ALMOST IDEAL:
Computational Epistemology and the Limits of Rationality for Finite Reasoners

By

Danilo Fraga Dantas
B.A. (Universidade Federal da Bahia) 2005
M.A. (Universidade Federal da Bahia) 2007
M.A. (Universidade Federal do Rio Grande do Sul) 2010
M.A. (University of California Davis) 2013

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Philosophy

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Bernard Molyneux, Chair

Hanti Lin

Adam Sennet

George J. Mattey II

Almerindo Ojeda

Committee in Charge

2016

Para Emilly.

Contents

List of Figures

List of Tables

Abstract

ALMOST IDEAL:
Computational Epistemology and the Limits of Rationality for Finite Reasoners

The notion of an ideal reasoner has several uses in epistemology. Often, ideal reasoners are used as a parameter of (maximum) rationality for finite reasoners (e.g. humans). However, the notion of an ideal reasoner is normally construed in such a high degree of idealization (e.g. infinite/unbounded memory) that this use is unadvised. In this dissertation, I investigate the conditions under which an ideal reasoner may be used as a parameter of rationality for finite reasoners. In addition, I present and justify the research program of computational epistemology, which investigates the parameter of maximum rationality for finite reasoners using computer simulations.

In chapter 1, I investigate the use ideal reasoners for stating the maximum (and minimum) bounds of rationality for finite reasoners. I propose the notion of a strictly ideal reasoner which coincides with the notion of maximum rationality. The notion of a strictly ideal reasoner is relative to a logic and a notion of beliefs (explicit, implicit, etc). I argue that, for some relevant logics, a finite reasoner may only approach maximum rationality at the limit of a reasoning sequence (stable beliefs)[1]. In chapter 2, I investigate the use of ideal reasoners in the zombie argument against physicalism (Chalmers, 2010). This notion is used in the principle that ideal negative conceivability entails possibility. The conclusion is that the zombie argument is neither an a priori nor a conclusive argument against physicalism. In chapter 3, I investigate the notion of maximum (and minimum) *epistemic* rationality for finite reasoners. Epistemic rationality is often related to maximizing true beliefs and minimizing false beliefs. I argue that most of the existing models of maximum epistemic rationality have problems in dealing with blindspots and propose a model in terms of the maximization of a function g, which evaluates sets

[1] Informally, the reasoning sequence of a reasoner is how the reasoner would reason from the available information if it had enough cognitive resources (e.g. time for reasoning). The notion of a stable belief is related to the beliefs that the reasoner would hold at the limit of a reasoning sequence.

of beliefs regarding truth/falsehood. However, function g may only be maximized at the limit of a reasoning sequence. In chapter 4, I argue that if maximum (epistemic) rationality for finite reasoners must be understood in terms of the limit of a reasoning sequence, then issues about the computational complexity of reasoning are relevant to epistemology. Then I propose the research program of computational epistemology, which uses computer simulations for investigating maximum (epistemic) rationality for finite reasoners and considers the computational complexity of reasoning. In chapter 5, I provide an example of an investigation in computational epistemology. More specifically, I compare two models of maximum rationality for situations of uncertain reasoning: theory of defeasible reasoning (Pollock, 1995) and Bayesian epistemology (Joyce, 2011).

Acknowledgments

For financial support: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and UCDavis. For epistemological support: Bernard Molyneux, G. J. Mattey, Hanti Lin, Luis Rosa, and people in LLEMMMa group. For mathematical/logical support: Aldo Antonelli, Alexandre Reggiolli, and Cássio Fraga Dantas. For emotional support: Emilly Dias and Taiz Maria Fraga da Silva. Finally, all those important people: Álvaro Ruy de Amorim Dantas, Cássio Fraga Dantas, Dona Risoleta (*in memoriam*), Dona Mira (*in memoriam*), Emilly Dias, Felipe Cerqueira, Laiz Fraga Dantas, Taiz Maria Fraga da Silva, Tânia Queiroz.

Introduction

The notion of an ideal reasoner is used in epistemology for several purposes. First, the notion of an ideal reasoner is used in the interpretation of some epistemic formalisms:

> Modal logic of an ordinary sort, when construed as a logic of belief or judgement, gives a schematic characterization of an ideally rational mind, or, as I shall call it, an ideal knower (Binkley, 1968, p. 127).

Second, the notion of an ideal reasoner is employed in stating epistemic norms:

> A logic provides a characterization of (at least a large subset of) the arguments that we ought to endorse. A logic hence provides an account of how an ideal reasoner reasons (Garfield, 1990, p. 103).

Third, the notion of an ideal reasoner is used in defining other notions in epistemology. For example, the notion of an ideal reasoner is used in defining the notion of warrant:

> Let us say that a proposition is warranted in an epistemic situation iff an ideal reasoner starting from that situation would be justified in believing the proposition (Pollock, 1987, p. 490).

For another example, ideal reasoners are used in the definition of ideal conceivability:

> Given this notion, we can say that S is *ideally* conceivable when the hypothesis expressed by S cannot be ruled out, a priori, even on ideal rational reflection (Chalmers, 2010, p. 143)[2].

In all three cases, an ideal reasoner is used as the parameter of (maximum) rationality. This "ideally rational mind" would endorse all (and only?) the arguments that it ought to endorse and would hold all (and only?) the beliefs which it has justification for holding[3].

Epistemology is usually concerned with the (ir)rationality of finite reasoners (e.g. humans)[4]. In this case, an ideal reasoner is used as the parameter of (maximum) rationality for finite reasoners. I think that this use is problematic for two reasons.

[2] The use of the notion of an ideal reasoner in the definition of ideal conceivability is even more explicit in Menzies (1998, p. 269): "[I]t is possible that p if and only if an ideal conceiver could conceive that p", where 'conceiving p' is understood as 'not being able to know $\neg p$ on a priori basis'.

[3] I develop the connection between ideal conceivability and maximum rationality in chapter 2, note 8.

[4] Roughly, a finite reasoner is a reasoner with cognitive limitations such as finite perceptual input, finite memory, and being able to execute only a finite number of inferential steps in a finite time interval.

The first reason is that the phrase ‘ideal reasoner’ does not mean the same thing in all its uses. There are at least two different notions of an ideal reasoner in the literature. There is a cognitive notion of an ideal reasoner as a reasoner without cognitive limitations:

> An ideal reasoner is unconstrained by a finite memory or processing capacity (Pollock, 1987, p. 504).

This cognitive notion is also found in Chalmers (2010, p. 143) and Menzies (1998).

There also exists an epistemic notion of an ideal reasoner as a reasoner that believes all logical consequences of its beliefs and does not have an inconsistent set of beliefs:

> [The ideal reasoner] avoids contradiction, is aware of all logical truths, and believes all the logical consequences of what he believes (Binkley, 1968).

This epistemic notion is also found in Halpern and Moses (1985), Grim (1988), Duc (1995), Giunchiglia and Giunchiglia (2001), and Stalnaker (2006, p. 179).

The cognitive notion of an ideal reasoner is insufficient as a parameter of rationality unless the notion specifies how that reasoner reasons. This is why most epistemologists that use the cognitive notion of an ideal reasoner presuppose that the ideal reasoner reasons in some way related to the epistemic notion (e.g. Chalmers, 2010, p. 193).

The second reason is that the notion of an ideal reasoner is sometimes construed in such a high degree of idealization that its use as a parameter of (maximum) rationality for finite reasoners is unadvised. This is the case for the cognitive notion of an ideal reasoner (even when it is supplemented by the epistemic notion). How a reasoner without cognitive limitations reasons is so different from how a finite reasoner reasons that the former may hardly be used as a parameter for the latter[5]. Those which defend the cognitive notion of an ideal reasoner seem to presuppose that an ideal reasoner in the epistemic sense must also have unlimited cognitive capacities (e.g. Chalmers, 2010, p. 143). However, I do not think that this is the case. An ideal reasoner in the epistemic sense may be a finite reasoner depending on which notion of beliefs is considered.

[5] For example, if there exists a procedure for checking guesses, a reasoner which is able to perform inferences instantaneously may solve any problem (instantaneously) by generating and checking (a large number of) random guesses. This is hardly a parameter of rationality for finite reasoners.

The possibility of finite ideal reasoners is often overlooked in the literature, but a finite ideal reasoner may be the relevant parameter of maximum rationality for finite reasoners. In this dissertation, I investigate the conditions under which the (epistemic) notion of an ideal reasoner may be used as a parameter of maximum rationality for finite reasoners.

Computational epistemology

This dissertation also has a methodological motivation: to present and justify the research program of computational epistemology (CE), which investigates the parameter of maximum rationality for finite reasoners using computer simulations.

In the last decades, the nature of epistemology has been subject to extensive discussion. More specifically, it has been discussed whether: (a) epistemology is a normative or a descriptive discipline; (b) epistemology is an a priori or an empirical (a posteriori) discipline; (c) epistemology is autonomous or should be seen as part of a different discipline. These question where introduced to mainstream epistemology by Quine's paper 'Epistemology Naturalized' (1969). In this paper, Quine recommends epistemologists to abandon the conceptual analysis and use the empirical methods of cognitive psychology in describing how humans form beliefs (Quine, 1969, p. 75):

> The relation between the meager [sensory] input and the torrential output [of beliefs] is a relation that we are prompted to study for somewhat the same reasons that always prompted epistemology... But a conspicuous difference between old epistemology and the epistemological enterprise in this new psychological setting is that we can now make free use of empirical psychology (Quine, 1969, p. 82-83).

