcement: "but if we are not to believe that it is the day of the exam, before the day of exam, then the exam cannot be given on Friday. Because if the exam would be given on Friday, we would know by Thursday evening that the exam had not been given before Friday. Then we would know (and thus believe) the exam would have to be given on Friday; so it would not be a surprise exam. Furthermore, since we know that the exam cannot be given on Friday, we know it cannot be given on Thursday either. Because Wednesday evening we would know that the exam was not given before Thursday and we already excluded Friday, so we would know the exam would have to be given on Thursday. Thus a Thursday exam would no be surprising either. And we can continue this reasoning until no days remain in the week. So you cannot give us a surprise exam next week."
>
> The teacher gave the exam on Tuesday, and the students were surprised.

The scenario above is considered paradoxical because the students have a seemingly sound argument for the absurd conclusion that a surprise exam cannot be given next week.

We should note that many variations of the surprise exam paradox exist. In early variations, the last sentence in the description was omitted, which lead several authors to (falsely) believe that the student correctly proved the impossibility of an announced surprise exam.[5] Further, there are variations in the definition of 'surprise', some authors defining it in terms of the students not having a justified belief about on which day the exam will be given[6], others in terms of the students' inability to predict the date of the exam[7]. We choose to define *surprise* in terms of the clever student not *believing* that the exam will be given before the day of the examination, so that the student is surprised if the exam is given on day d and the student does not believe the exam will be given on day d. This allows us to draw a direct comparison between the announcement in the surprise exam paradox and the canonical Moore sentence $p \wedge \neg \mathrm{B}\, p$.

4.1.2 Moore Sentence in the Surprise Exam Paradox

In the literature it is describe that a variation on the canonical Moore sentence $p \wedge \neg \mathrm{B}\, p$ plays a role in the surprise exam paradox. Binkley writes that for the students the teacher's announcement in the surprise exam paradox is similar to Moore's sentence on Thursday evening.[8] Because the announcement of the surprise exam next week is, on Thursday evening, equivalent to the announcement: 'Friday there will be an exam but you do not believe there will be an exam on Friday'. This sentence can be formalized in the same manner as the canonical Moore sentence, $p \wedge \neg \mathrm{B}\, p$, and we already prove that this sentence is a Moore sentence in section 3.4.1.

It is debated whether the teacher's announcement only becomes unbelievable for the student on Thursday evening or whether it is unbelievable for the student

[5] See e.g. O'Connor, 1948, or Alexander, 1950.

[6] Williams, 2007.

[7] Wright and Sudbury, 1977.

[8] Binkley, 1968.

right from the moment the teacher makes it. Binkley thought that the announcement became unbelievable only on the penultimate evening of the week. Most other authors have backed this position, or otherwise have argued that the announcement becomes unknowable only on Friday.[9] Quine on the other hand argued that, because the student can disprove the teacher's announcement with her backward induction argument on the Sunday before the test week, the announcement is unknowable right after the teacher makes it.[10]

The teacher's announcement can be formalized a number of different ways. Gerbrandy formalizes the announcement as $(d1 \wedge \neg \mathrm{K}\, d1) \vee (d2 \wedge \neg \mathrm{K}\, d2) \vee (d3 \wedge \neg \mathrm{K}\, d3) \vee (d4 \wedge \neg \mathrm{K}\, d4) \vee (d5 \wedge \neg \mathrm{K}\, d5)$ where $d1$ through $d5$ denote the propositions that the exam is given on day one through day five respectively (where day one refers to Monday), and $\mathrm{K}\,\phi$ is read as 'the student knows that ϕ'.[11] He thus interprets 'surprise' as the student not knowing that the exam will be given on the day that it is given. We choose to analyze surprise in terms of belief, so that the student is surprised if she does not believe the exam will be given on the day that it is given. Formally, we write the teacher's announcement as $(d1 \wedge \neg \mathrm{B}\, d1) \vee (d2 \wedge \neg \mathrm{B}\, d2) \vee (d3 \wedge \neg \mathrm{B}\, d3) \vee (d4 \wedge \neg \mathrm{B}\, d4) \vee (d5 \wedge \neg \mathrm{B}\, d5)$, where $\mathrm{B}\,\phi$ is read as 'the student believes ϕ'.[12] Henceforth we abbreviate this announcement as α. If the student believes α, and the exam is not given by Thursday evening, then her belief base consists of the sentence $\alpha \wedge \neg d1 \wedge \neg d2 \wedge \neg d3 \wedge \neg d4$, which reduces to $d5 \wedge \neg \mathrm{B}\, d5$. This sentence describes that the exam is on Friday and that the student does not believe this. Again, the latter sentence is obviously a Moore sentence for the student, and thus a sentence that the student cannot believe. Thus, our formalization of the surprise exam paradox suggests that Binkley is right in asserting that the announcement of the teacher is a Moore sentence after Thursday has passed an no exam has been given.

To settle Quine and Binkley's debate we then need to determine whether the announcement α of the teacher is a Moore sentence for the student directly after the teacher asserts it. We thus need to consider whether α is satisfiable with respect to a class of Kripke models suitable for modeling B, while $\mathrm{B}\,\alpha$ is not. Let us assume that belief distributes over conjunction and is consistent and that positive and negative introspection hold. Further, we should put strong constraints on our models to ensure that they represent the situation presented in the surprise exam paradox. We propose the following model to capture the situation described in the surprise exam paradox:

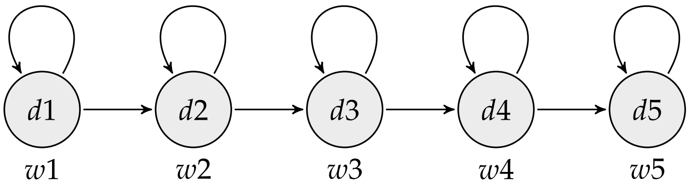
Take a model M with five worlds, $w1$, $w2$, $w3$, $w4$ and $w5$, that represent days Monday through Friday. Take an accessibility relation R such that all worlds access themselves and the worlds that are indexed with a number higher than their own index. Let the propositions $d1$ through $d5$ be true in worlds $w1$ through $w5$ respectively; these are the propositions 'the exam is on day one' through 'the exam is on day five'.

[9] See e.g. Wright and Sudbury, 1977 and Champlin, 1976.

[10] Quine, 1953.

[11] Gerbrandy, 2007, 24.

[12] Yet another possibility would be to interpret surprise as the students believing that the exam will not be given on the day that it is given. In this case the announcement could be formalized as $(d1 \wedge \mathrm{B}\,\neg d1) \vee (d2 \wedge \mathrm{B}\,\neg d2) \vee (d3 \wedge \mathrm{B}\,\neg d3) \vee (d4 \wedge \mathrm{B}\,\neg d4) \vee (d5 \wedge \mathrm{B}\,\neg d5)$.



[13]

This model represents an agent who on every day believes that the exam can be given on that day or any of the following days in the week, but not on days that already passed. For example, if the first three days have passed, the agent believes $d4 \lor d5$.[14] A result of this is that if the exam was not given on the first four days, and hence we evaluate the agents' beliefs at $d5$, the agent believes that the exam will be given on $d5$.

To assess whether α is a Moore sentence, we need to consider whether α is satisfiable and $\mathrm{B}\,\alpha$ is unsatisfiable on a suitable class of Kripke models. Since we constrained the class of models we work with to represent the surprise exam paradox to pretty much one model M, this amounts to considering whether α is true in at least one world in M while $\mathrm{B}\,\alpha$ is false in all worlds in M. It turns out that this is the case. In M, α $((d1 \land \neg\,\mathrm{B}\,d1) \lor (d2 \land \neg\,\mathrm{B}\,d2) \lor (d3 \land \neg\,\mathrm{B}\,d3) \lor (d4 \land \neg\,\mathrm{B}\,d4) \lor (d5 \land \neg\,\mathrm{B}\,d5))$ is true in worlds $w1$ through $w4$. But since α is not true in $w5$, and all worlds access $w5$, $\mathrm{B}\,\alpha$ is false in all worlds in M. Therefore, the teacher's announcement α can be said to be a Moore sentence under the assumption that M adequately represents the situation described in the surprise exam paradox.

Our formalization thus suggests that Quine was right after all, in spite of the majority of philosophers disagreeing with him.[15] The announcement of the teacher is a Moore sentence not only by Thursday evening, but directly after its assertion. The intuition behind this is in fact described beautifully in Quine's original article on the subject:[16]

> "If [the falsity of the teacher's announcement] is a conclusion which she is prepared to accept (though wrongly) in the end as a certainty, it is an alternative which she should have been prepared to take into consideration from the beginning as a possibility."

If the student is willing to accept as a certainty that the announcement would be false after Thursday, she has take the possibility that the announcement is false into consideration already in the beginning of the week; since she accepts that the announcement can turn out false she should not believe categorically that it is true.

4.1.3 The Twists

In the previous section we showed that, given some plausible assumptions about belief and about the surprise exam paradox, we can say that a Moore sentence is

[13] Note that the model is transitive; not all accessibility relation lines are drawn for clarity.

[14] Note that the fact that the first three days have passed in this case is not explicitly represented in the model, only the fact that the student believes that the first three days have passed is represented. To represent the fact that the first three days have indeed passed one needs a different system, for instance the dynamic epistemic logic framework presented in Gerbrandy, 2007.

[15] Quine, 1953, Chow, 1998.

[16] Quine, 1953, 65.

involved in the surprise exam paradox. But the Moore sentence in the surprise exam paradox may be (or at least *feel*) more complex or problematic than the canonical Moore sentence $p \wedge \neg \mathrm{B}\, p$, in at least three respects.

Incentives

First, the teacher's announcement has a built in incentive that the canonical Moore sentence lacks. Except for our desire to belief all true sentences, we have no readily available incentive to believe 'It's raining but I do not believe it', as it does not seem pertinent to us to have accurate beliefs about the weather and about our beliefs thereof (unless we happen to be preparing for a long outside walk or bike trip). But the teacher's announcement immediately engages our imagination and our memory of past exams, and recruits our association between (surprise) exams and studying properly beforehand, so that we can feel the importance of holding accurate beliefs about the teacher's announcement. So even though the canonical Moore sentence and the teacher's announcement are equally unbelievable, not being able to believe the teacher's announcement may *feel* worse than not being able to believe 'It is raining but I do not believe it', as incorrect beliefs about (our beliefs about) the weather are easier to accept and live with then incorrect beliefs about (our beliefs about) upcoming exams, because exams matter more than the weather.

More Complex

Second, the announcement of the teacher is a more complex Moore sentence than the canonical one, in the sense that it takes more effort to reveal its Moorean nature. In a sense, one is required to go through several reasoning steps to discover that the announcement cannot be believed, whereas this is immediately clear in the canonical sentence. Because the Moorean nature of the announcement is in this way obscured, one may also be more surprised when one finds out that the announcement cannot be believed, then one the canonical sentence cannot be believed.

No Justification

Third, consider the conjuncts p and $\mathrm{B}\, p$, and an agent A who is trying to determine which of these conjuncts he should believe. A can choose to believe one of the following four conjunctions:

i. $p \wedge \mathrm{B}\, p$
ii. $\neg p \wedge \neg \mathrm{B}\, p$
iii. $p \wedge \neg \mathrm{B}\, p$
iv. $\neg p \wedge \mathrm{B}\, p$

As we know, A cannot believe iii and iv are, because these are Moore sentences (assuming his beliefs are consistent, distribute over conjunction and that he is capable of positive introspection). This leaves i and ii, which both seem to be reasonable options if p is a placeholder for a sentence such as 'it is raining': either A believes that it is raining and that he believes that it is raining, or A believes that it is not raining and that he does not believe it is raining.

However, now consider the situation in which A is informed of the teacher's announcement. To simplify the discussion, consider the situation in which the exam has not been given on days one through four, so that the teacher's announcement is reduced to $d5 \wedge \neg \mathrm{B}\, d5$. Upon being told by the teacher that $d5 \wedge \neg \mathrm{B}\, d5$, A tries to determine which of the conjuncts $d5$ and $\mathrm{B}\, d5$ to believe. Suppose that, as is suggested in the description of the surprise exam paradox, A has no justification to believe that $d5$ is true except the teacher's announcement of $d5 \wedge \neg \mathrm{B}\, d5$.

In this case, none of the options i through iv are particularly satisfying for agent A. Again, options iii and iv are unbelievable for A. This leaves options i and ii. In this case, believing i amounts to believing that there will be an exam, but the exam will not be surprising. Of course it is possible for A to choose to believe this, but it seems that this belief cannot be justified in the situation. The only justification that the A has to believe that there will be an exam is that the teacher told him that $d5 \wedge \neg \mathrm{B}\, d5$. If A does not believe this announcement, he thus has no justification to believe $d5$, and consequently no justification to believe i. On the other hand, if A does believe the teacher's announcement, then he should not believe i, as the teacher's announcement of $d5 \wedge \neg \mathrm{B}\, d5$ directly contradicts i, which says that $d5 \wedge \mathrm{B}\, d5$. So it seems that, irrespective of whether A believes the teacher's announcement or not, A is not justified to believe i.

This leaves ii as the last option for A to believe. If we assume that exams are given rather infrequently in the school of the surprise exam paradox, then A seems to have a justification to believe that there will not be an exam next week, so he can be justified in believing ii. It thus seems that the only reasonable thing to do for A in this situation is to believe ii. But, for A, this is deeply unsatisfying, because he realizes that his believing ii is consistent with the announcement of the teacher being true. A realizes that if he believes ii then the teacher can make his announcement true after all by giving the exam. The only reasonable thing for A to do in this case is thus to believe that there will not be an exam, and accept that he will be caught off guard if the teacher decides to make his announcement true after all.