Quine also defends what is called 'the replacement thesis': "epistemology, or something like that, simply falls into place as a chapter of psychology and hence of natural science." (Quine, 1969, p. 82). Then, while traditional epistemology would be a normative, a priori and autonomous discipline, naturalized epistemology would be a descriptive and empirical discipline which is, in fact, a branch of cognitive psychology (non-autonomous).

Traditional epistemologists often argue that epistemology has an essential normative character which naturalized epistemology fails to ground. The idea is that epistemology deals with the notion of knowledge, the notion of knowledge implies justified belief, and justification would be a normative notion. The notion of justification would be normative in the sense of being related to the 'correct' or 'good' use of one's cognitive resources. But, they argue, the methods of empirical psychology can only describe how humans form beliefs, not what justifies them. As a consequence, naturalized epistemology would be a flawed discipline. This is often called 'the normative objection':

> If justification drops out of epistemology, knowledge itself drops out of epistemology. For our concept of knowledge is inseparably tied to that of justification. As earlier noted, knowledge itself is a normative notion. Quine's non-normative, naturalized epistemology has no room for our concept of knowledge (Kim, 1988, p. 389).

Similar objections are found in Sellars (1956, sec. 36) and Siegel (1996).

I think that the traditional epistemologists are correct in this criticism: evaluating beliefs and patterns of inference are essential roles of epistemology, but the fact that humans form beliefs in such and such way does not have normative force.

On the other hand, traditional epistemology is often criticized regarding its methods. In considering an analysis of 'knowledge', traditional epistemologists often describe a situation involving (lack of) knowledge and test how the theory describes the situation against our epistemic intuitions. The paradigmatic examples of this procedure are Gettier's counterexamples to the justified true belief analysis of 'knowledge' (Gettier, 1963). In this sense, intuitions play an evidential role in traditional epistemology.

The evidential role of intuitions in traditional epistemology is under general scrutiny since experimental philosophers presented empirical data suggesting that epistemi intuitions differ amongst individuals (Swain et al., 2008), groups with different ethnicity (Weinberg et al., 2001), or gender (Buckwalter and Stich, 2011), and, therefore, are not reliable evidence. In the literature, there are several attempts to defend the use of intuitions in traditional epistemology from the results of experimental philosophy, either by providing alternative interpretations to the results (Sosa, 2007; Nagel, 2012) or by pointing methodological problems in the experiments which produced the results (Bernstein, 2007; Nagel, 2012). In any case, (i) the evidential role of intuitions is controversial and (ii) epistemology would be in a better position if it employed additional (well-stablished) methods.

Finally, the most controversial claim in Quine (1969) is the replacement thesis. This thesis is endorsed by some epistemologists with naturalistic tendencies (Bishop and Trout, 2005; Kornblith, 1994), but it is rejected by most epistemologists, either with naturalistic (Haack, 1993; Goldman, 1992) or traditional tendencies (Almeder, 1998; BonJour, 1994; Foley, 1994; Fumerton, 1994). The criticism that I have against the replacement thesis is that epistemology seems to have a broader modal domain than cognitive psychology: it is interested not only in actual (ir)rational reasoners, but also in all (ir)rational reasoners. There are problems that are of interest to epistemology but not to cognitive psychology: evaluating patterns of inference, epistemic paradoxes, etc. The reduction of epistemology to cognitive psychology does not help to solve these problems, but only to overlook them.

In sum, traditional and naturalized epistemologies are distinguished over three axes:

	Goal	**Method**	**Reduction?**
Traditional epistemology	normative	a priori	no
Naturalized epistemology	descriptive	empirical	yes

Table 1: Traditional and naturalized epistemologies distinguished over their goal, method, and opinion in relation to epistemology being part of another discipline.

I don't think that having to choose between the two sides in this dispute is very fruitful. I think that both sides point to important traits, but also have serious defects. Also, both sides deny/lack the important traits pointed out by the other side. Traditional epistemology points to the normative role of epistemology and to the autonomy of epistemology, but gives too much weight to epistemic intuitions. Naturalized epistemology points to the necessity of having non-controversial methods and to the importance of taking into account empirical data, but fails to ground the normative role of epistemology and to guarantee the autonomy of epistemology.

In this dissertation, I intend to present and justify the research program of CE. Roughly, an investigation in CE has four stages. First, the choice and description of a relevant class of problems. Problems are collections of questions which are used to evaluate reasoning processes (e.g. chess, sudoku, SAT). Second, the choice and description of hypotheses of agents which would exhibit maximum rationality in dealing with problems in the class. The hypotheses are descriptions of how those agents would reason in solving problems in the class and may have different origins (e.g. existing theories, epistemi intuitions, empirical data). Third, the design and implementation of computer simulations of the agents solving problems in the class. Finally, the analysis of the data from the simulation given some parameters related to (maximum) rationality. The main inovation of CE is the inclusion of the efficiency of reasoning processes (computational complexity) among the parameters of analysis. An argument for the relevance of this parameter to epistemology and a more complete description of an investigation in CE are in chapter 4.

As I see it, CE has the good traits of traditional and naturalized epistemology and does not present their bad traits. In addition, CE provides epistemology with a new kind of consideration (efficiency), which may help enriching long standing disputes in the field.

The general features of CE are:

1. CE is (part of) a normative research program;
2. CE employs well-stablished methods and provides a bridge between methods and results of traditional, naturalized, and formal epistemologies;
3. CE construes epistemology as an autonomous discipline;
4. CE provides epistemology with a new kind of consideration (about efficiency), which may help enriching long standing disputes in the field.

First, (1) CE is (part of) a normative research program. In answering the normative objection, Quine uses what is called 'the engineering reply': epistemology would be best seen as a branch of engineering which is concerned with prescribing the best ways for using one's cognitive resources to achieve one's epistemic goals.

> Naturalization of epistemology does not jettison the normative and settle for the indiscriminate description of ongoing procedures. For me normative epistemology is a branch of engineering. It is the technology of truth-seeking, or, in more cautiously epistemological term, prediction... There is no question here of ultimate value, as in morals; it is a matter of efficacy for an ulterior

> end, truth or prediction. The normative here, as elsewhere in engineering, becomes descriptive when the terminal parameter [the goal] is expressed (Quine, 1998, p. 664-665).

In some sense, engineering is a normative discipline. For example, bridge engineers will apply theories of physics, geology, etc, in order to distinguish between good and bad projects for bridges (Wrenn, 2006, p. 68). In this sense, engineering epistemologists would use empirical data about how the world and the human mind work in order to distinguish between better and worse cognitive practices for achieving epistemic goals. For example, if you have the goal of maximizing true beliefs and minimizing false beliefs, the engineering epistemologist is able to prescribe you to ignore soothsayers based on the available scientific data on how the world and the human mind work (Quine, 1992, p. 19).

The engineering reply often does not persuade traditional epistemologists, who argue that it does not ground normativity in the relevant *categorical* sense (Wrenn, 2006). Epistemology would need to ground categorical normativity in the sense of delivering prescriptions which are independent of epistemic goals of individuals, but engineering epistemology would be only able to ground hypothetical normativity (i.e. dependent of epistemic goals of individuals)[6]. In the example, epistemology would need to be able to state that you ought to maximize true beliefs and minimize false beliefs and not only that you should do such and such *if you have this goal.* Engineering epistemology would not be able to ground categorical normativity because empirical data cannot ground categorical normativity and, given the replacement thesis, it does not have other means to do so.

CE may be seen as a kind of engineering epistemology without the replacement thesis. In this context, CE is able to ground categorical normativity if traditional epistemology is able to do so. This is the case because CE investigates formal properties of patterns of inference, leaving to traditional epistemology the investigation about the content of substantial epistemic norms and concepts. For example, CE investigates the maximum and minimum bounds of rationality for finite reasoners, leaving to traditional epistemology the investigation about the 'amount' of rationality involved in the notion of justification.

[6] Quine himself seems to agree with this interpretation: "There is no question here of ultimate value, as in morals..." (Quine, 1998, p. 664-665).

In addition, traditional epistemology has a role in an investigation in CE in, for example, stating of hypotheses of agents, which are tested using computer simulations.

Second, (2) CE employs well-stablished methods. In fact, when its comes to methods, I think that CE is in a better position than naturalized and traditional epistemology. CE is in a better position than naturalized epistemology because it is not restricted to empirical data about humans. The relevance of empirical data about humans for epistemology is hardly questioned (McCauley, 1988; Haack, 1993; Goldman, 1992)[7] and CE makes use of this kind of data in selecting hypotheses, but the restriction to about humans renders naturalistic epistemology unable to ground the normative character of epistemology. CE is in a better position than traditional epistemology because, in CE, intuitions play only a heuristic role: to state hypotheses to be tested using computer simulations. The use of computer simulations in CE may have good consequences for epistemology in general. First, Cummins (1998) points to the inexistence of an epistemically valuable calibration methods for intuitions. The use of computer simulations may provide such calibration. Second, many theories in traditional epistemology use counterfactuals (Sosa, 1999; Nozick, 1981). In CE, it is possible to evaluate counterfactuals by changing the initial conditions of a simulation. Finally, it is always better to have additional (well-stablished) methods[8].

In addition, CE provides a bridge between the methods and results of traditional, naturalized, and formal epistemologies. In the literature, it is often stressed the difficulty in relating the methods and results of traditional and formal epistemologies (e.g. Hendricks and Symons, 2006). The situation is not better for naturalized epistemology: while cognitive psychology is concerned with real reasoners (humans), formal epistemology deals with abstract reasoners and methods and results about abstract reasoners are difficult to adapt to (much more complex) real reasoners. This is not a good result because epistemic logic and formal epistemology have achieved important advances which would benefit epistemology in general. For example, advancces in modal epistemic logic (Hintikka, 1962), dynamic epistemic logic (van Ditmarsch et al., 2006), and Bayesian epistemology

[7] This position is reasonable, since there are other sciences which study knowledge (or cognition) and which have interesting results for epistemology: cognitive sciences, neuroscience, artificial intelligence.

[8] Similar techniques are used in several disciplines with similar purposes: computational linguistics (Ojeda, 2012), computational psychology (Sun, 2008), game theory (Tesfatsion, 2006), etc.

(Bovens and Hartmann, 2004). This is not a problem for CE. CE uses intuitions, empirical data, and existing epistemological theories in selecting of hypotheses, formal tools in formalizing hypotheses, and computer simulation in testing hypotheses.

Third, (3) CE construes epistemology as an autonomous discipline. It is true that the methods and goals of CE have some resemblance with those of cognitive psychology and AI, but CE is not part of those disciplines. CE and cognitive psychology often investigate reasoning using computer models, but the goals of CE and cognitive psychology are quite different. While the goal of cognitive psychology is to describe the patterns of inference of humans, the goal of CE is to describe the maximum and minimum bounds of rationality for finite reasoners in general (not necessarily humans). The goal of CE have some resemblance with the goal of AI. The goal of both CE and AI is to describe optimal patterns of inference, but not in the same sense. An investigation in AI is often concerned with finding the best strategy for solving specific problems. An investigation in CE, on the other hand, is not concerned with solving specific problems, but uses data about how agents deal with problems in order to investigate the maximum and minimum bounds of rationality for finite reasoners. The goal of CE is the same goal of epistemology in general: to set parameters from which we can evaluate patterns of inference of finite reasoners regarding rationality.

Finally, (4) CE provides epistemology with a new kind of consideration (efficiency), which may help enriching long standing disputes in the field. It may be the case, for example, that there exist two theories that prescribe otherwise equally reasonable patterns of inference which differ only in efficiency. I think that the best argument for (4) is to show a case like that and show how CE may help in such a case. I will make this case in chapter 5, where I present an example of investigation in CE which compares two theories about reasoning under uncertainty: nonmonotonic logic and Bayesian epistemology.

The last point that I want to stress is that most ideas in CE are not new to epistemology. First, the distinction between polynomial and exponential patterns of inference is already considered in the literature on epistemology. For example, Aaronson argues that the gap between polynomial and exponential may provide relevant distinctions for epistemology:

> I think of the polynomial/exponential gap as occupying a ‘middle ground’ between two other sorts of gaps: on the one hand, small quantitative gaps (such as the gap between n steps and $2n$ steps); and on the other hand, the gap between a finite number of steps and an infinite number. The trouble with small quantitative gaps is that they are too sensitive to ‘mundane’ modeling choices an the details of technology. But the gap between finite and infinite has the opposite problem: it is serenely insensitive to distinctions that we actually care about, such as that between finding a solution and verifying it, or between classical and quantum physics. The polynomial/exponential gap avoids both problems (Aaronson, 2011, p. 10).

The investigation of this gap would lead to new perspectives on the nature of mathematical knowledge, the problem of logical omniscience, Hume’s problem of induction, Goodman’s grue riddle, economic rationality, and other topics (Aaronson, 2011, p. 1).

There also exist research programs in epistemology which use computer simulations. For example, the computer agent OSCAR is an implementation of Pollock’s theories about defeasible reasoning, but also of his less well known ideas about intentions, interests, strategies for problem solving, and other cognitive architectural design (see Pollock, 1995).

Finally, there exist other proposals akin to CE in the literature on epistemology. For example, effective epistemology (Kelly, 1988) and android epistemology (Ford et al., 2006):

> On the theoretical side, android epistemology is an open and expanding body of theories and proposals about computational architectures for cognition, theories and proposals some of which are biologically or psychologically inspired, and others of which are free floating − never mind how *people* are able to do x − maybe a computational system could do x *this* way... On the other, more practical side, android epistemology is about how to use such ideas to extend human capacities (Ford et al., 2006, p. viii).

The same is true in the literature on AI, as in the idea of computational intelligence:

> The science of CI [computational intelligence] could be described as ‘synthetic psychology’, ‘experimental philosophy’ or ‘computational epistemology’. It can be seen as a way to study the old problem of the nature of knowledge and intelligence, but with a more powerful experimental tool than was previously available (Poole et al., 1998, p. 6).

However, as far as I know, the field of CE, as a branch of epistemology, still does not have a proper justification and presentation. In this dissertation, I intend to do both.

Overview

The dissertation has two parts. In the fist four chapters, I provide a justification for why CE is a relevant research program in epistemology. In chapters 1 to 3, I argue that finite resoners may only approach maximum rationality at the limit of a reasoning sequence[9]. In chapter 4, I argue that if this is the case, then the computational complexity of reasoning is relevant to epistemology. In this case, I argue, computer simulations would be a relevant tool for epistemology. Finally, I present more formally the research program of CE. In chapter 5, I provide an example of invetigation in CE about reasoning under uncertainty.

In chapter 1, I investigate the use of ideal reasoners for stating the maximum (and minimum) bounds of rationality for finite reasoners. I argue that this notion is unsatisfactory for modeling maximum rationality because an ideal reasoner may have random beliefs[10]. Then I propose the notion of a strictly ideal reasoner, i.e. an ideal reasoner which believes only the logical consequences of its initial beliefs. I argue that the notion of a strictly ideal reasoner coincides with the notion of maximum rationality. The notion of a strictly ideal reasoner is relative to a logic and to a notion of beliefs (explicit, implicit beliefs, etc). I investigate a strictly ideal reasoner for classical logic defined in terms of explicit beliefs. According to this notion, a strictly reasoner explicitly believes infinitely many logical tautologies, which a finite reasoner cannot do. A finite reasoner which does not believe anything is strictly ideal when this notion is defined in terms of implicit beliefs. I propose the notion of accessible beliefs and investigate a strictly ideal reasoner for classical logic defined in terms of accessible beliefs. This notion avoids the problems of the previous two notions. However, a strictly ideal reasoner for a nonmonotonic logic defined in terms of accessible beliefs may need to execute infinitely many inferential steps in a finite time interval, which a finite reasoner cannot do. I conclude that a finite reasoner may only

[9] Informally, the reasoning sequence of a reasoner is how the reasoner would reason from the available information if it had enough cognitive resources (e.g. time for reasoning).

[10] As long as those beliefs are not contradictory. I develop this issue in chapter 1.

approach maximum rationality at the limit of a reasoning sequence. The parameter of maximum rationality for finite reasoner must be a strictly ideal reasoner defined in terms of stable beliefs[11].

In chapter 2, I investigate the use of (strictly) ideal reasoners in the negative zombie argument against physicalism (Chalmers, 2010). The negative zombie argument is such that $p \wedge \neg q$ (the zombie sentence) is ideally negatively conceivable and, therefore, possible, which would entail that physicalism is false[12]. The notion of a strictly ideal reasoner is used as follows: ϕ is ideally negatively conceivable iff a strictly ideal reasoner for a relevant logic $\overset{\mathrm{x}}{\models}$ does not believe $\neg\phi$. I argue that, given this principle, the notion of a strictly ideal reasoner must be understood in terms of stable beliefs and that $\overset{\mathrm{x}}{\models}$ must be nonmonotonic. Then I argue that such a strictly ideal reasoner does not ideally negatively conceive $p \wedge \neg q$ for some choices of q (for example, for $q =$'someone is conscious'). I discuss the consequences of these results for the negative zombie argument and for the zombie arguments in general. The conclusion is that the negative zombie argument (and maybe the zombie arguments in general) is neither an a priori argument nor a conclusive argument against physicalism.

In chapter 3, I investigate the notion of maximum *epistemic* rationality for finite reasoners. In the literature, maximum epistemic rationality is often related to maximizing true beliefs and minimizing false beliefs[13]. I argue that most of the existing models of maximum epistemic rationality have problems in dealing with blindspots (Sorensen, 1988). Then I argue that maximum epistemic rationality is better modeled as the maximization of a function g which accepts a set of beliefs as input and returns a numeric evaluation. Then I discuss the properties of this function[14]. I argue that this model does not have problems with blindspots, but a finite reasoner is only able to maximize function g at the

[11] The stable beliefs are those beliefs that the reasoner would hold at the limit of a reasoning sequence.

[12] In the zombie sentence, p is the conjunction of the fundamental physical truths and laws and q is an arbitrary phenomenal truth. If $q =$'someone is conscious', then $\Diamond(p \wedge \neg q)$ asserts the possibility of a zombie world, i.e. a world which is physically identical to the actual world, but where no one is conscious.

[13] Similar claims are found in Lehrer (2000), Foley (1993), Plantinga (1993), Goldman (1986), Moser (1985), Alston (1985), BonJour (1985) Sosa (1985), and Chisholm (1982). List due to David (2001).

[14] Function g must vary directly as the number of true beliefs (t), vary inversely as the number of false beliefs (f), which entails that the function is defined for all values of t and f, and have an upper (lower) bound in the form of a supremum (infimum) but not a maximum (minimum).

limit of a reasoning sequence. In the discussion, I compare the function g model with other quantitative models of rationality, as, for example, the group of models usually known as 'epistemic utility theory' (Fitelson and Easwaran, 2015; Pettigrew, 2015a; Joyce, 1998).

In chapter 4, I argue that if finite reasoners are only able to approach maximum (epistemic) rationality at the limit of a reasoning sequence, then issues about the computational complexity of reasoning are relevant to epistemology. The argument exploits the facts that a finite reasoner cannot reach actual maximum (epistemic) rationality and that there exists a relevant difference in how a finite reasoner may approach maximum (epistemic) rationality depending on whether it has a polynomial or an exponential reasoning sequence. Then I present more formally the research program of CE. I discuss the formalization of problems, agents, the general features of the computer simulations, present and motivate the parameters used in the analysis of the data from the simulations.