4.2 The Knowability Paradox

Another paradox in which a Moore sentence is known to be involved is *the knowability paradox*, a paradox in epistemology discovered by Frederic Fitch.[17] The paradoxical result that Fitch found is that if it is possible to know all truths, then all truths are already known. Or contrapositively, if there are unknown truths, then there are truths that cannot be known. This result has been used to argue against the *verification thesis*, which states that all truths can be known, a thesis defended by a strand of metaphysical anti-realism.[18]

In this section we illustrate that Moore's paradox is involved in the knowability paradox. This in itself is a rather trivial result, as a knowledge variant of the traditional omissive Moore sentence $p \wedge \neg \mathrm{B}\, p$ is explicitly used is the proof, but it supports one of the main claims of this thesis, which is that Moore's paradox is involved in many existing paradoxes. Also, we discuss in what ways the Moore sentence involved in the knowability paradox differs from the canonical Moore sentences. Further, we distinguish between two interpretation of the verification thesis that all truths can be known, corresponding to two phases in learning. We argue

[17] Fitch, 1963.

[18] Hart and McGinn, 1976.

that the knowability paradox is an argument against the verification thesis under one interpretation of the thesis, but not on another interpretation, and the latter interpretation is the one that epistemilogists should care about.

4.2.1 Introducing the Knowability Paradox

The knowability paradox is usually presented as a derivation from two assumptions. The first assumption is that the anti-realist verification thesis holds. The verification thesis states that all truths are knowable and can be expressed formally as $\phi \rightarrow \Diamond \mathrm{K}\,\phi$, where $\Diamond$ is the dual of $\Box$.[19] The second assumption is that there exists at least one truth that is unknown. This assumption can be expressed as there being some p for which $p \wedge \neg \mathrm{K}\,p$ is true.[20]

The derivation of the knowability paradox is thus performed in a modal logic with metaphysical and epistemic modal operators. The metaphysical modality is constrained at least so that all validities are necessarily valid ($\models \phi \Rightarrow \models \Box\phi$), and the epistemic modality at least so that knowledge distributes over conjunction and is veridical.[21] The derivation of the proof of the knowability paradox can best be presented in two separate derivations. The first derivation shows that it is impossible to know a Moore sentence ($\neg\Diamond\,\mathrm{K}(\phi \wedge \neg \mathrm{K}\,\phi)$). In section 3.4.3 we provided a derivation to prove that $\mathrm{K}(p \wedge \neg \mathrm{K}\,p)$ is unsatisfiable under the assumptions that knowledge distributes over conjunction and is veridical. This implies that $\neg \mathrm{K}(p \wedge \neg \mathrm{K}\,p)$ is valid under these assumptions. Assuming that validities are necessarily valid, this implies that $\Box\neg \mathrm{K}(p \wedge \neg \mathrm{K}\,p)$ is valid. This implies that $\neg\Diamond\,\mathrm{K}(p \wedge \neg \mathrm{K}\,p)$ is valid, which means that it is impossible to know an epistemic Moore sentence.

The second derivation in the proof of the knowability paradox is to show that, from the assumptions mentioned above, it can be derived that it *is* possible to know an epistemic Moore sentence:

1. $\phi \rightarrow \Diamond \mathrm{K}\,\phi$ (assumption)
2. $p \wedge \neg \mathrm{K}\,p$ (assumption)
3. $(p \wedge \neg \mathrm{K}\,p) \rightarrow (\Diamond\,\mathrm{K}(p \wedge \neg \mathrm{K}\,p))$ (substitution, 1)
4. $\Diamond\,\mathrm{K}(p \wedge \neg \mathrm{K}\,p)$ (modus ponens, 2, 3)

The results of these two derivation contradict each other. Thus, assuming that knowledge distributes over conjunction and is veridical and that valid sentences are necessarily valid, a contradiction can be derived from the assumptions that all truths are knowable and that there is at least one unknown truth. Therefore, one of these assumptions needs to be rejected. As mentioned earlier, the assumptions that knowledge distributes over conjunction and is veridical and that logically valid sentences are necessarily valid are standard in (epistemic) modal logic, and are thus typically not rejected.[22] And since it seems highly implausible that there are no unknown truths, most epistemologists choose to reject the assumption that all truths are knowable.

[19] Fitch, 1963.

[20] Fitch, 1963.

[21] Fitch, 1963.

[22] Meyer, 2003, Berto and Plebani, 2015.

4.2.2 Moore Sentence in the Knowability Paradox and Twists

The derivation of the knowability paradox makes use of the sentence $\phi \wedge \neg K \phi$. As we showed in section 3.4.3, this is a Moore sentence. It is thus clear that Moore sentences play a key role in the knowability paradox.

Further, the result of the knowability paradox is similar to that of the canonical Moore sentence. The result of the canonical Moore sentence is that there are sentences that can be true but that cannot be believed (and since knowledge is usually taken to imply belief, this implies that there are sentences that can be true but that cannot be known).[23] The result of the knowability paradox is that not all truths can be known. Although these results are quite similar, the result of the knowability paradox differs from that of the canonical Moore sentence in at least two respects. Firstly, in the knowability paradox the existence of true Moore sentences is postulated. Secondly, in the knowability paradox the result of the canonical Moore sentence is considered from a wider perspective.

Postulating True Moore Sentences

In the knowability paradox the existence of true Moore sentences is postulated. Whereas Moore and subsequent philosophers merely wondered about the fact that sentences like 'it is raining but I do not believe it' cannot be believed, even *if* they are true, the knowability paradox postulates that there are such sentences that *in fact* are true. Specifically, it postulates that there is a true sentence such that we do not know it, i.e. a sentence p such that $p \wedge \neg K p$ is true. This postulate constitutes a completely plausible assumption, as nearly anyone would concede that humans collectively do not know all truths. Still, this assumption ups the stakes of Moore's paradox, because its result is transformed from there being unbelievable sentences that *can be true*, to there being unbelievable sentences that *are true*. This result may be more problematic than the original result, as it seems more forgivable to be incapable of believing sentences that can be true than to be incapable of believing sentences that are true.

A Wider Perspective

Further, in the knowability paradox the result of the canonical Moore sentence is considered from a wider perspective. This wider perspective is achieved by phrasing the result of the paradox as a universal statement rather than an existential statement. That is, instead of saying that there exist sentences that cannot be known, it is said that not all truths can be known. This latter phrasing puts the result in a wider perspective, in the sense that the universally stated result makes the consequences of this result for science and epistemology more readily available. If not all truths can be known, then advancements in human knowledge must come to a halt before reaching their ultimate goal (at least if we conceive of their goal as exhaustively listing all true sentences). Progress in human knowledge must eventually run into an impenetrable wall made up of unknowable true sentences, after which further inquiry will be futile. Or at least, this is what suggested by those who take the knowability paradox to be a refutation of the verification thesis that all truths can be known.[24]

[23] Cf. Williamson's knowledge first account Williamson, 2002.

[24] Hart and McGinn, 1976

4.2.3 Distinguishing Two Phases in the Knowability Paradox

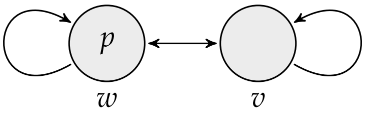
In the introduction we claimed that there are two different interpretations of the anti-realist claim that *all truths* can be known. If we think about knowledge in a dynamic sense, these two interpretation correspond to two phases, the phase before we learn new information and the phase after it. We can interpret 'all truths' to refer to all sentences that are true now, or we can interpret it to refer to all sentences that are true after we learn new information. This distinction is relevant in cases when sentences change truth values based on our learning of new information. This turns out to be the case for the Moore sentence used in the knowability paradox. Therefore this distinction should be kept in mind when we think about the consequences of the knowability paradox for the verification thesis that all truths can be known.

To illustrate this, it is useful to enrich the logical language we have been working with so far. Let's introduce a *public announcement operator* into our language: $[\phi!]$. Public announcement operators are used in dynamic logics to capture the interaction between agents' belief bases and new incoming information in the form of public announcements.[25] Public announcement operators, as they are typically used, reduce the domain W of a model M as follows:

$$W[\phi!] = \{w \in W | M, w \models \phi\}$$

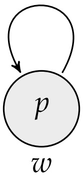
The public announcement $[\phi!]$ can thus be understood as eliminating all worlds in the model M where ϕ is not true, indicating that the belief base modeled by M is updated with the information that ϕ is true.

Consider the following epistemic model M with two worlds w and v such that wRw, wRv, vRw and vRv, and let p be true in w, which is the actual world:



Let Σ be the set of all sentences that are true in world the actual world w in M. Σ then contains $p \wedge \neg \mathrm{K}\, p$, which is a Moore sentence, and importantly is an unknowable truth. In this model the claim that all truths can be known is thus false, if we interpret 'all truths' as all sentences that are currently true.

However, consider what happens if the agent learns that p is true, for example via a public announcement. That is, suppose that the computation $W[p!]$ is performed on M. Then the resulting model M' is:



[25] Baltag and Renne, 2016.

Let Ω denote the set of all sentences that are true in w in M'. Ω then contains p, $\mathrm{K}\,p$, $\mathrm{K}\,\mathrm{K}\,p$, and so forth. Clearly, Ω does not contain any unknowable sentences. Since M' is the result of a simple learning action being performed on M, there is an important sense in which it is true that all truths can be known in M. If we interpret 'all truths' as all sentences that are true after we learn that p, then all truths can be known.

From this is should be clear that the verification thesis that all truths can be known, is attacked by the knowability paradox only in a limited sense. The result of the knowability paradox is that it is impossible to know all sentences that are currently true (given that not all sentences are already known). But, the knowability paradox does not prove that there it is impossible to perform a learning action (or a sequence of learning actions), so that all sentences that are true after the performance of this action are known. Therefore, anti-realists can endorse the claim that all truths can be known in spite of the knowability paradox, as long as they interpret this thesis as follows: it is possible to perform a learning action (or sequence of learning actions) after which all sentences that are then true are known.

4.2.4 Van Benthem's Solution to the Knowability Paradox

In the previous section we already downplayed the importance of the result of the knowability paradox by pointing out the limited sense in which it is a problem for anti-realists. Van Benthem has argued that the result of the knowability paradox is not only not a problem, it is actually an interesting result about human communication.[26]

Van Benthem introduces a dynamic reading of the verification thesis: "What is true may come to be known".[27] With this dynamic reading van Benthem shifts the discussion from what can be known to what can be learned. To speak of learnability in formal terms, van Benthem uses public announcement logic.[28] Van Benthem defines a *learnable proposition* ϕ as a proposition for which there exists a true sentence ψ such that ϕ becomes known after the public announcement of ψ.[29]

At first sight it may seem that under this definition, every true proposition is learnable. One may suspect that for every true proposition ϕ, one can simply publicly announce ϕ so that ϕ becomes known. That is, one may suspect that what van Benthem calls the *Learning Principle (LP)* holds:

> "Announcing ϕ publicly in a group G makes ϕ common knowledge, or in a dynamic-epistemic formula: $[\phi!]C_G\phi$"[30,31]

But, it turns out that not all propositions become known upon being announced. For example, if the epistemic Moore sentence $p \wedge \neg\,\mathrm{K}\,p$ is announced, it does not become known. This is apparent, because after the announcement of $p \wedge \neg\,\mathrm{K}\,p$, p is true

[26] Van Benthem, 2004.

[27] Van Benthem, 2004, 96.

[28] Van Benthem, 2004, 97.

[29] Van Benthem, 2004, 101.

[30] Van Benthem, 2004, 99.

[31] Common knowledge of ϕ in a group G means that every agent in group G knows ϕ, and every agent in G knows that every agent in G knows ϕ, and every agent knows that every agent knows that every agent knows that ϕ, and so forth.

globally, and thus $\neg \mathrm{K}\, p$ is false globally. Van Benthem writes about this observation that:[32]

> "Upon reflection, the Learning Principle just seems too hasty an assertion, and the given counter-example [of an epistemic Moore sentence] seems very natural. Indeed, announcements of ignorance are not just philosophical conundrums. They are made frequently, and they can be very useful."

In van Benthem's interpretation, the result of the knowability paradox is that the Learning Principle, and its static counterpart the verification thesis, are too hasty. As it turns out, not all truths can be learned or known. In particular, Moore sentences cannot be learned or known. But this does not mean that Moore sentences cannot be informative. Moore sentences are declarations of our ignorance; they describe that something is the case and that we do not know it. Naturally, when we 'learn' a Moore sentence, we do not learn that we are ignorant about a particular fact; we learn that we *were* ignorant to a particular fact. Immediately upon 'learning' a Moore sentence, which expresses that a particular fact is true but that we are ignorant about that fact, we stop being ignorant about that fact. Moore sentences thus have a non-straightforward communicative role: we do not learn these sentences themselves, we learn the fact that they describe our ignorance of. As I'm told that I do not know that it is raining even though it is, I learn that it's raining.

The interesting role of Moore sentences in communication forces us to reconsider the naive verification thesis that all truths can be known. Clearly, some true sentences are such that they cannot be known, because some true sentences can serve a more complicated communicative goal than simply informing one about the truth of that sentence. But this failure of the classical verification thesis should not be considered too problematic; rather it should be understood as the starting point of an inquiry into human knowledge that takes into account the intricate roles that some sentences can play in communication.