In chapter 5, I provide an example of an investigation in computational epistemology. More specifically, I compare two models of maximum rationality in situations of uncertain reasoning: the theory of defeasible reasoning (Pollock, 1995) and Bayesian epistemology (Joyce, 2011). The goal is to show how considerations about computational complexity may help enrich disputes in epistemology. In order to compare these theories, I introduce an epistemic version of the Wumpus World, a class of problems described in Russell and Norvig (2010) as a tool for investigating uncertain reasoning. I design agents based on these models, implement these agents in a computer simulation of the epistemic Wumpus World, and analyze the data from the simulation. I analyze how much these agents succeed in solving problems in that class of problems, the cost of their solutions, and the computational complexity of their reasoning processes. The conclusion is that, under the conditions of the epistemic Wumpus World, the theory of defeasible reasoning provides a better model of uncertain reasoning. I discuss the general conditions under which these results hold.

The appendices contain some proofs, a survey on computational complexity theory, and a modal interpretation for the notion of accessible beliefs introduced in chapter 1.

Chapter 1

The limits of finite reasoning

> The ideal reasoner would, when he had once been shown a single fact in all its bearings, deduce from it not only all the chain of events which led up to it but also all the results which would follow from it.
> Sherlock Holmes, *The Five Orange Pips*.

Consider the following goals for a research program in analytic epistemology (AE):

(g1) investigate which are the epistemic norms for finite reasoners (e.g. humans)[1];

(g2) provide formalizations of epistemic practices (e.g. patterns of inference) in a way which contributes to the advancement of AE.

I think that both (g1) and (g2) are reasonable goals for a research program in AE. About (g1), the investigation of the notion of epistemic justification is usually accepted as a major goal of AE (Chisholm, 1977). But it is also usually accepted that the notion of epistemic justification is closely related to the fulfillment of epistemic norms: for example, some epistemologists claim that a belief is justified iff it is permissible given the correct epistemic norms (Pollock and Cruz, 1999, p. 123). Then the investigation of epistemic norms is already a goal of AE. In addition, AE is usually concerned with the epistemic practices of humans (who are finite reasoners): skeptical arguments often exploit the cognitive limitations of humans and epistemologists often exploit these limitations in their

[1] Roughly, a finite reasoner is a reasoner with cognitive limitations such as finite perceptual input, finite memory, and being able to execute only a finite number of inferential steps in a finite time interval.

theories[2]. For these reasons, I think that it is a reasonable goal for a research program in AE to (g1) investigate which are the epistemic norms for finite reasoners.

About (g2), it is usually accepted that to provide precise analyses of the concepts of knowledge, belief, and epistemic justification is an important goal of AE. In providing such analyses, epistemologists often employ formal reconstructions of concepts. For example, modal logic is often employed in the formal reconstruction of the concepts of knowledge, belief, and epistemic justification (Halpern et al., 2009). In the last decades, it has been argued that AE should also be concerned with the formal reconstruction of epistemic practices (Pollock, 1998). Pollock argues that there exist at least two notions of epistemic justification which are of interest to AE. The first is epistemic justification as "what turns true belief into knowledge", which concerns the traditional conceptual analysis. But, Pollock argues, there also exists a notion of epistemic justification as "the correct ways of reasoning", which concerns procedural epistemology.

> Viewed from this perspective, we can roughly divide the interests of epistemology into procedural and descriptive epistemology. Procedural epistemology is directed at how to build the system of cognition, whereas descriptive epistemology concerns how to describe what the system is doing once it is running. Thus rules for reasoning become part of procedural epistemology, but the analyses of "*S* knows that *P*" is assigned to descriptive epistemology (Pollock, 1998, p. 18).

Then to provide formalizations of epistemic practices is a goal of (procedural) AE. The clause 'in a way which contributes to the advancement of AE' is in place because it is a requirement for using a formal tool in some field of research that the use contributes to the advancement that field; otherwise, it is only an empty use. For these reasons, I think that it is a goal of a research program in AE to (g2) provide formalizations of epistemic practices in a way which contributes to the advancement of AE.

[2] For example, epistemologists often argue from the fact that finite reasoners (in fact, humans) do not make inferences instantaneously: "Can we argue that knowledge distribute over conjunction? It would scarcely be relevant to argue that one will always infer the conjuncts from the conjunction, for, since making an inference takes time, one might still violate distributivity before completing the inference. If knowledge of the conjunction causes knowledge of the conjuncts, and the cause is not simultaneous with the effect, then for an intermediate period one would know the conjunction without knowing the conjuncts; that period would be a counterexample to distribution" (Williamson, 2000, p. 281).

In this chapter, I investigate which notion of an ideal reasoner may be used in a research program about stating the maximum and minimum bounds of rationality *for finite reasoners* as to fulfill goals (g1) and (g2)[3]. Due to (g1), the relevant notion of an ideal reasoner must be able to figure in some *informative* epistemic norms about which beliefs a finite reasoner ought to hold and may hold in order to be maximally rational and which beliefs a finite reasoner may hold in order to be minimally rational[4]. In order to be informative, the epistemic norms in the research program cannot be known a priori to be fulfilled/not-fulfilled by the whole class of finite reasoners. Due to (g2), the notion of an ideal reasoner must contribute to the advancement of AE in the sense of generating notions of maximum and minimum rationality for finite reasoners which may be investigated using the best (most fruitful) methods available.

In section 1.1, I propose a model of a reasoner which is used in most definitions in the chapter. For example, this model is used in the definition of a finite reasoner, an ideal reasoner, the epistemic norms, etc. I then investigate some notions of an ideal reasoner which are common in the literature and argue that these notions are not adequate models of maximum rationality for finite reasoners. I propose the related notion of a strictly ideal reasoner and argue that this notion is an adequate model of maximum rationality in our framework. In section 1.2, I argue that the notion of a strictly ideal reasoner is relative to a logic and to a notion of beliefs. Then I investigate a strictly ideal reasoner for classical logic defined in terms of explicit, implicit, and accessible beliefs. I reject the definitions in terms of explicit and implicit beliefs on the basis of goal (g1). In section 1.3, I investigate a strictly ideal reasoner for a nonmonotonic logic with some essential properties defined in terms of accessible beliefs and reject this definition on the basis of goal (g2). In section 1.4, I introduce the notion of stable beliefs and use this notion in the definition of strictly ideal reasoners for both classical and nonmonotonic logic. I argue that the definitions in terms of stable beliefs fulfill goals (g1) and (g2) and should be adopted in the research program.

[3] In the following, I refer to this research program as ‘the research program’.

[4] In the following, ‘may believe ϕ’ means that believing ϕ is (epistemically) permissible, ‘ought to believe ϕ’ means that believing ϕ is mandatory, ‘may-not believe ϕ’ means that believing ϕ is impermissible, and ‘may believe not-ϕ’ means that believing not-ϕ is permissible.

1.1 Definitions

The model of a reasoner $\mathcal{R}$ is composed of a formal language ($\mathcal{L}$), an input ($\texttt{INPUT}$), a knowledge base ($\texttt{KB}$), and a pattern of inference (π). $\texttt{INPUT}$ and $\texttt{KB}$ are sets of sentences in $\mathcal{L}$ and π is a function which has two sets of sentences in $\mathcal{L}$ ($\texttt{INPUT}$ and $\texttt{KB}$) and a positive integer as inputs and outputs a set of sentences in $\mathcal{L}$ (an updated $\texttt{KB}$)[5]. Roughly, the sentences in $\texttt{INPUT}$ model the perceptual input of the reasoner, the sentences in $\texttt{KB}$ model the (explicit) beliefs of the reasoner (memory), and function π models the dispositions of the reasoner about how to reason from the available information ($\texttt{INPUT}$ and $\texttt{KB}$). Consider the model of a reasoner:

Definition 1.1.1 (Reasoner ($\boldsymbol{\mathcal{R}}$)). A reasoner $\mathcal{R} = \langle \mathcal{L}, \texttt{INPUT}, \texttt{KB}, \pi \rangle$ is a 4-tuple, where $\mathcal{L}$ is a formal language, $\texttt{INPUT}$ and $\texttt{KB}$ are sets of sentences in $\mathcal{L}$, and π is a function $\pi : 2^{\mathcal{L}} \times 2^{\mathcal{L}} \times \mathbb{Z}^+ \to 2^{\mathcal{L}}$.

The work of function π may be thought of as making a copy of $\texttt{KB}$, executing some operations on that copy, and returning an updated $\texttt{KB}$. The operations executed by function π may either add or delete sentences to/from the copy of $\texttt{KB}$ given some conditions. For example, consider the operation scheme $\frac{\phi \in \texttt{INPUT}, \neg\phi \notin \texttt{KB}}{\text{add } \phi \text{ to } \texttt{KB}}$ (if ϕ is in $\texttt{INPUT}$ and $\neg\phi$ is not in $\texttt{KB}$, then add ϕ to $\texttt{KB}$). In the following, the execution of function π for a given input is the sequence of instantiations of operation scheme for terms and sentences in $\mathcal{L}$ that the function executes in order to go from the input to the corresponding output. The execution of function π for a given input is finite iff the respective sequence is finite. The definition of function π is a (full) description of the general work of the function that specifies which execution is produced for each possible input (e.g. a computer program). The definition of function π is finite iff there exists a finite description of such kind[6].

In this context, the notion of a finite reasoner is the following:

[5] The phrase 'knowledge base' denotes (as in AI) the module of a reasoner which stores information and is not directly related to the epistemic notion of knowledge. I use the term 'function π' for readability. Actually, π is a set of instructions for calculating a function. The numeric parameter of π models the fact that a reasoner may perform different inferences from the same premises (see sections 1.2 and 1.4).

[6] For example, consider the function *even*, which accepts a natural number as input and returns 1 if the input is even and 0 otherwise. Then function *even* is the set $\{\langle 0,1\rangle, \langle 1,0\rangle, \langle 2,1\rangle, \ldots\}$. The definition of *even* is finite: 'if (input mod 2 = 0), return 1; else return 0'. The execution of *even* for any input is also finite: for any input, function *even* executes only one modulo operation (input mod 2).

Definition 1.1.2 (Finite reasoner). A reasoner $\mathcal{R} = \langle \mathcal{L}, \texttt{INPUT}, \texttt{KB}, \pi \rangle$ is finite iff $\mathcal{L}$ is finitely specifiable, the sentences in INPUT and KB and those sets are finite, the definition of π is finite, and the execution of π is finite for the relevant inputs[7].

The language $\mathcal{L}$ being finitely specifiable is a requirement for a finite reasoner to learn $\mathcal{L}$ (see Davidson, 1965). The sentences in INPUT and INPUT itself being finite model finite perceptual input. The sentences in KB and KB itself being finite model finite memory. The definition of function π being finite is also a consequence of finite memory. The execution of function π being finite for the relevant inputs models the fact that finite reasoners may execute only a finite number of inferential steps in a finite time interval. For simplicity, I deal mostly with a priori reasoners in this chapter. An a priori reasoner is a reasoner with $\texttt{INPUT} = \varnothing$. Then I will stop referring to INPUT and partially to $\mathcal{L}$[8]. I will treat reasoners as pairs $\langle \texttt{KB}, \pi \rangle$ with $\pi : 2^{\mathcal{L}} \times \mathbb{Z}^{+} \rightarrow 2^{\mathcal{L}}$.

There exist at least two different notions of an ideal reasoner in the literature. The first notion is a cognitive notion of an ideal reasoner as a reasoner without cognitive limitations (Chalmers, 2010; Pollock, 1987). The cognitive notion of an ideal reasoner is not adequate for modeling maximum rationality for finite reasoners for two reasons. The first reason is that a purely cognitive notion of an ideal reasoner is insufficient for describing maximum rationality unless the notion specifies how that reasoner reasons. The second reason is that the patterns of inference of a reasoner without cognitive limitations are fundamentally different from those of a finite reasoner. For example, if there is a procedure for checking guesses, a reasoner which is able to perform inferences instantaneously solves any problem (instantaneously) simply by generating and checking (a large number of) random guesses.

The second is an epistemic notion of an ideal reasoner:

Definition 1.1.3 (Ideal reasoner). A reasoner $\mathcal{R}$ is ideal iff

(i) $\mathcal{R}$ believes all logical consequences of its epistemic situation;

(ii) $\mathcal{R}$ has a nontrivial set of beliefs,

[7] The relevant inputs are those used in generating the beliefs of the reasoner (see sections 1.2 and 1.4).

[8] In the following, I always discuss reasoners in the context of a logic. Then I will always presuppose that $\mathcal{L}$ is an arbitrary language for that logic (with a finite number of semantic primitives).

where the epistemic situation of an a priori reasoner is its initial `KB`.

In discussing the surprise test paradox, Binkley (1968) proposes a notion of an ideal reasoner with these requirements. Also do Halpern and Moses (1985), Grim (1988), Duc (1995), Giunchiglia and Giunchiglia (2001), and Stalnaker (2006)[9]. The epistemic notion of an ideal reasoner is not adequate for modeling maximum rationality because a reasoner with features (i) and (ii) may still have all sorts of random beliefs[10]. In other words, definition 1.1.3 has a very strong completeness requirement (closure), but only a very weak soundness requirement (nontriviality)[11].

As much as the notion of an ideal reasoner is not an adequate model of maximum rationality for finite reasoners, that notion may still be used in the research program in defining adequate notions of maximum and minimum rationality for finite reasoners. In standard deontic logic, the notions of ought and may are modeled as 'ought ϕ' being true iff ϕ is true in all accessible (relevant) possible worlds and 'may ϕ' being true iff ϕ is true in some accessible (relevant) possible worlds (see von Wright, 1951)[12]. The relevant possible worlds in deontic logic are often referred as "ideal worlds" or "deontically perfect worlds" (see von Wright, 1986). In this case, the notion of an ideal reasoner may be used in a related definition of the epistemic notions of ought and may in which ideal reasoners play the role of relevant possible worlds and the relation 'being in the same epistemic situation' is the accessibility relation.

Definition 1.1.4 (Epistemic ought). A reasoner ought to believe ϕ iff all ideal reasoners in the same epistemic situation believe ϕ.

[9] In the literature, requirement (ii) is usually expressed as '$\mathcal{R}$ has a consistent set of beliefs'. A set of beliefs is consistent iff it does not entail a contradiction. A set of beliefs is nontrivial iff it does not entail every sentence in $\mathcal{L}$. If we accept the principle of explosion (see section 1.3), these requirements coincide. Grim, Duc and Giunchiglia and Giunchiglia drop requirement (ii), but this requirement is important for blocking a (fully credulous) reasoner which believes all sentences in the language from being ideal.

[10] As long as those beliefs do not contradict requirement (ii). It may be difficult to make sense of this use of 'random beliefs' when 'beliefs' are understood as explicit beliefs. If explicit beliefs are modeled as the sentenes in `KB`, then an ideal reasoner would not be able to have random beliefs without believing all of their logical consequences. In sections 1.2 and 1.4, I introduce other notions of beliefs for which this use of 'random beliefs' makes sense (accessible and stable beliefs).

[11] This explanation was suggested by an anonymous reviewer. In the following, 'ideal reasoner' refers to the epistemic notion of an ideal reasoner in definition 1.1.3.

[12] In other words, ought is box-like and may is diamond-like.

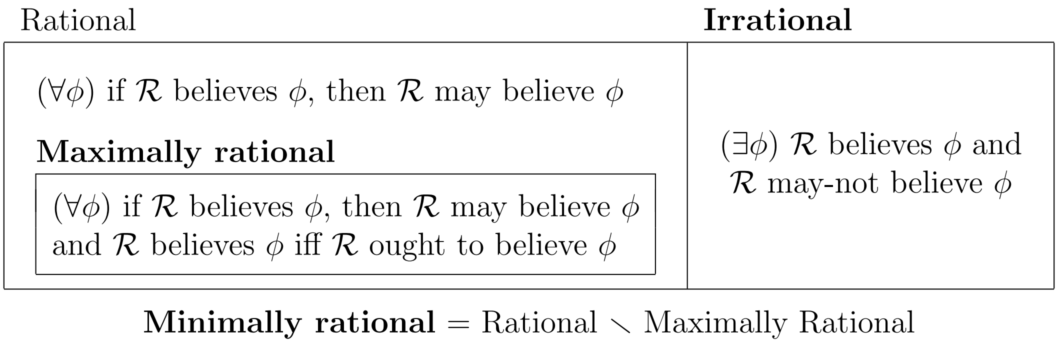

Rational
Irrational
$(\forall\phi)$ if $\mathcal{R}$ believes ϕ, then $\mathcal{R}$ may believe ϕ
Maximally rational
$(\forall\phi)$ if $\mathcal{R}$ believes ϕ, then $\mathcal{R}$ may believe ϕ and $\mathcal{R}$ believes ϕ iff $\mathcal{R}$ ought to believe ϕ
$(\exists\phi)$ $\mathcal{R}$ believes ϕ and $\mathcal{R}$ may-not believe ϕ
Minimally rational = Rational $\setminus$ Maximally Rational

Figure 1.1: Mutually exclusive classes of reasoners: maximally and minimally and irrational reasoners. Minimally rational reasoners are rational but not maximally rational. According to goal (g1), none of these classes may be a priori empty for finite reasoners.

Definition 1.1.5 (Epistemic may). A reasoner may believe ϕ iff there exists an ideal reasoner in the same epistemic situation which believes ϕ.

The epistemic notions of ought and may would be used in stating the notions of a maximally rational, a minimally rational, and an irrational reasoner as following[13]:

Definition 1.1.6 (Maximally rational). A reasoner $\mathcal{R}$ is maximally rational iff, for all ϕ, $\mathcal{R}$ believes ϕ iff $\mathcal{R}$ ought to believe ϕ and $\mathcal{R}$ believes ϕ only if $\mathcal{R}$ may believe ϕ.

Definition 1.1.7 (Minimally rational). A reasoner $\mathcal{R}$ is minimally rational iff, for all ϕ, $\mathcal{R}$ believes ϕ only if $\mathcal{R}$ may believe ϕ and $\mathcal{R}$ is not maximally rational.

Definition 1.1.8 (Irrational). A reasoner $\mathcal{R}$ is irrational iff, for some ϕ, $\mathcal{R}$ believes ϕ and $\mathcal{R}$ may-not believe ϕ.

Maximally rational reasoners fulfill requirements (i) and (ii) for ideal reasoners (see proof C.1). In addition, they believe only the logical consequences of their epistemic situation. Then all maximally rational reasoners are ideal reasoners, but the opposite is not true[14]. Minimally rational reasoners fulfill requirement (ii) and may also fulfill requirement (i) as long as they do not believe only the logical consequences of their epistemic situation. Then some (but not all) minimally rational reasoners are ideal reasoners and

[13] The definitions have the consequence that a maximally rational reasoner is not minimally rational. It would be more natural to talk about, respectively, perfect and imperfect rationality. I do not use these terms because I am interested in the maximum and minimal bounds of rationality for finite reasoners.