4.3 Conclusion

In this chapter we discussed under which assumptions it can be said that Moore sentences are involved in the surprise exam paradox and the knowability paradox. We contributed to a long standing debate surrounding the surprise exam paradox, about which sentence exactly plays the role of a Moore sentence. We also reconsidered the consequences of the knowability paradox for the verification thesis that all truths can be known, and argued that under one interpretation of 'all truths' the knowability paradox is an argument against this claim, but under another interpretation it is not.

In the next chapter we draw upon some of the results presented in the current chapter. We discuss the Toxin problem and Newcomb's problem and the role Moore sentences play in them, and introduce two new paradoxes that are inspired by the knowability paradox. Further, we suggest how solutions that are modeled after van Benthem's solution to the knowability paradox can be used to solve these two new paradoxes. And, we suggest how the distinction of two phases we made in the knowability paradox can be used to address these new paradoxes.

[32] Van Benthem, 2004, 99.

Chapter 5

The Toxin problem and Newcomb's Problem

In this chapter we discuss the Toxin problem and Newcomb's problem. Both problems are similar in that everyone agrees that they have simple, even trivial solutions, with half the people agreeing on one solution and the other half agreeing on another. In this chapter we contribute to the long-standing debates on these problems. Firstly, by showing that in both problems, a Moore sentence is involved. Secondly, by providing new variations of both problems that pose challenges that so far have not been addressed in the literature. Thirdly, we introduce two paradoxes related to the Toxin problem, Newcomb's problem and the knowability paradox, and also suggest how these can be addressed.

5.1 The Toxin Problem

The Toxin problem, introduced by Kavka, concerns the following situation:[1]

> A billionaire comes up to you and offers you a challenge. He shows you a small bottle and explains that it contains a poison that induces a day long illness but does not cause any chronic averse health effects. The billionaire offers to give you one million euro if at midnight tonight you have the *intention* to drink the poison tomorrow morning (assume that he has a device that accurately measures your intentions in real time, such as a high tech brain scanner). He stresses that it does not matter whether you actually drink the poison tomorrow or not, you need only to have the intention to do so tonight at midnight.

Kavka notes that being offered this challenge may seem as good fortune. Even being offered the challenge to *drink* the poison for one million euro would have been good fortune, as obtaining a million euro certainly is worth being ill for a day. The current challenge may seem even more delightful, as a million euro can be earned without even drinking the poison. As the billionaire stresses, the poison does not need to be drunk to obtain the million, only the intention to do so needs to be formed. A clever agent could thus simply intend to drink the poison tonight, and change his mind tomorrow, to obtain one million euro without becoming ill. Seemingly, this challenge is thus all upside and no downside.

But, Kavka points out that the billionaire's challenge has a catch. The billionaire issues his reward when the challenge is completed, directly after midnight. And if

[1] Kavka, 1983.

the agent at midnight intended to drink the poison, he would not have any reason to carry out his intention to drink the poison tomorrow morning. After midnight, drinking the poison only has a substantial downside, namely becoming ill for a day. It seems that a rational person, after forming the intention to drink the poison, would never follow up on his intention and choose to actually drink the poison.

This is problematic, because presumably a rational person also foresees that even if he tonight has the intention to drink the poison tomorrow, he will not follow up on his intention tomorrow. The agent thus knows that he will never drink the poison. Kavka suggests that this implies that the agent cannot form the intention to drink the poison in the first place, because an intention from which it is known beforehand that it will not be carried out is not a real intention. Thus, Kavka suggests that it can be doubted whether a rational agent can form the intention to drink the poison in the Toxin problem. It may be that, if the agent is rational, he cannot complete the billionaire's challenge.

Kavka concludes that the billionaire's challenge is rather devious; initially it seems to be a blessing that the reward is issued for merely intending to drink the poison rather than actually drinking it, but further consideration reveals that this might make the challenge impossible to fulfill for rational agents.

5.1.1 Similarities between the Toxin Problem and Moore's Paradox

Because the Toxin problem concerns intentions rather than belief, it is not generally considered to be related to Moore's paradox. However, Goldstein and Cave argue that it is.[2] They illustrate this by formalizing what they call the *jackpot scenario*, i.e. the scenario that is most preferable for the agent who is issued the challenge of the billionaire in the Toxin problem. They let p read as 'I drink the poison tomorrow' and $\mathrm{I}\phi$ as 'I intend to make ϕ true', and formalize the sentence that describes the jackpot scenario as follows:

(1) $\mathrm{I}\,p \wedge \neg p$

The sentence (1) describes the jackpot scenario because it describes the situation in which the reward is obtained, since the intention to drink the poison is formed, but the agent does not become ill for a day because he does not actually drink the poison. Moreover, they remark that (1) describes a situation that is logically possible, as it is possible for someone to have an intention and subsequently fail to realize it. But, they add, although (1) describes a possible situation, the following sentence does not:

(2) $\mathrm{I}(\mathrm{I}\,p \wedge \neg p)$

Goldman and Cave claim that the situation described by (2) is not possible because it is what they call *deliberatively unstable*. Goldstein and Cave do not define when a situation is deliberatively unstable, but they describe it as "a situation where the choice to p is trumped by the choice to not-p, which is then trumped by the choice to p, and so on, *ad infinitum*".[3] They also provide a reference to an article from Sober, in

[2] Goldstein and Cave, 2008.

[3] Goldstein and Cave, 2008, 3.

which he describes deliberatively instable situations in multi-agent decision problems as decision problems without a *Nash equilibrium*.[4] A Nash Equilibrium is a concept in game theory, which denotes a strategy profile for which no player in the game can earn a higher payoff by deviating from this strategy profile (given that the other agents also do not deviate from this strategy profile).

For example, consider an open information version of the decision problem *rock, paper, scissors* with two agents A and B, which is deliberatively unstable. Both agents can choose either of three actions, rock, paper or scissors, and both players choose their action simultaneously. Further, in this open information version of rock, paper, scissors, both agents announce which action they choose before actually choosing it. Now, if A announces that he will choose rock, then B will choose paper. But, when B announces that he will choose paper, A will change plans and choose scissors instead. But when A tells B that he switched plans and will now choose scissors, this again prompts B to switch plans too and choose rock. And so forth, ad infinitum. Open information rock, paper scissors is thus deliberatively instable (as Sober uses this term) in the sense that every time one agent settles on a strategy, this gives the other agent a reason to change his strategy.

We take it that Goldstein and Cave suggest that a similar phenomenon occurs in the single agent setting of the Toxin problem. The suggested similarity seems to be as follows. Suppose that the agent plans to drink the poison. Then, he realizes he successfully formed the intention to drink the poison. Since he will obtain the reward for merely having the intention to drink the poison rather than for actually drinking it, he decides he no longer needs to actually drink the poison. Therefore, he plans not to drink the poison after all. But, by revising his plans, he loses the intention to drink the poison, and with that also the prospects of obtaining the reward. Therefore, he again plans to drink the poison. This way, the agent goes back and forth between planning to drink the poison and planning not to. This situation resembles Goldstein and Cave's description of a deliberatively unstable situation as a "a situation where the choice to p is trumped by the choice to not-p, which is then trumped by the choice to p, and so on, *ad infinitum*".[5]

We are not sure that the deliberatively instable situation described by Goldstein and Cave in the single agent setting is completely analogous with the deliberatively instable situation described by Sober in the multi-agent setting in this case, but this need not bother us at this point. The important thing is that we can note that Goldstein and Cave recognize that although (1) is possible, (2) is problematic. And, more importantly for this thesis, Goldstein and Cave remark that (2) is deliberatively unstable in the same manner as a canonical Moore sentence.

We should make a clarifying remark on the syntactic similarity of (1), (2) and the canonical Moore sentences. Goldstein and Cave claim that (1) is syntactically similar to the *commissive* Moore sentence $p \wedge \mathrm{B}\,\neg p$. However, (1) is actually more similar to the *omissive* Moore sentence $p \wedge \neg\,\mathrm{B}\,p$. This is apparent if in (1) we substitute p for $\neg p$, eliminate double negation and switch the conjuncts around:

[4] Sober, 1998.

[5] Goldstein and Cave, 2008, 3.

(3) $p \wedge \mathrm{I}\,\neg p$

Further, Goldstein and Cave do not substantiate their claim that (1) is similar to a Moore sentence, and seem to rely on the assumption that these sentences are similar in a substantial sense because they are syntactically similar.[6] But, as we discussed in Chapter 3, the fact that a sentence is similar in syntactic structure to a canonical Moore sentence, does not guarantee that it is also a Moore sentence. According to our definition, a sentence ϕ is a Moore sentence to an agent A iff ϕ contains a modal operator $\Box_A$, ϕ is satisfiable on a class of Kripke models K appropriate for modeling $\Box_A$ and $\Box_A\phi$ is unsatisfiable on K. In the following section we discuss what an appropriate class of Kripke model to model I might look like, so that we can determine whether (1) is similar to $p \wedge \mathrm{B}\,\neg p$ in a substantive sense.

5.1.2 Formal Properties of Intentions

In this section we present some formal properties that intention may have. Our goal is not to argue for a particular account of intentions. Rather, we aim to present some properties that intention may have, so that we can refer back to these properties when we discuss under what assumptions about intention we can say that a Moore sentence is involved in the Toxin problem in the next section.

The formal properties of intentions have been studied extensively, and many different formalisms of intention exist.[7] In this thesis we take intention to be a modal operator I_A that ranges over sentences. We read $\mathrm{I}_A\,\phi$ as 'agent A intends to make ϕ true'. This is in contrast with for example a proposal of Cohen and Levesque to take I_A to be an operator that ranges over actions (modeled as propositions), and read I_A as 'agent A intends to do action ϕ'.[8] Cohen and Levesque note that their account of intentions is not apt to deal with all intentions agents may have. For example, one may intend to become rich, but becoming rich can hardly be describes as an action (especially if the agent in question has no concrete ideas about how he will become rich).[9] For these cases, they propose to take I_A as an operator that ranges over sentences (that describe the state of affairs that the agent intends to bring about). $\mathrm{I}_A\,\phi$ is then read as 'agent A intends to bring about ϕ', which is synonymous with our reading.

One of the properties of intention that is mentioned frequently in such terms, is *internal consistency*: agents should not intend to bring about contradictions ($\models \neg\,\mathrm{I}\,\bot$).[10] Other principles that are mentioned frequently concern the relation between intention and belief. For example, it is often mentioned that if an agent intends to make ϕ true, then the agent believes that it is possible for ϕ to be true, and does not believe that he will not make ϕ true.[11] In general, it is important to be aware of the relation between beliefs, intentions and even desires when dealing with intentions.[12]

[6] Goldstein and Cave, 2008, 366.

[7] Roy, 2008, Shoham, 2009 and Lorini and Herzig, 2008.

[8] Cohen and Levesque, 1990, 245.

[9] Cohen and Levesque, 1990, 247.

[10] Roy, 2008, 139, Shoham, 2009, 639.

[11] Cohen and Levesque, 1990, 218.

[12] So-called 'BDI' logics (which stands for 'Belief, Desire and Intention' logics) aim to deal specifically with these relations. For for a brief introduction to BDI logics see Herzig et al., 2017, for a more encompassing introduction see Georgeff et al., 1998.

Two other formal principles are collection of intention over conjunction ($(\mathrm{I}\,\phi \wedge \mathrm{I}\,\psi) \rightarrow \mathrm{I}(\phi \wedge \psi)$) and distribution of intention over conjunction ($\mathrm{I}(\phi \wedge \psi) \rightarrow (\mathrm{I}\,\phi \wedge \mathrm{I}\,\psi)$). Bratman argues for the importance of *agglomerativity*, the act of putting intentions together into larger intentions. Bratman writes that "Given the role of intentions in coordination, there is rational pressure for the agent to put his intentions together into a larger intention."[13] Formally, agglomerativity can be expressed as collection of intention over conjunction (but, cf. Yaffe who argues that collection of intention over conjunction should only be required of agents in cases when this collection can expose inconsistencies between different intentions).[14]

Given the importance of putting intentions together into larger intentions to facilitate coordination that Bratman describes, one may be weary of accepting the distribution of intention over conjunction, which describes the breaking up of large intentions into smaller intentions. Further, one could argue against distribution of intention over conjunction that it is possible for an agent to intend to make true a conjunctive sentences, without intending to make true either sentence individually. For example, I might intend to make it true that I get a high salary job and retire early. However, from this it should not be inferred that I have the individual intention to make it true that I retire early; for if I do not get a high salary job, I might not have the money to sustain myself through a prolonged retirement. Nor should it be inferred that I have the individual intention to make it true that I get a high salary job; because if I know that I somehow will not have the opportunity to retire early, I might choose to take a more enjoyable job with a lower salary. I thus only have the intention to make true both that I get a high salary job and that I retire early, I do not have intentions to make either of these true individually.

We are not entirely convinced by this argument. Even if we grant that the objection above is sound, it seems to follow only that there are special cases in which it is improper to say of an agent who has an intention to make $\phi \wedge \psi$ true, that he intends to make ϕ true and also intends to make ψ true. These special cases seem to be when the intention to $\phi \wedge \psi$ is such that the intention to ϕ is dependent on the (expectation of the) success of the intention to ψ; I intend to make it true that I get a high salary job only if I (expect that I) can also make it true that I retire early, and vice versa. Clearly, intentions to make conjunctions true do not always involve conjuncts of which the individual intention are in this way conditional on each other. For example, I can intend to make it true that I am at school at 12AM and that I eat dinner at 6PM. If I now learn that the school is closed today, so that I can no longer (expect to) make true the first conjunct of my intentions, this need not change the fact that I intend to make it true that I eat dinner at 6PM. My intention to make it true that I have dinner at 6PM is not dependent on my intention that I am at school at 12AM, and vice versa. Since the intentions in this example are not conditional on each other's (expected) success, it also seems reasonable to say that I do have the individual intention to be at school at 12AM and the individual intention to eat dinner at 6PM. Since the objection presented above only seems to point out that there is a set of special cases in which distribution of intention over conjunction fails, there seems to be room for a restricted version of distribution of intention over conjunction. An example of such a restricted principle could be that intentions distribute over conjunction, provided

[13] Bratman, 1987, 134.