[14] The ideal reasoners with 'all sorts of random beliefs' discussed above are not maximally rational.

vice versa. Irrational reasoners do not fulfill requirement (ii), independently of fulfilling requirement (i). According to goal (g1), none of these three classes may be known a priori to be empty for the whole class of finite reasoners[15].

How a maximally rational reasoner relates to an ideal reasoner suggests the notion of a *strictly* ideal reasoner which is an adequate model of maximum rationality. Roughly, a strictly ideal reasoner is an ideal reasoner which believes only the logical consequences of its epistemic situation. The definition of a strictly ideal reasoner is the following:

Definition 1.1.9 (Strictly ideal reasoner). A reasoner $\mathcal{R}$ is strictly ideal iff

(i) $\mathcal{R}$ believes all and only the logical consequences of its epistemic situation;

(ii) $\mathcal{R}$ has a nontrivial set of beliefs.

A reasoner $\mathcal{R}$ is maximally rational iff $\mathcal{R}$ is strictly ideal (see proof C.2). This seems to be a good reason to think that the notion of a strictly ideal reasoner is an adequate model of maximum rationality. If a notion of an ideal reasoner generates adequate models for the maximum and minimum bounds of rationality *for finite reasoners* depends on whether this notion fulfills goal (g1) of the research program.

1.2 Classical reasoners

The notions of an ideal and of a strictly ideal reasoner are not precise for two reasons. First, these notions are relative to a logic because the notions of logical consequence and triviality are relative to a logic. Second, the term 'belief' is vague in these definitions because there exist several notions of beliefs in the literature: explicit, implicit beliefs, etc[16]. In this section, I am concerned with (strictly) ideal reasoners for classical logic, where 'classical logic' means first-order logic. I investigate which notion of beliefs generates a notion of a classical ideal reasoner which fulfills goal (g1) of the research program[17].

Informally, a reasoner explicitly believes ϕ iff a token of ϕ is stored in its memory. Consider the notion of explicit beliefs in the model:

[15] This goal is related to the discussion about epistemic ought implying cognitive can (see Neta, 2013).

[16] In fact, the notions of an ideal reasoner, a strictly ideal reasoner, the epistemic ought and may, a maximally, a minimally and an irrational reasoner are all relative to a logic and to a notion of beliefs.

[17] In the following, $\overset{x}{\models}$ and $\overset{c}{\models}$ are, respectively, an arbitrary and the classical consequence relation.

Definition 1.2.1 (Explicit belief (belief$_{ex}$)). A reasoner $\mathcal{R} = \langle \mathtt{KB}, \pi \rangle$ believes$_{ex}$ ϕ iff $\phi \in \mathtt{KB}$.

It follows from definitions 1.2.1 and 1.1.2 that no finite reasoner has infinitely many beliefs$_{ex}$. This is the case because having infinitely many beliefs$_{ex}$ requires having infinitely many sentences in $\mathtt{KB}$ (def. 1.2.1, beliefs$_{ex}$), but a finite reasoner cannot have infinitely many sentences in $\mathtt{KB}$ (def. 1.1.2, $\mathtt{KB}$ is finite). I will refer to this fact as 'finite memory'.

This is the definition of an ideal reasoner for $\overset{\text{x}}{\models}$ defined in terms of beliefs$_{ex}$[18]

Definition 1.2.2 (Ideal$_{ex}$ reasoner for $\overset{\text{x}}{\models}$). A reasoner $\mathcal{R} = \langle \mathtt{KB}, \pi \rangle$ is ideal$_{ex}$ for $\overset{\text{x}}{\models}$ iff

(i) $\mathcal{R}$ believes$_{ex}$ all logical consequences of its epistemic situation ($\mathtt{KB} \overset{\text{x}}{\models} \phi \rightarrow \phi \in \mathtt{KB}$);

(ii) $\mathcal{R}$ has a nontrivial set of beliefs$_{ex}$ ($\exists \phi (\mathtt{KB} \overset{\text{x}}{\not\models} \phi)$).

Using the notion of an ideal$_{ex}$ reasoner for $\overset{\text{c}}{\models}$ (classical ideal$_{ex}$ reasoner) in the research program has the consequence of making the class of maximally rational finite reasoners a priori empty, which conflicts with goal (g1)[19]. Suppose that there exist a maximally rational finite reasoner $\mathcal{R}$. Then $\mathcal{R}$ is strictly ideal$_{ex}$ for $\overset{\text{c}}{\models}$. There exist infinitely many logical theorems in $\overset{\text{c}}{\models}$. Then $\mathcal{R}$ has infinitely many beliefs$_{ex}$ (def. 1.2.2, ideal$_{ex}$). Therefore $\mathcal{R}$ is not a finite reasoner (finite memory), which is a contradiction. Therefore, there does not exist a maximally rational finite reasoner in this model[20].

This problem may be avoided using a different notion of beliefs in the definition of a classical ideal reasoner. For example, there exist a notion of implicit beliefs in the

[18] The phrase 'logical consequence' in the definition of an ideal reasoner may be understood using syntatic ($\vdash$) or semantic consequence ($\models$). I do not focus on this distinction because I am mostly dealing with sound and complete logical systems. I acknowledge that there may be advantages in using $\vdash$. For example, this would render possible to talk about ideal reasoners for logical systems which are not (cannot be) sound and complete. My choice of using $\models$ is due to the fact that most discussions abotu properties of logical systems are done in terms of $\models$ (see section 1.3 and chapter 2).

[19] This problem is closely related to the problem of logical omniscience as discussed in the literature (see Stalnaker, 1991, 2006; Fagin et al., 1995, ch. 9).

[20] Alternatively, suppose that there exists a maximally rational finite reasoner $\mathcal{R}$. There exist infinitely many logical theorems in $\overset{\text{c}}{\models}$. Then all ideal$_{ex}$ reasoners for $\overset{\text{c}}{\models}$ believe$_{ex}$ the infinitely many logical theorems in $\overset{\text{c}}{\models}$ (def. 1.2.2, ideal$_{ex}$). Then $\mathcal{R}$ ought to believe$_{ex}$ the infinitely many logical theorems in $\overset{\text{c}}{\models}$ (def. 1.1.4, ought). If $\mathcal{R}$ ought to believe$_{ex}$ ϕ, then $\mathcal{R}$ believes$_{ex}$ ϕ (def. 1.1.6, maximally rational). Then $\mathcal{R}$ has infinitely many beliefs$_{ex}$. Therefore $\mathcal{R}$ not a finite reasoner (finite memory), what is a contradiction. Another problem with defining an ideal reasoner in terms of beliefs$_{ex}$ is that, in this definition, there is no difference between an ideal$_{ex}$ reasoner and a strictly ideal$_{ex}$ reasoner.

literature which is often used in dealing with similar problems (Fagin et al., 1995, p. 363). Informally, a reasoner implicitly believes ϕ iff ϕ is a logical consequence of its beliefs$_{ex}$. Consider the notion of implicit beliefs in the model:

Definition 1.2.3 (Implicit belief (belief$_{im}$)). A reasoner $\mathcal{R} = \langle \mathtt{KB}, \pi \rangle$ believes$_{im}$ ϕ iff $\mathtt{KB} \overset{\mathrm{x}}{\models} \phi$.

This is the definition of an ideal reasoner for $\overset{\mathrm{x}}{\models}$ in terms of beliefs$_{im}$:

Definition 1.2.4 (Ideal$_{im}$ reasoner for $\overset{\mathrm{x}}{\models}$). A reasoner $\mathcal{R}=\langle \mathtt{KB}, \pi \rangle$ is ideal$_{im}$ for $\overset{\mathrm{x}}{\models}$ iff

(i) $\mathcal{R}$ believes$_{im}$ all logical consequences of its epistemic situation ($\mathtt{KB} \overset{\mathrm{x}}{\models} \phi \rightarrow \mathtt{KB} \overset{\mathrm{x}}{\models} \phi$);

(ii) $\mathcal{R}$ has a nontrivial set of beliefs$_{im}$ ($\exists \phi(\{\psi \mid \mathtt{KB} \overset{\mathrm{x}}{\models} \psi\} \overset{\mathrm{x}}{\not\models} \phi)$)[21].

Using the notion of an ideal$_{im}$ reasoner for $\overset{\mathrm{c}}{\models}$ in the research program does not have the consequence of making the class of maximally rational finite reasoners a priori empty. But avoiding this problem comes with the price of 'trivializing' the notion of a (strictly) ideal reasoner: requirement (i) of the notion of an ideal$_{im}$ reasoner is obviously trivial. For this reason, using the notion of an ideal$_{im}$ reasoner for $\overset{\mathrm{c}}{\models}$ in the research program has the consequence of conflating rational and maximally rational reasoners, leaving the class of minimally rational finite reasoners a priori empty, which conflicts with goal (g1)[22].

Using the notion of accessible beliefs in the research program avoids both problems. Informally, a reasoner has the accessible belief that ϕ iff the reasoner may believe$_{ex}$ ϕ after some amount of reasoning. Consider the notion of accessible beliefs in the model:

Definition 1.2.5 (Accessible belief (belief$_{ac}$)). A reasoner $\mathcal{R} = \langle \mathtt{KB}, \pi \rangle$ believes$_{ac}$ ϕ iff $\phi \in \pi(\mathtt{KB})$.