[14] Roy, 2008, 139, Yaffe, 2004, 511-512.

that the intentions towards the conjuncts are not conditional on the (expected) success of the other conjuncts.

For our discussion of the Toxin problem the question then is whether the conjuncts of the agent's intention are conditional on each other. We argue that they are not. The first conjunct describes that the agent has the intention to make it true that he drinks the poison. The second conjunct describes that the agent does not drink the poison. The agent intends to make the first conjunct true regardless of whether (he expects that) he will make the second conjunct true; because making the first conjunct true is what the agent will obtain the million euro reward for. And the agent intends to make the second conjunct true regardless of whether he succeeded in making the first conjunct true. If the agent did succeed in making the first conjunct true, then the agent obtains the jackpot scenario by making the second conjunct true. If the agent did not succeed in making the first conjunct true, the agent did not obtain the reward, and certainly has no reason to make the second conjunct true. We think it is therefore fair to say that in the Toxin problem, the intention to obtain the jackpot scenario distributes over conjunction, so that the agent has an individual intention to make it true that he has an intention to drink the poison, and an individual intention to make it true that he does not drink the poison.

Aside from the principle of internal consistency that we mentioned earlier, a principle of *external consistency of intentions* can be considered: an agent who intends to make ϕ true, does not also intend to make $\neg\phi$ true ($\mathrm{I}\,\phi \rightarrow \neg\,\mathrm{I}\,\phi$).[15] This principle can actually be derived from the assumptions of internal consistency and collection of intention over conjunction. To see this, consider that if an agent intends to make ϕ true and also intends to make $\neg\phi$ true, by collection of intentions over conjunction he intends to make $\phi \wedge \neg\phi$ true, which violates internal consistency of intentions. Therefore, an agent cannot intend to make ϕ true and also intend to make $\neg\phi$ true if intentions are internally consistent and collect over conjunction.

Another principle concerning intentions is the principle of *success of second order intentions*: if an agent intends to make true the sentence that describes that he intends to make ϕ true, then he intends to make ϕ true ($\mathrm{I}\,\mathrm{I}\,\phi \rightarrow \mathrm{I}\,\phi$).[16] This principle is not discussed in the literature as far as we know, but it describes an interesting idea. It describes the idea that agents are always successful in fulfilling their intentions if these intentions concern their own intentions (we could say more poetically that agents are omnipotent when it comes to their intentions to bring about other intentions). More precisely it describes the idea that if an agent succeeded in forming the intention to form the intention to make ϕ true, he in fact already formed the intention to make ϕ true, because the agents intentions respond immediately to his second order intentions.

Note that the principle of success of second order intentions describes that agents form any intention that they intend to form, but not necessarily any intention that

[15] This principle is endorsed for example by Roy, 2008, 139-141, and Cohen and Levesque, 1990, 250.

[16] Second order intentions are described by sentences in which an I operator that is not in the range of another I operator, ranges over a (sub-)sentence that contains another I operator, such as $\mathrm{I}\,\mathrm{I}\,\phi$ or $\mathrm{I}(\psi \vee \mathrm{I}\,\chi)$. First order intentions are described by sentences in which a I operator that is itself not in the range of another I operator ranges over a (sub-)sentence that contains no other I operators, such as $\mathrm{I}\,\phi$ or $\mathrm{I}(\psi \wedge \chi)$. Note that for example the sentence $\mathrm{I}\,\phi \wedge \mathrm{I}\,\mathrm{I}\,\chi$ describes both a first order intention and a second order intention.

they *want* to form. That is, we should distinguish the principle of success of second order intentions from the principle of *fulfillment of desires to intend*: ($\mathrm{D}\,\mathrm{I}\,\phi \rightarrow \mathrm{I}\,\phi$). This latter principle may also be true, but to claim this one would need to independently argue for this claim. We think it is important to distinguish these two principles because, if they are conflated, objections against fulfillment of desires to intend may be taken as objections against success of second order intentions. In particular, we anticipate the objection that in everyday life, we often fail to have intentions that we would like to have. Consider the example of man on the couch who would like to have the intention to go and work out, but unfortunately does not have that intention. This example is an objection against the fulfillment of desires to intend, but not necessarily against the success of second order intentions. In order for the example to be an objection to success of second order intentions, it would need to be argued that the man on the couch successfully formed the intention to intend to go and work out, and still does not have the intention to go and work out.

Nevertheless, we are not convinced that the principle of success of second order intentions is unproblematic. In particular, we think that Williamson's objection from time delay against positive introspection of knowledge and beliefs, can be raised against the principle of success of second order intentions as well. It seems that even if the idea behind the principle is correct and agents always succeed in converting a second order intention into a first order intention, this conversion must take some time. If an agent formed the intention to make $\mathrm{I}\,\phi$ true, he arguably still needs some time to make $\mathrm{I}\,\phi$ true. After all, the forming or changing of mental states is a process that happens over time, rather than instantaneously. Therefore, it seems that there can at least be short moments in which an agent has the intention to make $\mathrm{I}\,\phi$ true, but did not yet form the intention to make ϕ true. In these moments the principle of success of second order intentions fails.

Another principle concerning intentions, one that reminds one of positive introspection, is that if one intends to make ϕ true, one also intends to make true that one intends to make ϕ true ($\mathrm{I}\,\phi \rightarrow \mathrm{I}\,\mathrm{I}\,\phi$). Whether intentions are characterized by this principle is debatable, although to us it seems that they are not. In fact, it can be argued that intentions may be characterized by the opposing principle that if one intends to make ϕ true, one does not intend to make true that one intends to make ϕ true ($\mathrm{I}\,\phi \rightarrow \neg\,\mathrm{I}\,\mathrm{I}\,\phi$). This principle we call *anti introspection for intentions*. A possible argument to support this principle is that it seems plausible to assume that one can only intend to make true sentences that one does not believe to be true already ($\mathrm{I}\,\phi \rightarrow \neg\,\mathrm{B}\,\phi$).[17] If we combine this assumption with the assumption that if an agent intends to make ϕ true, he also believes that he intends to make ϕ true ($\mathrm{I}\,\phi \rightarrow \mathrm{B}\,\mathrm{I}\,\phi$), then we can derive that $\mathrm{I}\,\phi \rightarrow \neg\,\mathrm{I}\,\mathrm{I}\,\phi$. To see this, suppose that an agent intends to make ϕ true. Then he also believes that he intends to make ϕ true. Since be believes it is already the case that he intends to make ϕ true, he cannot intend to make it true that he intends to make ϕ true, because one cannot intend to make true sentences that one believes are true already.

We should note, that it may be wise not to adopt both the principles success of second order intentions and anti introspection for intentions, as together they trivialize second order intentions. If both these principles are assumed, an agent cannot have a second order intention. To see this, suppose for contradiction that an agent

[17] Cohen and Levesque, 1990, 234.

has the second order intention $\mathrm{I}\,\mathrm{I}\,\phi$. Then, by the success of second order intentions, $\mathrm{I}\,\phi$ is true. But, then by anti introspection for intentions, this implies that $\neg\,\mathrm{I}\,\mathrm{I}\,\phi$ is true. This contradicts our initial assumption, so an agent cannot have the second order intention $\mathrm{I}\,\mathrm{I}\,\phi$.

It is important to repeat that we do not necessarily endorse any of the principles concerning intention mentioned in this section. We will only make use of these principles to make clear in the following section under which assumptions it can be said that a Moore sentence is involved in the Toxin problem.

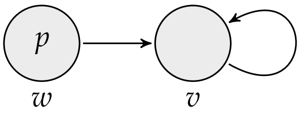
5.1.3 Moore Sentence in the Toxin Problem

As mentioned earlier, whether the sentence that represents the jackpot scenario in the Toxin problem, $p \wedge \mathrm{I}\,\neg p$, is a Moore sentence depends on our assumptions about the formal properties of intentions. First we should note that if we assume that if intentions are characterized by distribution over conjunction, consistency and positive introspection for intentions $p \wedge \mathrm{I}\,\neg p$ is a Moore sentence. For the proof we direct the reader to section 3.4.1, where we proved that $p \wedge \mathrm{B}\,\neg p$ is a Moore sentence if we assume that belief is characterized by distribution over conjunction, consistency and positive introspection.

In the previous section we mentioned the principle of success of second order intentions ($\mathrm{I}\,\mathrm{I}\,\phi \rightarrow \mathrm{I}\,\phi$). Assuming, that this principle holds, and intentions are further characterized at least by consistency and distribution over conjunction, $p \wedge \mathrm{I}\,\neg p$ is a Moore sentence. We prove this by proving that (i) $p \wedge \mathrm{I}\,\neg p$ is satisfiable on a class of Kripke models characterized by seriality and density (for all worlds w and v such that wRv, there is a world u such that wRu and uRv), and (ii) $\mathrm{I}(p \wedge \mathrm{I}\,\neg p)$ is unsatisfiable, if intentions are characterized at least by consistency, distribution over conjunction and the success of second order intentions.

Proof

(i) Take a model M with two possible worlds w and v and the accessibility relation R such that wRv and vRv, and let p be true in w but false in v. Then M serial and dense and $p \wedge \mathrm{I}\,\neg p$ is true in w.



(ii)
(1) $\mathrm{I}(p \wedge \mathrm{I}\,\neg p)$ (assumption for reductio)
(2) $\mathrm{I}\,p \wedge \mathrm{I}\,\mathrm{I}\,\neg p$ (distribution over conjunction, 1)
(3) $\mathrm{I}\,p$ (elimination of conjunction, 2)
(4) $\mathrm{I}\,\mathrm{I}\,\neg p$ (elimination of conjunction, 2)
(5) $\mathrm{I}\,\neg p$ (success of second order intentions, 4)
(6) $\neg\,\mathrm{I}\,\neg p$ (consistency of intentions, 3)
(7) $\perp$ (introduction of conjunction, 5 and 6)

This is an interesting result because it shows that Moore sentences do not need to explicitly refer to beliefs, or propositional attitudes that behave similarly, such as knowledge. We modeled intention to behave quite differently from how belief is normally modeled (we dropped positive introspection and instead adopted the success of second order intentions), and we were still able to find a Moore sentence concerning intentions. This is a promising result that encourages research to Moore sentences concerning other propositional attitudes that behave differently from belief.

5.1.4 The Twists

Both the canonical Moore sentence and the Moore sentence in the Toxin problem can plausibly be said to be Moore sentences. But, even baring the fact that one concerns beliefs and the other concerns intentions, there may be some important differences between these sentences. We discuss these differences in this section.

Incentives

In Chapter 4 we explained that the Moore sentence in the surprise exam paradox differs from the canonical Moore sentences in that the students have an incentive to believe the former, while the canonical Moore sentences have no incentives associated with them. Because of the incentive associated with the Moore sentence in the surprise exam paradox, the fact that the students cannot believe the Moore sentence in the surprise exam paradox may *feel* more problematic than the fact that agents cannot believe canonical Moore sentences. The incentive emphasizes the fact that agents cannot believe Moore sentences.

In the Toxin problem, the agent who accepts the billionaire's challenge also has an incentive to intend to make true a Moore sentence. And compared to the surprise exam paradox, the incentive in the Toxin problem is humorously dramatic. The incentive for the contestant to intend to make a Moore sentence true, is one million euro. This incentive emphasizes that the agent cannot intend to make a Moore sentence true, regardless of how badly he wants to.

Temporal Aspect

The Moore sentence in the Toxin problem also differs from the canonical Moore sentences in that it involves an interesting temporal aspect. If an agent would like to believe the canonical Moore sentence $p \wedge \mathrm{B}\,\neg p$, he needs to believe p and $\mathrm{B}\,\neg p$ at the same moment. In the Toxin problem, if the agent wants to intend to make the Moore sentence $p \wedge \mathrm{I}\,\neg p$ true, the agent does not necessarily need to intend to make p and $\mathrm{I}\,\neg p$ true simultaneously. The agent would also obtain the jackpot scenario if he made $\mathrm{I}\,\neg p$ true tonight, and p tomorrow. That is, the agent obtains the reward if he forms the intention to drink the poison tonight, and then changes his mind and does not drink the poison after all tomorrow. Thus, to believe the canonical Moore sentence $p \wedge \mathrm{B}\,\neg p$ an agent needs to believe both of its conjuncts simultaneously, but to intend to make $p \wedge \mathrm{I}\,\neg p$ true, an agent can intend to first make the second conjunct true and then make the first conjunct true. We come back to this temporal aspect of the Toxin problem in sections 5.2.4 and 5.5.