The set of beliefs$_{ac}$, $\pi(\mathtt{KB})$, is the union of the outputs of function π for $\mathtt{KB}$ and all positive integers i. Formally, $\pi(\mathtt{KB}) = \bigcup \pi(\mathtt{KB}, i)$.

This is the definition of an ideal reasoner for $\overset{\mathrm{x}}{\models}$ in terms of beliefs$_{ac}$:

[21] If $\overset{\mathrm{x}}{\models}$ exhibits cautious monotony, then $\exists \phi(\{\psi \mid \mathtt{KB} \overset{\mathrm{x}}{\models} \psi\} \overset{\mathrm{x}}{\not\models} \phi)$ entails $\exists \phi(\mathtt{KB} \overset{\mathrm{x}}{\not\models} \phi)$ (see section 1.3).

[22] In this model, all reasoners with nontrivial set of beliefs$_{im}$ (including all with $\mathtt{KB} = \varnothing$) are maximally rational and all others are irrational. A reasoner with $\mathtt{KB} = \varnothing$ is strictly ideal$_{im}$ for every $\overset{\mathrm{x}}{\models}$.

Definition 1.2.6 (Ideal$_{ac}$ reasoner for $\overset{\mathrm{x}}{\models}$). A reasoner $\mathcal{R} = \langle \texttt{KB}, \pi \rangle$ is ideal$_{ac}$ for $\overset{\mathrm{x}}{\models}$ iff

(i) $\mathcal{R}$ believes$_{ac}$ all logical consequences of its epistemic situation ($\texttt{KB} \overset{\mathrm{x}}{\models} \phi \rightarrow \phi \in \pi(\texttt{KB})$);

(ii) $\mathcal{R}$ has a nontrivial set of beliefs$_{ac}$ ($\exists \phi (\pi(\texttt{KB}) \overset{\mathrm{x}}{\not\models} \phi)$).

The notion of an ideal$_{ac}$ reasoner for $\overset{\mathrm{c}}{\models}$ allows for the existence of finite reasoners which are maximally rational, minimally rational, and irrational in the research program. Consider a reasoner $\mathcal{R}_c = \langle \texttt{KB}, \pi \rangle$ and an axiomatic system for $\overset{\mathrm{c}}{\models}$ (e.g. Antonelli, 2015)[23]. Make $\texttt{KB} = \varnothing$. For each axiom schema in the axiomatization, consider an operation schema with no premises and with the axiom schema as conclusion. For example, if the axiomatization contains the schema $x = x$, consider the operation schema $\frac{\top}{\text{add } x=x \text{ to } \texttt{KB}}$. For each rule in the axiomatization, consider an operation schema with the same premises and conclusion. For example, if the axiomatization contains the rule $\frac{\phi, \phi \rightarrow \psi}{\psi}$, consider the operation schema $\frac{\phi \in \texttt{KB}, \phi \rightarrow \psi \in \texttt{KB}}{\text{add } \psi \text{ to } \texttt{KB}}$. Consider some ordering for the terms and the sentences in $\mathcal{L}$. In this context, the execution of function π for $\texttt{KB}$ and some positive integer i may be (roughly) defined as follows: for $j = 0$ to $j = i$, (i) execute all operation scheme without premises instantiated to all possible combinations of the terms and sentences with positions $\leq j$ then (ii) execute all operation scheme with premises instantiated for all possible combinations of sentences in $\texttt{KB}$ before step (ii) and terms with positions $\leq j$[24].

The reasoner $\mathcal{R}_c$ is finite. Since $\texttt{KB} = \varnothing$, all sentences in $\texttt{KB}$ and $\texttt{KB}$ itself are finite. The definition of function π is also finite (c.f. last paragraph). The execution of π for $\texttt{KB}$ and an arbitrary integer i is finite. For $i = 1$, the number of operations executed by π is a product of the number of operation schema without premises, the number of operation schema with premises, the number of sentences in $\texttt{KB}$ before step (ii) and i. Since the number of operation schema is finite (the number of axiom schema and rules in the axiomatization is finite), this number is finite. For $i = n + 1$, the number of executed operations is the number of operations for $i = n$ plus the number of operations for $i = n + 1$. This number is finite for the same reasons as before. Therefore, $\mathcal{R}_c$ is finite.

[23] The restriction to axiomatic systems is due to the fact that dealing with hypothetical derivations would require some adaptations in the model, but this restriction is not crucial for the argument.

[24] The existence of such a function is a consequence of theoremhood in classical logic being recursively enumerable (see Church, 1936b).

The reasoner $\mathcal{R}_c$ is maximally rational because it is strictly ideal$_{ac}$ for $\overset{c}{\models}$. This is the case because function π generates all and only the theorems of classical logic for the inputs $\texttt{KB} = \varnothing$ and a positive integer. The loop generates all theorems of classical logic because it executes all rules of classical logic for all combinations of sentences in $\mathcal{L}$ and the axiomatization is complete (Gödel, 1930)[25]. The loop generates only the theorems of classical logic because the axiomatization is sound (Gödel, 1930). $\mathcal{R}_c$ is maximally rational because (i) $\mathcal{R}_c$ believes$_{ac}$ all and only the logical consequences of its epistemic situation (the tautologies) and (ii) has a nontrivial set of beliefs$_{ac}$ (e.g. $\perp$ is not a tautology)[26]. Then $\mathcal{R}_c$ is a maximally rational finite reasoner. From $\mathcal{R}_c$ it is easy to construct a minimally rational (not considering an operation schema) and an irrational finite reasoner (adding $\perp$ to $\texttt{KB}$). Then using the notion of an ideal$_{ac}$ reasoner for $\overset{c}{\models}$ fulfills goal (g1).

1.3 Nonmonotonic reasoners

The main difference between classical logic and a nonmonotonic (first-order) logic is that classical logic is monotonic. In other words, the consequence relation $\overset{c}{\models}$ has the property:

(p0) **Monotony:** if $\texttt{KB} \overset{c}{\models} \phi$, then $\texttt{KB} \cup \texttt{INPUT} \overset{c}{\models} \phi$,

where $\texttt{INPUT}$ is an arbitrary set of sentences. Monotony states that if ϕ follows from $\texttt{KB}$, then ϕ also follows from the larger set $\texttt{KB} \cup \texttt{INPUT}$. Classical logic is monotonic because if ϕ is true in every model in which all sentences in $\texttt{KB}$ are true, then ϕ is true in every model in which all sentences in the larger set $\texttt{KB} \cup \texttt{INPUT}$ are true.

Some features of $\overset{c}{\models}$ restrict the epistemic situations for which a strictly ideal reasoner for $\overset{c}{\models}$ is an adequate model for maximum rationality. First, the principle of explosion ($\texttt{KB} \overset{x}{\models} \perp \rightarrow \forall\phi(\texttt{KB} \overset{x}{\models} \phi)$) being true of $\overset{c}{\models}$ precludes the existence of strictly ideal reasoners for $\overset{c}{\models}$ with inconsistent epistemic situations (i.e. $\texttt{KB} \overset{c}{\models} \perp$). Consider, for example, a classical strictly ideal$_{ex}$ reasoner $\mathcal{R}$ with $\texttt{KB} \overset{c}{\models} \perp$. From the principle of explosion, it follows that $\forall\phi(\texttt{KB} \overset{c}{\models} \phi)$. Then $\mathcal{R}$ is not an ideal$_{ex}$ reasoner for $\overset{c}{\models}$ (def. 1.2.2, req. (ii))

[25] Then, for every tautology ϕ, there exists an i such that $\phi \in \pi(\texttt{KB}, i)$.

[26] Alternatively, $\mathcal{R}_c$ (i) may believe$_{ac}$ all the sentences that it believes$_{ac}$ ($\mathcal{R}_c$ itself is an ideal$_{ac}$ reasoner) and (ii) $\mathcal{R}_c$ believes$_{ac}$ all and only the sentences that it ought to believe (the tautologies).

and, consequently, $\mathcal{R}$ is not a strictly ideal$_{ex}$ reasoner for $\overset{c}{\models}$. But it may be thought that a maximally rational reasoner must be able to deal with inconsistencies.

Second, there may exist strictly ideal reasoners for $\overset{c}{\models}$ in epistemic situations with incomplete information, but monotony makes their patterns of inference, in some sense, unsatisfactory[27]. In those situations, it may be useful to make reasonable guesses and (possibly) retract these guesses given new information. For example, believing 'Most P are Q' and Pa, a reasoner may conclude that Qa and possibly retract that conclusion given new information (e.g. $\neg Qa$). But monotony prevents strictly ideal reasoners for $\overset{c}{\models}$ from having these patterns of inference[28]. In these cases, using strictly ideal reasoners for $\overset{c}{\models}$ as a model of maximum rationality disregards a whole class of successful patterns of inference. This fact is accepted in the literature:

> Many people consider this feature of FOL (monotony) as inadequate to capture a whole class of inferences typical of everyday (as opposed to mathematical or formal) reasoning and therefore question the descriptive adequacy of FOL when it comes to representing common sense inferences. In everyday life, we quite often reach conclusions tentatively, only to retract them in the light of further information (Antonelli, 2005, p. 6).

This second problem may be avoided using strictly ideal reasoners for a nonmonotonic logic as a model of maximum rationality[29]. A nonmonotonic logic $\overset{n}{\models}$ is a logic which does not exhibit (p0) monotony. For a nonmonotonic logic, there are situations in which $\texttt{KB} \overset{n}{\models} \phi$ and $\texttt{KB} \cup \texttt{INPUT} \overset{n}{\not\models} \phi$. Then a strictly ideal reasoner for a nonmonotonic logic may withdraw beliefs given new information.