5.2 Newcomb's Problem

A problem that is related to Toxin's paradox is Newcomb's problem:[18]

> You are on a delightfully simple game show. You can choose to obtain the belongings in box A, or to take the belongings in both box A and box B. Box B is transparent and visibly contains 1.000 euro. Box A is opaque. The host of the game show tells you that for the last month, a psychologist followed around your every move to determine what kind of a person you are, a one-boxer (someone who picks only box A) or a two-boxer (someone who picks both box A and box B). The psychologist made her final prediction yesterday and put it in a sealed envelope. If she predicted that you are a one-boxer, she put 1.000.000 euro in box A. If she predicted you are a two-boxer, she put 0 euro in box A. To be perfectly clear, the psychologist made her prediction yesterday and already allocated the money over the boxes; your decision to pick either one or two boxes cannot change how much money is in either box from this point forward. This kind of game show has been played many times before and the prediction of the psychologist matched the decision of the previous players 99% of the time. In this scenario, do you one-box or two-box?[19]

There are two approaches to answer the question posed by Newcomb's problem: an evidentialist approach and a causalist approach. Evidentialists argue that we should one-box, because if we one-box, there is a 99% probability that the psychologist predicted that we would one-box, and thus there is a 99% chance that we obtain 1.000.000 euro. Whereas if we two-box, there is a 99% chance the psychologist predicted we would do so, and thus there is a 99% chance that we obtain only 1.000 euro.[20] To put it in game theoretic terms, the expected utility of one-boxing is approximately 1.000.000 euro and the expected utility of two-boxing is only approximately 1.000 euro. Therefore, if we are rational, we one-box.[21]

Causalists on the other hand, argue that we should two-box. Their reasoning is that, regardless of what the psychologist predicted, we obtain more money by two-boxing. If the psychologist predicted that we would two-box, there is 0 euro in box A and 1.000 euro in box B. In this case, we obtain a higher reward if we take both boxes than if we take only box A. And if the psychologist predicted we would one-box, there is 1.000.000 euro in box A and 1.000 euro in box B: but this does not change the fact that we obtain 1.000 euro more by choosing both boxes rather than just box A. We are thus guaranteed to obtain more money by two-boxing than by one-boxing. Therefore, if we are rational, we two-box.[22]

[18] The similarity between the Toxin problem and Newcomb's problem is noted by Kavka when he introduces the Toxin problem in Kavka, 1983, 35.

[19] This problem was first introduced by Nozick, 1969.

[20] Strictly speaking the problem description of Newcomb's problem does not specify how the psychologists' errors are distributed over one-boxers and two-boxers, so these percentages are not necessarily correct given the problem description. We take it that these percentages nevertheless reflect the intention behind the problem description.

[21] Papineau, 2001.

[22] Nozick, 1969.

We do not pick a side in the debate between evidentialists and causalists. Rather, we discuss the role of Moore sentences in Newcom's problem, and from this discussion derive a variation of the problem that provides a new challenge for both causalists and evidentialists.

5.2.1 Moore Sentence in Newcomb's Problem

Since Newcomb's problem does not concern beliefs or knowledge, its similarity to Moore's paradox has gone largely unnoticed in the literature. Only Goldstein and Cave have described the similarity, in the same article that they discuss the similarity between Moore's paradox and the Toxin problem. Just like in the Toxin problem, they identify a 'jackpot scenario' for the agent in Newcomb's problem, which they describe in terms of desires. They say that the jackpot scenario is for the agent to want to one-box, and then two-box. If we let p read 'I one-box' and, since one-boxing and two-boxing are mutually exclusive and are the only two options in Newcomb's problem, interpret $\neg p$ as 'I two-box', then we can formalize Goldstein and Cave's jackpot scenario as follows:

(5) $\mathrm{D}\, p \wedge \neg p$

The reading of (5) is 'I desire that it is true that I one-box, but I two-box'. Goldstein and Cave note its apparent similarity to the canonical Moore sentence $p \wedge \neg\, \mathrm{B}\, p$. But again, we should correct them, because (5) is more similar to the commissive Moore sentence $p \wedge \mathrm{B}\, \neg p$ than to the omissive $p \wedge \neg\, \mathrm{B}\, p$. This is apparent if in (5) we substitute p for $\neg p$, eliminate double negation and switch the conjuncts around:

(6) $p \wedge \mathrm{D}\, \neg p$

The sentence (6) is particularly interesting because, as mentioned in Chapter 3, scholars have recently started to study Moore sentences involving desires. Williams and Wall both argue that (6) is a Moore sentence.[23] They do so from a philosophical and psychological perspective, considering whether (6) could be desired by a rational agent, according to our intuitions about rationality. Whether (6) is a Moore sentence according to our formal definition of course depends on the formal properties of desire.

5.2.2 Formal Properties of Desire

In this section we present some formal properties that desire may have. Our goal is not to argue for a particular account of desires. Rather, we aim to present some properties that desires may have, so that we can refer back to these properties when we discuss under what assumptions about desire we can say that a Moore sentence is involved in Newcomb's problem in the next section.

One property concerning desires is consistency. Cohen and Levesque write that "Whereas desires can be inconsistent, agents do not typically adopt intentions that they believe conflict with their present- and future-directed intentions."[24] It is not entirely clear from this context in which sense Cohen and Levesque allow desires to be inconsistent. They may tolerate external inconsistency of desires ($\mathrm{D}\, \phi \wedge \mathrm{D}\, \neg \phi$),

[23] Wall, 2012, Williams, 2014.

[24] Cohen and Levesque, 1990, 218.

internal inconsistency of desires ($\mathrm{D}(\phi \wedge \neg\phi)$), or both.

Dubois et al. develop a logic of desires. They do not allow for external inconsistency of desires.[25] They also do not allow for internal inconsistency of desires, although this remark requires some explanation. In their logic of desires, disjunctive sentences $\phi \vee \psi$ have a similar role as conjunctive sentences $\phi \wedge \psi$ in most epistemic logics. They explain the rationale behind this decisions as follows:

> "Indeed, as we shall claim in this paper, while believing ϕ and believing ψ amounts to believing $\psi \wedge \psi$, both desiring ϕ and desiring ψ amounts to desiring $\phi \vee \psi$, and conversely. This is because when an agent discovers new desires, it enlarges the number of desirable situations, while accumulating beliefs reduces the number of possible worlds."[26]

The idea is thus that if an agent believes ϕ is true and believes that ψ is true, then the worlds of which the agent believes that they could be the actual world are (a subset of) those worlds in which ϕ and ψ are both true. But, if an agent desires that ϕ is true and desires that ψ is true, then the worlds that are desirable to that agent are (a subset of) all worlds in which at least ϕ is true, or at least ψ is true. These considerations lead Dubois et al. to assume the formal principle $(\mathrm{D}\,\phi \wedge \mathrm{D}\,\psi) \rightarrow \mathrm{D}(\phi \vee \psi)$. Note that in their setting, $\mathrm{D}_A(\phi \vee \psi)$ is read in natural language as 'agent A desires *ϕ and ψ*'.[27]

Alternatively, one could model desires to be more similar in semantics to beliefs. In this case, one could adopt the distribution and collection of desire over conjunction: $\mathrm{D}(\phi \wedge \psi) \rightarrow (\mathrm{D}\,\phi \wedge \mathrm{D}\,\psi)$ and $(\mathrm{D}\,\phi \wedge \mathrm{D}\,\psi) \rightarrow \mathrm{D}(\phi \wedge \psi)$ respectively. To justify these assumptions one would of course have to propose an alternative account of desire. One such account might be the following. Dubois et al. think of desiring ϕ and desiring ψ as finding any world in which ϕ or ψ is true desirable. The rationale is that if a world fulfills *some* of my desires, that world is desirable. A competing account of belief might take only the world that fulfills *all* my desires to be desirable. In such an account, desiring ϕ and desiring ψ amounts to desiring $\phi \wedge \psi$.

Dubois et al. further assume that one can only desire to make true sentences that one believes to be false ($\mathrm{D}\,\phi \rightarrow \mathrm{B}\,\neg\phi$).[28]. They also assume that agents cannot believe contradictions ($\models \neg\phi \Rightarrow \models \neg\,\mathrm{B}\,\phi$), and from this it follows that valid sentences cannot be desired ($\models \phi \Rightarrow \models \neg\,\mathrm{D}\,\phi$).[29] The justification for this last principle is that to desire a tautology to be true "is in contradiction with the longing aspect of desires (...) There is no point desiring the truth of tautologies, because you can only desire to make true propositions that you believe to be false in your state of affairs, and tautologies are never believed so by a rational agent."[30]

Further, Dubois et al. assume that if an agent desires that ϕ is true, and ψ is true exclusively in situations in which ϕ is true, then the agent also desires that ψ is true ($\mathrm{D}\,\phi$ and $\models \psi \rightarrow \phi \Rightarrow \mathrm{D}\,\psi$).[31] The justification for this principle is that Dubois et

[25] Dubois, Lorini, and Prade, 2017, 207.
[26] Dubois, Lorini, and Prade, 2017, 201.
[27] Dubois, Lorini, and Prade, 2017, 207.
[28] Dubois, Lorini, and Prade, 2017, 203.
[29] Dubois, Lorini, and Prade, 2017, 206.
[30] Dubois, Lorini, and Prade, 2017, 206.
[31] Dubois, Lorini, and Prade, 2017, 205.

al. only deal with what they call *unconditional desires*: sentences that an agent desires to be true unconditionally. They give the example of an agent who desires to drink red wine, but only under the condition that she is also eating red meat. This is an example of a *conditional desire*, because the desire to drink red wine is conditioned on the eating of red meat. An agent who desires to drink red wine, regardless of the conditions, thus in any and all possible situations, has an unconditional desire to drink red wine. Given that they aim to deal only with unconditional desires, they justify the principle mentioned above as follows:

> "Clearly, if an agent desires a certain proposition ϕ to be unconditionally true, then all situations in which ϕ is true should be desirable for the agent. Suppose $\psi \rightarrow \phi$ is valid. Thus, all situations in which ψ is true are situations in which ϕ is true. Consequently, if all situations in which ϕ is true are desirable for the agent, then all situations in which ψ is true should be desirable for the agent. The latter means that the agent desires ψ."[32]

We are not entirely convinced by this justification, as we think agents can sometimes desire a sentence to be true without desiring the logical consequences of this sentence to be true. Even if we restrict ourselves to the consequences that follow from a sentence in virtue of the logical validities (rather than also consider what follows from a sentence and other sentences that happen to be true), it seems that an agent need not desire all the consequences of his desire, for example because he fails to believe some logical validities (of course this cannot happen in the framework of Dubois et al. because agents believe all validities in their framework; but this just means that we doubt that this is a realistic assumption).[33]

Other than these principles, Dubois et al. follow Cohen and Levesque in claiming that desire is unrestricted. The motivation for this seems to be that desires are not as constrained by rationality as for example beliefs or intentions.[34] We think this is reasonable, as many other constraints that could be imposed on desire seem to be clearly false.

For example, positive introspection for desires ($\mathrm{D}\,\phi \rightarrow \mathrm{D}\,\mathrm{D}\,\phi$) has famously been argued against by Frankfurt.[35] Frankfurt argues that if I desire to smoke, this does not mean that I desire to have a desire to smoke. In fact, Frankfurt argued that it is a hallmark of a rational person (a *non-wanton*) that he is able to have and reflect upon conflicts between his first and second order desires.[36],[37]

Negative introspection for desires ($\neg\,\mathrm{D}\,\phi \rightarrow \mathrm{D}\,\neg\,\mathrm{D}\,\phi$) seems similarly flawed. That I do not desire to go to the gym to work out right now, does not imply that I desire that I do not desire to go to the gym. Again, there seems to be room for conflicts

[32] Dubois, Lorini, and Prade, 2017, 206.

[33] Cf. the problem of logical omniscience Stalnaker, 1991, Solaki, 2017.

[34] Cohen and Levesque, 1990, 201.

[35] Frankfurt, 1971, 12.

[36] Frankfurt, 1971, 11.

[37] Second order desire are described by sentences in which a D operator that is not in the range of another D operator, ranges over a (sub-)sentence that contains another D operator, such as $\mathrm{D}\,\mathrm{D}\,\phi$ or $\mathrm{D}(\psi \vee \mathrm{D}\,\chi)$. First order desires are described by sentences in which a D operator that is itself not in the range of another D operator ranges over a (sub-)sentence that contains no other D operators, such as $\mathrm{D}\,\phi$ or $\mathrm{D}(\psi \wedge \chi)$. Note that for example the sentence $\mathrm{D}\,\phi \wedge \mathrm{D}\,\mathrm{D}\,\chi$ describes both a first order desire and a second order desire.

between second order desires and first order (non-)desires in rational agents. And success of second order desires, ($\mathrm{D}\,\mathrm{D}\,\phi \rightarrow \mathrm{D}\,\phi$) seems to fail in cases similar to the example of Frankfurt. Just like one can have a desire to smoke without having the desire to have that desire, it seems that one can fail to have a desire (not) to smoke while one has the desire to have that desire.

But, as mentioned, Williams and Wall have independently argued for the existence of Moore sentences concerning desires.[38] In their arguments, they mention some rationality constraints for desire. Williams formulates a principles in terms of self-frustrating desires. A desire that I have is *self-frustrating* "if it cannot obtain if I desire it".[39] Williams proposes the following rationality constraint on desires:

> "Do not form —or continue to have— a specific desire that you can be reasonably expected to see is self-frustrating."[40]

We can choose to formalize this principle a number of different ways. Let a desire for a sentence ϕ to be true be self-frustrating iff ϕ is not true if it is desired to be true ($\mathrm{D}\,\phi \rightarrow \neg\phi$). Then we can capture the idea that a rational agent does not have self-frustrating desires as $\mathrm{D}\,\phi \rightarrow \neg\phi) \rightarrow \neg\,\mathrm{D}\,\phi$. This reads that if $\mathrm{D}\,\phi$ is self-frustrating in the sense described above, then a rational agent does not desire ϕ to be true. However, this requirement seems too strong to model desire as observed in real world agents, because it implies that desires are veridical ($\mathrm{D}\,\phi \rightarrow \phi$). To see this, suppose that $(\mathrm{D}\,\phi \rightarrow \neg\phi) \rightarrow \neg\,\mathrm{D}\,\phi$ is true. Then its contrapositive $\mathrm{D}\,\phi \rightarrow \neg(\mathrm{D}\,\phi \rightarrow \neg\phi)$ is also true. Suppose now that $\mathrm{D}\,\phi$ is true. Then by modus ponens $\neg(\mathrm{D}\,\phi \rightarrow \neg\phi)$ is true. By the definition of implication this implies that $\mathrm{D}\,\phi$ is true and that $\neg\phi$ is false, which is to say that ϕ is true. Thus, if $(\mathrm{D}\,\phi \rightarrow \neg\phi) \rightarrow \neg\,\mathrm{D}\,\phi$ is true, then $\mathrm{D}\,\phi$ implies ϕ.

Clearly, desires are not veridical in the real world, as it is possible for real agents to desire that a sentence is true while that sentence is not true. It thus seems that Williams' requirement, formalized as $(\mathrm{D}\,\phi \rightarrow \neg\phi) \rightarrow \neg\,\mathrm{D}\,\phi$, poses an unrealistically strong constraint on desires.

Alternatively, we could incorporate belief in our formalization of Williams' principle. Williams says that you should not form a specific desire that *you can be reasonably expected to see* is self-frustrating. 'See' in this phrase can be read as 'believe', and since we are talking about rational agents it is not implausible to assume that agents actually believe whatever they 'can be reasonably expected to believe'. Under these assumptions, we can formalize Williams' principle as $\mathrm{B}(\mathrm{D}\,\phi \rightarrow \neg\phi) \rightarrow \neg\,\mathrm{D}\,\phi$, which reads that if an agent believes a desire $\mathrm{D}\,\phi$ to be self-frustrating, then he does not desire to make ϕ true.

Again, we note that we do not necessarily endorse any of the principles concerning desire mentioned in this section. We will only make use of these principles to make clear in the following section under which assumptions it can be said that a Moore sentence is involved in Newcomb's problem.

[38] Wall, 2012, Williams, 2014.

[39] Williams, 2014, 20.

[40] Williams, 2014, 20.

5.2.3 Moore Sentence in Newcomb's Problem

As we mentioned in the previous section, Williams and Wall argue that Moore sentences concerning desires exist. And we can extend their argument to include the sentence that describes the jackpot scenario in Newcomb's problem, which we formally express as $p \wedge \mathrm{D}\,\neg p$. In this section we discuss under which assumptions about desire this sentence can indeed be said to be a Moore sentence.

First of all, if we assume that desire is characterized at least by consistency, distribution over conjunction and positive and negative introspection for desires, then $p \wedge \mathrm{D}\,\neg p$ is a Moore sentence (for the proof see section 3.4.1, where we proved that $p \wedge \mathrm{B}\,\neg p$ is a Moore sentence if belief is characterized at least by consistency, distribution of conjunction and positive and negative introspection for belief). Also, if we assume that desire is characterized at least by consistency, distribution over conjunction and success of second order desires, $p \wedge \mathrm{D}\,\neg p$ is a Moore sentence (for the proof see section 5.1.3, where we proved that $p \wedge \mathrm{I}\,\neg p$ is a Moore sentence if intention is characterized at least by consistency, distribution over conjunction and success of second order intentions).

However, it is not too plausible that these sets of assumptions characterize desire as observed in real world agents, as we discussed in the previous section. Furthermore, it is difficult to find another set of assumptions that does characterize desire as observed in real world agents, and under which $p \wedge \mathrm{D}\,\neg p$ is a Moore sentence. For example under Williams' assumption that rational agents do not form desires that are self-frustrating, formalized as $(\mathrm{D}\,\phi \rightarrow \neg\phi) \rightarrow \neg\,\mathrm{D}\,\phi)$, $p \wedge \mathrm{D}\,\neg p$ is not a Moore sentence. To see this, consider that $(\mathrm{D}\,\phi \rightarrow \neg\phi) \rightarrow \neg\,\mathrm{D}\,\phi)$ implies that desire is veridical ($\mathrm{D}\,\phi \rightarrow \phi$), as we proved in the previous section. And in section 3.4.3 we proved that $p \wedge \mathrm{K}\,\neg p$ is not a Moore sentence if knowledge is characterized at least by veridicality.

As we mentioned, Williams' principle can also be formalized as $\mathrm{B}(\mathrm{D}\,\phi \rightarrow \neg\phi) \rightarrow \neg\,\mathrm{D}\,\phi)$. Intuitively, under this assumption $p \wedge \mathrm{D}\,\neg p$ may not necessarily be a Moore sentence, unless the agent also believes that $\mathrm{D}(p \wedge \mathrm{D}\,\neg p) \rightarrow \neg(p \wedge \mathrm{D}\,\neg p)$. If the agent believes this and Williams' principle holds, $\mathrm{D}(p \wedge \mathrm{D}\,\neg p)$ is clearly unsatisfiable with respect to any class of Kripke models that at least guarantees the consistency of desires, since by $\mathrm{D}(p \wedge \mathrm{D}\,\neg p) \rightarrow \neg(p \wedge \mathrm{D}\,\neg p)$ and modus ponens on Williams principle we immediately get that $\neg\,\mathrm{D}(p \wedge \mathrm{D}\,\neg p)$.

There is something to be said for the assumption that the agent believes that $\mathrm{D}(p \wedge \mathrm{D}\,\neg p) \rightarrow \neg(p \wedge \mathrm{D}\,\neg p)$. First note that if we assume that desire distributes over conjunction and is consistent, $\mathrm{D}(p \wedge \mathrm{D}\,\neg p) \rightarrow \neg(p \wedge \mathrm{D}\,\neg p)$ is true.

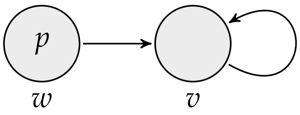
Proof

(1) $\mathrm{D}(p \wedge \mathrm{D}\,\neg p)$ (assumption)
(2) $\mathrm{D}\,p$ (elimination of conjunction, 1)
(3) $\neg\,\mathrm{D}\,\neg p$ (consistency of desires, 2)
(4) $\neg p \vee \neg\,\mathrm{D}\,\neg p$ (introduction of disjunction, 3)
(5) $\neg(p \wedge \mathrm{D}\,\neg p)$ (de Morgan, 4)
(6) $\mathrm{D}(p \wedge \mathrm{D}\,\neg p) \rightarrow \neg(p \wedge \mathrm{D}\,\neg p)$ (introduction of implication, 1 and 5)

Thus, there is a relatively simple proof that $\mathrm{D}(p \wedge \mathrm{D}\neg p) \rightarrow \neg(p \wedge \mathrm{D}\neg p)$ is true. Arguably, this implies that a rational agent also believes that it: $\mathrm{B}(\mathrm{D}(p \wedge \mathrm{D}\neg p) \rightarrow \neg(p \wedge \mathrm{D}\neg p))$. And if Williams' principle formalized as $\mathrm{B}(\mathrm{D}\phi \rightarrow \neg\phi) \rightarrow \neg\mathrm{D}\phi)$ holds, this implies that $\neg\mathrm{D}(p \wedge \mathrm{D}\neg p)$ is true.

Thus, if we assume that Williams' principle holds and the agent believes that $\mathrm{D}(p \wedge \mathrm{D}\neg p) \rightarrow \neg(p \wedge \mathrm{D}\neg p)$, $\mathrm{D}(p \wedge \mathrm{D}\neg p)$ is unsatisfiable with respect to any class of Kripke models that at least guarantees the consistency of desires and the distribution of desire over conjunction. And clearly, $p \wedge \mathrm{D}\neg p$ is satisfiable with respect to a class of Kripke models that is serial (which guarantees the consistency of desires and the distribution of desire over conjunction).

Proof
Take a model M with two possible worlds w and v and the accessibility relation R such that wRv and vRv, and let p be true in w. Then M is serial, and $p \wedge \mathrm{D}\neg p$ is true in w.



Thus, under the assumptions that agents do not form desires that they believe to be self-refuting ($\mathrm{B}(\mathrm{D}\phi \rightarrow \neg\phi) \rightarrow \neg\mathrm{D}\phi)$), the assumption that the agent in Newcomb's problem believes that the desire to obtain the jackpot scenario is self-refuting ($\mathrm{D}(p \wedge \mathrm{D}\neg p) \rightarrow \neg(p \wedge \mathrm{D}\neg p)$), and the further assumptions that desire distributes over conjunction and is consistent, the sentence describing the jackpot scenario in the Toxin problem $p \wedge \mathrm{D}\neg p$ can be said to be a Moore sentence.

5.2.4 The Twists

The sentence that describes the jackpot scenario in Newcomb's problem can be said to be a Moore sentence under the assumptions we specified in the previous section. Nevertheless, even baring the fact that it concerns desires instead of beliefs, the Moore sentence in the Newcomb problem differs from the canonical Moore sentence $p \wedge \mathrm{B}\neg p$ in some interesting ways. We discuss these differences in this section.

Incentives

Alike the Moore sentences in the surprise exam paradox and the Toxin problem, the Moore sentence in Newcomb's problem differs from the canonical Moore sentence in that it has an incentive associated with it. This incentive is monetary, in a similarly amusing manner as in the Toxin problem. The Moore sentence in Newcomb's problem describes the situation in which the contestant on the game show (likely) obtains a million and a thousand euro. Associating incentives with Moore sentence may make it *feel* more problematic than the canonical Moore sentence that lacks such an incentive, because the incentive highlights the fact that the agent cannot believe (or desire) the Moore sentence to be true.

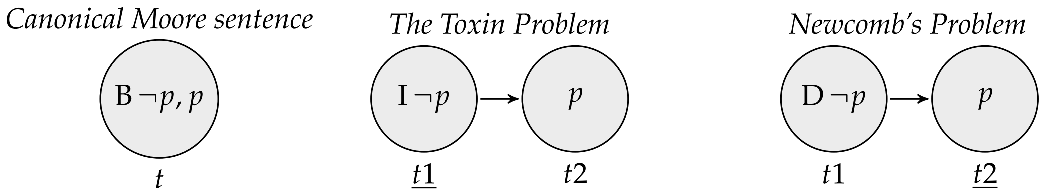
Temporal Aspect

Similar to the Moore sentence in the Toxin problem, the Moore sentence in Newcomb's problem comes with an interesting temporal aspect that is not present in the canonical Moore sentence. As we mentioned in the final subsection of section 5.1.4, to believe $p \land \mathrm{B}\,\neg p$, an agent would need to believe p and $\mathrm{B}\,\neg p$ at the same moment. In the Toxin problem, for an agent to intend to make $p \land \mathrm{I}\,\neg p$ true, the agent could intend to make $\mathrm{I}\,\neg p$ true first and p true second.

The Moore sentence in Newcomb's problem is similar to the Moore sentence in the Toxin problem, but also differs from it in an interesting respect. In Newcomb's problem, the agent wants to desire $p \land \mathrm{D}\,\neg p$ to be true. As in the Toxin problem, the agent does not necessarily desire both conjuncts of this sentence to be true simultaneously. Rather, the agent wants $\mathrm{D}\,\neg p$ to be true first, at the moment that the psychologist predicts whether he is a one-boxer or a two-boxer. Later, during the game show, the agent wants p to be true, so that he can maximize his prize money.

The interesting difference with the Toxin problem is that in Newcomb's problem, the agent is already in the game show, thus the point at which he desires $\mathrm{D}\,\neg p$ to be true already lies behind him. However, it could be argued that during the game show the agent still wants $\mathrm{D}\,\neg p$ to be true, because it is suggested in the description of Newcomb's problem that there is a strong correlation between the truth of $\mathrm{D}\,\neg p$ during the game show and the truth of $\mathrm{D}\,\neg p$ when the psychologist makes her prediction (this correlation is suggested by the 99% accuracy of the psychologist's prediction of what the agent will end up doing). But even if this is so, the agent's desire for $\mathrm{D}\,\neg p$ to be true during the game show derives from his desire for $\mathrm{D}\,\neg p$ to be true earlier.

We can summarize the differences in temporal aspects of the canonical Moore sentence and the Moore sentences in the Toxin problem and Newcomb's problem, in the following picture:



Canonical Moore sentence

The Toxin Problem

Newcomb's Problem

If an agent wants to believe the canonical Moore sentence, he wants to believe p and $\mathrm{B}\,\neg p$ simultaneously. If an agent wants to obtain the jackpot scenario in the Toxin problem, he wants to align his intentions so that $\mathrm{I}\,\neg p$ is true at $t1$ and p is true at $t2$, and the agent is at $t1$. If an agent wants to obtain the jackpot scenario in Newcomb's problem, he wants to align his desires so that $\mathrm{D}\,\neg p$ is true at $t1$ and p is true at $t2$, and the agent is at $t2$. This conceptualization of the difference between the Toxin problem and Newcomb's problem seems not to have been suggested in the literature before, and we briefly return to this point in the final section of this chapter.

5.3 Omissive Variants of the Toxin Problem and Newcomb's Problem

In Chapter 2 we mentioned that there are two different kinds of doxastic Moore sentences that require slightly different approaches to be solved or explained away: the commissive Moore sentence $p \wedge \mathrm{B}\,\neg p$ and the omissive Moore sentence $p \wedge \neg\,\mathrm{B}\,p$. Both the Toxin problem and Newcomb's problem make use of Moore sentences with a commissive syntactic structure. Prima facie it seems that we can create new variations of these problems by considering the omissive variations of their Moore sentences. We do this in the coming two sections.

5.3.1 The Potion Problem

Based on the omissive variant of the sentence in the Toxin problem, $p \wedge \neg\,\mathrm{I}\,p$, we can try to create a new problem description for what we might call *the Potion problem* (or, for a more informative name with less literary appeal, the *omissive Toxin problem*):

> A billionaire comes up to you and offers you a challenge. He shows you a small bottle and explains that it contains a potion that delivers tremendous health benefits, and on top of that is the most delicious tasting liquid known to mankind. The billionaire offers to give you one million euro if, by midnight tonight, you have *not* formed the intention to drink the potion tomorrow morning. He stresses that it is not important whether you end up drinking the potion tomorrow morning; if at midnight tonight you do not intend to drink the potion tomorrow, you immediately obtain the reward, regardless of whether you end up drinking the potion (assume that he has a device that accurately measures your intentions in real time, such as a high tech brain scanner).

In this scenario the jackpot scenario for the agent is to not have the intention to drink the potion tomorrow by midnight tonight, but still drink the potion tomorrow. The question is whether this challenge can be completed by a rational agent.

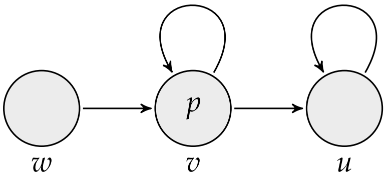
We should note that, just like in the Toxin problem, it is possible for the jackpot scenario of the Potion problem to obtain. It is possible that an agent does not have the intention to drink the potion tomorrow at midnight, but ends up drinking the potion anyway.

However, it is not clear that the agent can intend to bring the jackpot scenario about. One could argue that if tonight the agent intends to drink the potion tomorrow without having the intention tonight, he tonight actually does have the intention to drink the potion tomorrow. The reason for this would be that if an agent intends to make ϕ true unintentionally, he still intends to make ϕ true.

It seems that the sentence that describes the jackpot scenario in the Potion problem thus is such that an agent may not be able to (successfully) intend to make it true. To determine whether this sentence is a Moore sentence, we need to make our assumptions about intention explicit.

We formalize the sentence that describes the jackpot scenario in the Potion problem as: $p \wedge \neg\,\mathrm{I}\,p$. Let us consider whether this sentence is a Moore sentence under the same assumptions about intention we made when we considered whether the sentence that describes the jackpot scenario in the Toxin problem is a Moore sentence. These assumptions are that intentions are characterized by distribution over conjunction, consistency, and success of second order intentions. Under these assumptions, $p \wedge \neg \mathrm{I}\, p$ is *not* a Moore sentence. We prove this by proving that $I(p \wedge \neg \mathrm{I}\, p)$ is satisfiable on a class of Kripke models characterized by seriality and density.

Proof
Take a model M with three possible worlds w, v and u and the accessibility relation R such that wRv, vRu, vRv, and uRu. Let p be true in v but false in u. Then M is serial and dense and $\mathrm{I}(p \wedge \neg \mathrm{I}\, p)$ is true in w.



This is an interesting result, because it turns out that under the assumptions mentioned above, the Toxin problem involves a Moore sentence but the Poison problem does not. Even though, we argued, both the Toxin problem and the Potion problem involve a jackpot scenario of which it is debatable whether the agent can successfully intend to bring it about. This indicates that the Toxin problem and the Potion problem pose slightly different challenges to scholars that study intentions.

5.3.2 The Omissive Newcomb Problem

We may also wonder whether there is an omissive variant of Newcomb's problem. One way to set up such a scenario is to alter the condition of the game show slightly. In the original Newcomb problem, the psychologist put a million euro in box A if she has predicted that the contestant is a one-boxer. To create an omissive Newcomb problem, we can change this condition so that the psychologist put a million in box A if she predicted the contestant is *not a two-boxer*. This is different from the original Newcomb problem if we assume that the psychologist always makes either of three predictions: she predicts that you are a one-boxer, that you are a two-boxer, or that you are neither. She predicts that you are a one-boxer if you are the kind of person that, if you could play the game show multiple times, would one-box every time. She predicts that you are a two-boxer if you are the kind of person that, if you were to play the game multiple times, would always two-box. And she predicts that you are neither a one-boxer nor a two-boxer if you are the kind of person that, if you were to play the game show multiple times, would sometimes one-box and sometimes two-box.

To be clear, in the omissive Newcomb problem the contestant still plays in the game show only once. But, the psychologist bases her prediction not (exclusively) on what you will do in the game show. Rather, she bases her prediction on how you would behave if you were presented with this scenario multiple times. The following example makes clear how this differs from the original Newcomb problem.

Suppose that Robert is invited to play in the game show. Robert, being a good study, decides to prepare himself for the game show by reading up on the philosophical literature about it. He figures that he will adopt the strategy that the great thinkers of his time agree is the best. But, to Robert's surprise, the philosophical literature offers no such agreement. Some scholars suggest you should one-box, and others that you should two-box. Robert, after exhausting a great many philosophical journals in search of a conclusive argument, decides that he cannot determine whether he should one-box or two-box. Defeated, he decides in the end to flip a coin, and one-box if it comes up heads and two-box if it comes up tails.

Robert is the kind of person who, if he were to play in the game show multiple times, would sometimes one-box and sometimes two-box, since the coin flip sometimes comes up heads and sometimes come up tails. In the original Newcomb problem, the psychologist put the million in the box if she predicted the contestant is a two-boxer. It is debatable what this means for Robert. Robert is not a two-boxer in the sense that he will reliably two-box. This could prompt the psychologist not to predict that he is a two-boxer. At the same time, the chance that Robert will two-box is about 50%. In an attempt to reflect this fact in her prediction, the psychologist could choose to flip a coin of her own, and predict that Robert is a one-boxer if the coin comes up heads and predict that he is a two-boxer if the coin comes up tails.

In the omissive Newcomb problem, it is clear that the psychologist will predict that Robert is neither a one-boxer nor a two-boxer. Therefore, the 'strategy' that Robert ends up playing, is highly effective in the omissive Newcomb problem. Since Robert is neither a one-boxer nor a two-boxer, she psychologist will likely predict that he is not a two-boxer and put a million euro in box A. If that happens, Robert has a 50% chance that he ends up choosing only one box in which case he obtains a million euro, and a 50% chance he ends up two-boxing in which case he obtains a million euro plus an additional thousand euro.

It is thus clear that Robert's strategy is effective in the omissive Newcomb problem, but it remains unclear whether the same strategy would work as well in the original Newcomb problem. Since the two problems seem to allow for different solutions, we can say that they are different problems. Thus, the omissive Newcomb problem seems to pose a different challenge than the original Newcomb problem.

5.4 The Intentionability Paradox and the Desirability Paradox

In Chapter 4 we discussed the result of the knowability paradox: if there are unknown truths, there are truths that cannot be known. We argue that we can derive a similar *intentionability paradox*. The result of this paradox is that if it is possible that there are truths that are not intended to be true, then there are truths that cannot be intended to be true. And we can also derive a *desirability paradox*, which states that if there are undesired truths, then there are undesirable truths.

The knowability paradox is derived using an epistemic Moore sentence: $p \wedge \neg \mathrm{K}\, p$. This sentences is such that it can be true, but it cannot be known. The intentionaility paradox can be derived using a Moore sentence concerning intentions: $p \wedge \neg \mathrm{I}_A\, p$. This sentence describes that a sentence is true although agent A does not intend to make it true. If we assume that there can be sentences that are true even though A does not intend them to be true, $p \wedge \neg \mathrm{I}_A\, p$ can be true. And as we argued in the previous section, A cannot intend to make $p \wedge \neg \mathrm{I}_A\, p$ true. Thus, if there are truths that *are* not intended to be true, then there are truths that *cannot* be intended to be true.

Similarly, we can derive the desirability paradox using a Moore sentence concerning desires: $p \wedge \neg \mathrm{D}\, p$.[41] If there can be truths that are not desired, then this sentence can be true. And, as we discussed in the previous section, we can plausibly say that this sentence cannot be desired to be true. Thus, if there are truths that *are* not desired to be true, then there are truths that *cannot* be desired to be true.

As we discussed, the knowability paradox is really only a paradox insofar as the optimistic verification thesis is plausible in the first place. Similarly, the intentionability paradox and the desirability paradox are paradoxes insofar as optimistic theses about intentions and desires are plausible. A candidate optimistic thesis about intentions is that for all sentences that can be true, agents can intend to make them true. This thesis is contradicted by the intentionability paradox, but prima facie it may have seen quite plausible. Before reading this thesis, the reader may not have seen any reason to doubt that agents can intend to make any sentence true, provided at least that the sentence can be true.

A candidate optimistic thesis about desires may be that all sentences that can be true can be desired to be true. This thesis is contradicted by the desirability paradox. But again, before reading this thesis, the reader may not have been aware of any counterexamples to the thesis that any sentence that can be true can be desired to be true.

It thus seems that from our discussion about the knowability paradox, and Moore sentences involving intentions and desires, emerged two new paradoxes. Fortunately, we also know that solutions to the knowability paradox have been offered, which can now be reapplied to the intentionability and desirability paradoxes. What we are left with in the end is then a more mature understanding of the limits of intentions and desires.

5.5 Solution to the Intentionability and Desirability Paradoxes

The knowability paradox shows the failure of the verification thesis that all truths can be known. Van Benthem argued that, upon reflection, the failure of this thesis can be accepted with appreciation rather than remorse.[42] The reason for this is that it was rather naive of us to assume that human knowledge is constrained only by what is true. It is also constrained by the dynamic aspects of how information is conveyed in communication. Since humans can never know about their own ignorance about particular facts, statements of their ignorance cannot function the straight-forward communicative role of conveying the information they contain. Statements of ignorance are not meant to make someone aware that he is ignorant; they are meant to

[41] In this section we follow Williams and Wall in assuming that this sentence is a Moore sentence, and thus cannot be desired.

[42] Van Benthem, 2004.

take away his ignorance.

Similarly, the intentionability and desirability paradoxes highlight that there are limits to what humans can intend and desire. And these limits are more restrictive then we may have thought initially. Just like it initially seemed plausible that we can know anything that is true, it may have seemed like we could intend to make true any sentence that can be true, and desire any sentence to be true that can be true. As the knowability paradox paved the way for scientific research based on a more realistic and mature assumption about the limits of human knowledge, we hope that the intentionability paradox and desirability paradox spark similar research towards intentions and desires.

One suggestion for future research is a consideration of the role of Moore sentences concerning desire and intention in human cooperation. Moore sentences concerning belief seem to be useful in communication, because an agent A can point out to an agent B that B fails to believe some fact that is true, thereby teacher B that fact. The multi-agent setting is crucial here, because as our discussion in this thesis showed many times over, a single agent can make little use of a doxastic Moore sentence that is about himself. But in the multi-agent setting doxastic Moore sentences become useful devices to inform one another about ignorance and false beliefs. We suspect the same is true for Moore sentences concerning intention and desire. In the single agent setting, an agent can make no use out of such Moore sentences, at least in the sense that he cannot intend to make them true or desire them to be true. But, an agent A can intend to make true that p is true while agent B does not intend to make p true. This is interesting with regards to human cooperation, because the sentence $p \wedge \neg \mathrm{I}_B\, p$ is a sentence about B, that B cannot intend to make true himself. Yet, A can intend to make it true for him. And the same is true about desires. $p \wedge \neg \mathrm{D}_B\, p$ is a sentence about B that B cannot desire to be true, but another agent A can desire it to be true for him.

We think that the result that there are sentences about agents that they can never intend to make true, while other agents can intend to make it true for them, is rather deep and may lead to many different discoveries about intentions. One example of a result that can be derived from this, is a new proposal for a solution to the Toxin problem. Suppose that agent A is posed the Toxin problem and thus wants to make the sentence $p \wedge \mathrm{I}_A \neg p$ true. A, having read this thesis, realizes that he cannot intend to make this sentence true himself, asks his friend B whether B can intend to make it true for him. After all, $\mathrm{I}_B(p \wedge \mathrm{I}_A\, p)$ is unproblematic. B could intend to make $p \wedge \mathrm{I}_A \neg p$ true in various ways, for example by preaching to A that drinking the poison is the right thing to do (without actually making A drink it in the end). Of course, the interesting critical question that can be asked about this solution is whether A's decision to ask B to help him, is in any way problematic. If A intends to make it true that B intends to make $p \wedge \mathrm{I}_A\, p$ true, does A then somehow intend to make $p \wedge \mathrm{I}_A\, p$ true himself? Does $\mathrm{I}_A\, \mathrm{I}_B(p \wedge \mathrm{I}_A\, p)$ imply $\mathrm{I}_A(p \wedge \mathrm{I}_A\, p)$? These questions are all interesting for follow up research towards the role of Moore sentences concerning intentions in human cooperation.

Of course, similar questions can be asked about the role of Moore sentences concerning desires in human cooperation. A may not be able to desire that $p \wedge \mathrm{D}_A\, p$ is true, but B is. If A is playing in the Newcomb game show, he cannot desire that he wins the jackpot; but his friend B can desire for him to win There are things that we cannot want for ourselves, but that others can want for us. We think this fact may lead to many interesting discovering in research towards desires.

To offer one more suggestion for future research, we would like to point out how our discussion of the interpretation of the knowability paradox can be used to think about the Toxin problem and Newcomb's problem. In our discussion of the knowability paradox, we distinguished two phases, before and after a learning action. This distinction proved especially relevant for the consequences of the knowability paradox for the verification thesis that all truths can be known, because our learning action changes which sentences are true. 'All truths' thus refers to different sentences in the phase before the learning action than in the phase after it.

We already pointed out in section 5.2.4 that the Toxin problem and Newcomb's problem both consist of two phases as well. In the Toxin problem, there is the phase that lasts up to midnight and the phase after midnight. In Newcomb's problem there is the phase up to the prediction of the psychologist, and the phase after this prediction. In both problems, the agent's propositional attitude gets measured at the end of the first phase, and whatever propositional attitude is measured then determines the allocation of the rewards. Whatever propositional attitude the agent forms after this measurement is irrelevant for the reward he receives.

In the Toxin problem, the agent is situated in phase one of his problem, which means that his intentions towards drinking the poison tomorrow are still relevant when the wager is offered to him. He is also told that at midnight, he will transition to phase two, in which his intention will no longer be relevant for what reward he will obtain. One way to think about this transition from phase one to phase two at midnight, as the agent *learning* that whatever propositional attitude he has, does not matter anymore. Note that this learning does not consist of the agent receiving more information, because the conditions of the wager were all explained fully during phase one. But, the indexical sentence 'the intention I have *now* regarding the drinking of the poison tomorrow determines what reward I will obtain' becomes false during the transition from the first to the second phase. At midnight it is true that the agent's intentions determine his reward, and the agent knows this. After midnight it has become false that the agent's intentions determine his reward, and the agent has learned this.

In the knowability paradox it is important to distinguish between the phases before and after the learning action because after the learning action, different sentences are true than before the learning action, and this affects which sentences can be known. Similarly, in the Toxin problem we should distinguish between the phases before and after midnight, because at midnight the agent learns that his intentions no longer determine his reward, and this knowledge affects which sentences the agent can (rationally) intend to make true. It seems that before midnight it might be rational for the agent to intend to make it true that he drinks the poison tomorrow, but after he learns that his intention no longer matters, this intention may no longer be rational.

Newcomb's problem is interesting because the agent is informed of the situation he is in, when he is already in phase two. That is to say, when the agent is informed of his situation, the desire that determines the allocation of the rewards over the boxes has already been measured. What makes Newcomb's problem so puzzling is that it is debatable to what extent the agent's desire in the second phase is still 'relevant' for the reward he will obtain. After all, the psychologist is said to be 99% accurate, which seems impossible unless the desires of contestants at the moment of prediction tend to concord with their desires during the game show. We suggest that distinguishing the two phases of Newcomb's problem and sorting out exactly how the two phases are related to each other may be the key to solving Newcomb's problem.

Chapter 6

Conclusion

In this thesis we discussed Moore's paradox and its relation to other paradoxes and problems. For this purpose we developed a formal definition of Moore sentences that also applies to Moore sentences of deviant syntactical structures or semantic contents. This enabled us to single out Moore sentences in epistemic and non-epistemic paradoxes.

We discussed the role of Moore sentences in the surprise exam paradox, the knowability paradox, the Toxin problem and Newcomb's problem, and also contributed to the discussions about these problems. We provided a new argument in the long-standing debate about which sentence, precisely, in the surprise exam paradox is a Moore sentence. We argued that the knowability paradox is only an argument against the verification thesis under some but not all of its interpretations. We pointed to some interesting temporal aspects of the Toxin problem and Newcomb's problem, and we introduced new variations of both these problems, as well as two new paradoxes akin to the knowability paradox concerning intention and desire. We also suggested how these paradoxes can be addressed by considering the role of Moore sentences in human cooperation.

A strong personal motivation for this thesis has been the value and enjoyment that is intrinsic to studying the subtle ways in which different paradoxes and problems are similar to one another. But, this has not been the sole reason for conducting this study. Paradoxes challenge existing dogma's and demand creative solutions that can catalyze scientific progress. Understanding the relations between different paradoxes that we have described in this thesis can add to this process. We hope that many more discoveries about individual paradoxes and their relations follow from our results, and that these discoveries in turn further inspire research on the topics that these paradoxes concern.

Bibliography

Alexander, Peter (1950). “Pragmatic paradoxes”. In: *Mind* 59.236, pp. 536–538.

Almeida, Claudio (2001). “What Moore’s paradox is about”. In: *Philosophy and Phenomenological Research* 62.1, pp. 33–58.

Baltag, Alexandru and Bryan Renne (2016). “Dynamic Epistemic Logic”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.

Berto, Francesco and Matteo Plebani (2015). *Ontology and metaontology: A contemporary guide*. Bloomsbury Publishing.

Binkley, Robert (1968). “The surprise examination in modal logic”. In: *The Journal of Philosophy* 65.5, pp. 127–136.

Bloom, Allan and Adam Kirsch (2016). *The republic of Plato*. Basic Books.

Bratman, Michael (1987). “Intention, plans, and practical reason”. In:

Champlin, TS (1976). “Quine’s judge”. In: *Philosophical studies* 29.5, pp. 349–352.

Cholbi, Michael (2009). “Moore’s paradox and moral motivation”. In: *Ethical theory and moral practice* 12.5, p. 495.

Chow, Timothy Y (1998). “The surprise examination or unexpected hanging paradox”. In: *The American Mathematical Monthly* 105.1, pp. 41–51.

Clark, Ron (1994). “Pragmatic paradox and rationality”. In: *Canadian journal of philosophy* 24.2, pp. 229–242.

Cohen, Philip R and Hector J Levesque (1990). “Intention is choice with commitment”. In: *Artificial intelligence* 42.2-3, pp. 213–261.

Collins, Arthur W (1996). “Moore’s paradox and epistemic risk”. In: *The Philosophical Quarterly (1950-)* 46.184, pp. 308–319.

DeRose, Keith (1991). “Epistemic possibilities”. In: *The philosophical review* 100.4, pp. 581–605.

— (1997). “Editor’s introduction”. In: *Notre Dame ournal of Formal Logic* 38, pp. 481–487.

Dubois, Didier, Emiliano Lorini, and Henri Prade (2017). “The Strength of Desires: A Logical Approach”. In: *Minds and Machines* 27.1, pp. 199–231.

Fitch, Frederic B (1963). “A logical analysis of some value concepts”. In: *The journal of symbolic logic* 28.2, pp. 135–142.

Frankfurt, Harry G. (1971). “Freedom of the Will and the Concept of a Person”. In: *Journal of Philosophy* 68.1, pp. 5–20.

Georgeff, Michael et al. (1998). “The belief-desire-intention model of agency”. In: *International Workshop on Agent Theories, Architectures, and Languages*. Springer, pp. 1–10.

Gerbrandy, Jelle (2007). “The surprise examination in dynamic epistemic logic”. In: *Synthese* 155.1, pp. 21–33.

Goldstein, Laurence and Peter Cave (2008). “A Unified Pyrrhonian Resolution of the Toxin Problem, the Surprise Examination, and Newcomb’s Puzzle”. In: *American Philosophical Quarterly* 45.4, pp. 365–376.

Gombay, André (1988). “Some paradoxes of Counterprivacy”. In: *Philosophy* 63.244, pp. 191–210.

Green, Mitchell S and John N Williams (2007). *Moore's paradox: new essays on belief, rationality, and the first person*. Oxford University Press on Demand.

Hart, WD and Colin McGinn (1976). "Knowledge and necessity". In: *Journal of philosophical logic* 5.2, pp. 205–208.

Heal, Jane (1994). "Moore's paradox: A Wittgensteinian approach". In: *Mind* 103.409, pp. 5–24.

Herzig, Andreas et al. (2017). "BDI logics for BDI architectures: old problems, new perspectives". In: *KI-Künstliche Intelligenz* 31.1, pp. 73–83.

Hintikka, Jaakko (1962). "Knowledge and belief". In:

Irwin, Terence (1979). "Gorgias". In:

Kavka, Gregory S (1983). "The toxin puzzle". In: *Analysis* 43.1, pp. 33–36.

Levinson, Stephen C (1983). "Pragmatics Cambridge University Press". In: *Cambridge UK*.

Levy, Ken (2009). "The solution to the surprise exam paradox". In: *The Southern Journal of Philosophy* 47.2, pp. 131–158.

Linsky, Leonard (1968). "On Interpreting Doxastic Logic". In: *The Journal of Philosophy* 65.17, pp. 500–502.

Lorini, Emiliano and Andreas Herzig (2008). "A logic of intention and attempt". In: *Synthese* 163.1, pp. 45–77.

Martinich, Aloysius P (1980). "Conversational maxims and some philosophical problems". In: *The Philosophical Quarterly (1950-)* 30.120, pp. 215–228.

Menzel, Christopher (2017). "Possible Worlds". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University.

Meyer, John-Jules Ch (2003). "Modal epistemic and doxastic logic". In: *Handbook of philosophical logic*. Springer, pp. 1–38.

Moore, George Edward (1942). "A Reply to My Critics". In: *The Philosophy of G. E. Moore*. Ed. by Paul Arthur Schilpp. Open Court.

— (1944). "Russell's" Theory of Descriptions."" In:

Nozick, Robert (1969). "Newcomb's problem and two principles of choice". In: *Essays in honor of Carl G. Hempel*. Springer, pp. 114–146.

O'Connor, Donald J (1948). "Pragmatic paradoxes". In: *Mind* 57.227, pp. 358–359.

Papineau, David (2001). "Evidentialism reconsidered". In: *Nous* 35.2, pp. 239–259.

Priest, Graham (1992). "What is a non-normal world?" In: *Logique et analyse* 35.139/140, pp. 291–302.

Quine, WV (1953). "On a so-called paradox". In: *Mind* 62.245, pp. 65–67.

Rantala, Veikko (1982). "Impossible worlds semantics and logical omniscience". In: *Acta Philosophica Fennica* 35, pp. 106–115.

Roy, Olivier et al. (2008). *Thinking before acting: intentions, logic, rational choice*. Institute for Logic, Language and Computation.

Russell, Bertrand (1905). "On denoting". In: *Mind* 14.56, pp. 479–493.

Santas, Gerasimos (1964). "The Socratic Paradoxes". In: *The Philosophical Review*, pp. 147–164.

Schick, Frederic (2000). "Surprise, self-knowledge, and commonality". In: *The Journal of philosophy* 97.8, pp. 440–453.

Scriven, Michael (1951). "Paradoxical announcements". In: *Mind* 60.239, pp. 403–407.

Shoemaker, Sydney (1995). "Moore's paradox and self-knowledge". In: *Philosophical Studies* 77.2-3, pp. 211–228.

— (1996). *The first-person perspective and other essays*. Cambridge University Press.

Shoham, Yoav (2009). "Logical theories of intention and the database perspective". In: *Journal of Philosophical Logic* 38.6, p. 633.

Sim, Kwang Mong (1997). "Epistemic logic and logical omniscience: A survey". In: *International Journal of Intelligent Systems* 12.1, pp. 57–81.

Smullyan, Raymond M (2012). *Forever undecided*. Knopf.

Sober, Elliott (1998). "To give a surprise exam, use game theory". In: *Synthese* 115.3, pp. 355–373.

Solaki, Anthia (2017). "Steps out of Logical Omniscience". PhD thesis. Universiteit van Amsterdam.

Sorensen, Roy (2000). "Moore's problem with iterated belief". In: *The Philosophical Quarterly* 50.198, pp. 28–43.

Sorensen, Roy A (1988). *Blindspots*.

Stalnaker, Robert (1991). "The problem of logical omniscience, I". In: *Synthese* 89.3, pp. 425–440.

Stevenson, Charles L. (1942). "Arguments against ethical naturalism". In: *The Philosophy of G. E. Moore*. Ed. by Paul Arthur Schilpp. Open Court.

Van Benthem, Johan (2004). "What one may come to know". In: *Analysis* 64.282, pp. 95–105.

Van Ditmarsch, Hans et al. (2015). "An introduction to logics of knowledge and belief". In: *arXiv preprint arXiv:1503.00806*.

Wall, David (2012). "A Moorean paradox of desire". In: *Philosophical Explorations* 15.1, pp. 63–84.

Williams, John N (2007). "The surprise exam paradox: Disentangling two reductios". In: *Journal of Philosophical Research* 32, pp. 67–94.

— (2014). "Moore's Paradox in Belief and Desire". In: *Acta Analytica* 29.1, pp. 1–23.

— (2015a). "Moore's Paradox in Speech: A Critical Survey". In: *Philosophy Compass* 10.1, pp. 10–23.

— (2015b). "Moore's Paradox in Thought: A Critical Survey". In: *Philosophy Compass* 10.1, pp. 24–37.

Williamson, Timothy (1999). "Rational failures of the KK Principle". In: *The logic of strategy*, pp. 101–118.

— (2002). *Knowledge and its Limits*. Oxford University Press on Demand.

Wittgenstein, Ludwig, Gertrude Elizabeth Margaret Anscombe, and Ludwig Wittgenstein (1963). *Philosophical Investigations... Translated by GEM Anscombe.[A Reprint of the English Translation Contained in the Polyglot Edition of 1958.].* Basil Blackwell.

Wright, Crispin and Aidan Sudbury (1977). "The paradox of the unexpected examination". In: *Australasian Journal of Philosophy* 55.1, pp. 41–58.

Yaffe, Gideon (2004). "Trying, intending, and attempted crimes". In: *Philosophical Topics* 32.1/2, pp. 505–532.

Yalcin, Seth (2007). "Epistemic modals". In: *Mind* 116.464, pp. 983–1026.