In order to make the notion of a strictly ideal reasoner for $\overset{n}{\models}$ precise, we need to know the properties of $\overset{n}{\models}$. Many nonmonotonic formalisms have been proposed in the literature: circumscription (McCarthy, 1980), autoepistemic logic (Moore, 1985), default logic (Reiter, 1980), etc. These formalisms are implemented using different tools: circumscription

[27] Roughly, an epistemic situation has incomplete information when not all relevant conclusions may be deduced from the available information.

[28] Suppose that $\mathcal{R}$ is a strictly ideal$_{ac}$ reasoner for $\overset{c}{\models}$. Monotony states that if $\texttt{KB} \overset{c}{\models} \phi$, then $\texttt{KB}{\cup}\texttt{INPUT} \overset{c}{\models} \phi$. Then $\mathcal{R}$ is such that if $\phi \in \pi(\texttt{KB})$, then $\phi \in (\texttt{KB} \cup \texttt{INPUT})$ (def. 1.1.9, strictly ideal$_{ac}$), what means that $\mathcal{R}$ cannot withdraw beliefs$_{ac}$ given new information.

[29] The first problem may be avoided using a paraconsistent logic.

uses a second-order framework, autoepistemic logic uses a modal framework, default logic uses a classical framework augmented with default rules, etc. This fact make it difficult to access the general features of a strictly ideal reasoner for $\overset{\text{n}}{\models}$ using any of these formalisms.

In the literature, there exist four properties which are often regarded as essential for a well-behaved nonmonotonic logic $\overset{\text{n}}{\models}$ (Antonelli, 2005; Stanalker, 1994; Makinson, 1994):

Definition 1.3.1. (Features of a well-behaved nonmonotonic logic $\overset{\text{n}}{\models}$):

(p1) **Supraclassicality:** if $\texttt{KB} \overset{\text{c}}{\models} \phi$, then $\texttt{KB} \overset{\text{n}}{\models} \phi$;

(p2) **Reflexivity:** if $\phi \in \texttt{KB}$, then $\texttt{KB} \overset{\text{n}}{\models} \phi$;

(p3) **Cut:** if $\texttt{KB} \overset{\text{n}}{\models} \texttt{KB}'$ and $\texttt{KB} \cup \texttt{KB}' \overset{\text{n}}{\models} \phi$, then $\texttt{KB} \overset{\text{n}}{\models} \phi$;

(p4) **Cautious monotony (CM):** if $\texttt{KB} \overset{\text{n}}{\models} \texttt{KB}'$ and $\texttt{KB} \overset{\text{n}}{\models} \phi$, then $\texttt{KB} \cup \texttt{KB}' \overset{\text{n}}{\models} \phi$,

where $\texttt{KB} \overset{\text{n}}{\models} \texttt{KB}'$ means that if $\phi \in \texttt{KB}'$, then $\texttt{KB} \overset{\text{n}}{\models} \phi$.

Supraclassicality states that $\overset{\text{n}}{\models}$ extends $\overset{\text{c}}{\models}$: if ϕ follows from $\texttt{KB}$ in $\overset{\text{c}}{\models}$, then ϕ follows from $\texttt{KB}$ in $\overset{\text{n}}{\models}$. Reflexivity states that if $\texttt{KB}$ contains ϕ, then ϕ follows from $\texttt{KB}$ in $\overset{\text{n}}{\models}$. Cut is a form of transitivity: it states that adding logical consequences of $\texttt{KB}$ back to $\texttt{KB}$ does not lead to a increase in inferential power. CM is the converse of cut: it states that adding consequences of $\texttt{KB}$ back to $\texttt{KB}$ does not lead to a decrease in inferential power.

Properties (p1)$-$(p4) are identified as essential for $\overset{\text{n}}{\models}$ because these properties enable the existence of well-behaved semantic interpretations. For example, Kraus et al. (1990) propose a nonmonotonic logic **C** with properties (p1)$-$(p4) which is proved sound and complete over preferential models, ordered using a preference relation $\prec$. In the research program, properties (p1)$-$(p4) are also essential because they enable the existence of a well-behaved strictly ideal$_{ac}$ reasoner for $\overset{\text{n}}{\models}$, a reasoner $\mathcal{R}$ such that (p1) $\mathcal{R}$ is an ideal$_{ac}$ reasoner for $\overset{\text{c}}{\models}$, (p2) $\mathcal{R}$ accesses all of its beliefs$_{ex}$ (if $\mathcal{R}$ believes$_{ex}$ ϕ, then $\mathcal{R}$ believes$_{ac}$ ϕ), and (p3, p4) $\mathcal{R}$ has 'stable' beliefs$_{ac}$ (see table 1.1)[30].

Given that $\overset{\text{n}}{\models}$ is (p1) supraclassical, an ideal$_{ac}$ reasoner for $\overset{\text{n}}{\models}$, and, consequently, a strictly ideal$_{ac}$ reasoner for $\overset{\text{n}}{\models}$ are classical ideal$_{ac}$ reasoners. Consider a nonmonotonic

[30] I use 'stable' between quotes because I use this term later with a different meaning. Here, having 'stable' beliefs$_{ac}$ is having a set of beliefs$_{ac}$ which do not change over any amount of a priori reasoning.

Logic $\overset{\mathrm{x}}{\models}$	Strictly ideal$_{ac}$ for $\overset{\mathrm{x}}{\models}$
Supraclassicality	Classical ideality$_{ac}$
Reflexivity	$\mathtt{KB} \subseteq \pi(\mathtt{KB})$
Cut	$\pi(\pi(\mathtt{KB})) \subseteq \pi(\mathtt{KB})$
CM	$\pi(\mathtt{KB}) \subseteq \pi(\pi(\mathtt{KB}))$

Table 1.1: Relation between properties of a logic and properties of a strictly ideal$_{ac}$ reasoner for that logic. For a modal interpretations of these properties, see appendix A.

ideal$_{ac}$ reasoner $\mathcal{R} = \langle \mathtt{KB}, \pi \rangle$. Then (i) $\mathcal{R}$ believes$_{ac}$ all classical consequences of its epistemic situation: $\mathtt{KB} \overset{\mathrm{c}}{\models} \phi \rightarrow \mathtt{KB} \overset{\mathrm{n}}{\models} \phi$ (supraclassicality) and $\mathtt{KB} \overset{\mathrm{n}}{\models} \phi \rightarrow \phi \in \pi(\mathtt{KB})$ (def. 1.2.6, req. (i)), then $\mathtt{KB} \overset{\mathrm{c}}{\models} \phi \rightarrow \phi \in \pi(\mathtt{KB})$. Also (ii) $\mathcal{R}$ has nontrivial beliefs$_{ac}$: $\exists\phi(\pi(\mathtt{KB}) \overset{\mathrm{n}}{\not\models} \phi) \rightarrow \exists\phi(\pi(\mathtt{KB}) \overset{\mathrm{c}}{\not\models} \phi)$ (supraclassicality) and $\exists\phi(\pi(\mathtt{KB}) \overset{\mathrm{n}}{\not\models} \phi)$ (def. 1.2.6, req. (ii)), then $\exists\phi(\pi(\mathtt{KB}) \overset{\mathrm{c}}{\not\models} \phi)$. Given that $\overset{\mathrm{n}}{\models}$ is (p2) reflexive, an ideal$_{ac}$ reasoner for $\overset{\mathrm{n}}{\models}$, and, consequently, a strictly ideal$_{ac}$ reasoner for $\overset{\mathrm{n}}{\models}$ access all of their beliefs$_{ex}$ ($\mathtt{KB} \subseteq \pi(\mathtt{KB})$): $\phi \in \mathtt{KB} \rightarrow \mathtt{KB} \overset{\mathrm{n}}{\models} \phi$ (reflexivity) and $\mathtt{KB} \overset{\mathrm{n}}{\models} \phi \rightarrow \phi \in \pi(\mathtt{KB})$ (def. 1.2.6, req. (i)), then $\phi \in \mathtt{KB} \rightarrow \phi \in \pi(\mathtt{KB})$ and $\mathtt{KB} \subseteq \pi(\mathtt{KB})$.

Given that $\overset{\mathrm{n}}{\models}$ exhibits (p3) cut and (p2) reflexivity, the beliefs$_{ac}$ of a strictly ideal$_{ac}$ reasoner for $\overset{\mathrm{n}}{\models}$ do not increase over any amount of a priori reasoning ($\pi(\pi(\mathtt{KB})) \subseteq \pi(\mathtt{KB})$). Suppose that $\phi \in \pi(\pi(\mathtt{KB}))$. Then $\pi(\mathtt{KB}) \overset{\mathrm{n}}{\models} \phi$ (def. 1.1.9, strictly ideal$_{ac}$). Then $\mathtt{KB} \cup \pi(\mathtt{KB}) \overset{\mathrm{n}}{\models} \phi$ (reflexivity)[31]. For all $\psi \in \pi(\mathtt{KB})$, $\mathtt{KB} \overset{\mathrm{n}}{\models} \psi$ (strictly ideal). Then $\mathtt{KB} \overset{\mathrm{n}}{\models} \pi(\mathtt{KB})$. $\mathtt{KB} \overset{\mathrm{n}}{\models} \pi(\mathtt{KB})$ and $\mathtt{KB} \cup \pi(\mathtt{KB}) \overset{\mathrm{n}}{\models} \phi$, what entails $\mathtt{KB} \overset{\mathrm{n}}{\models} \phi$ (cut). $\mathtt{KB} \overset{\mathrm{n}}{\models} \phi$, then $\phi \in \pi(\mathtt{KB})$ (strictly ideal$_{ac}$). Therefore $\pi(\pi(\mathtt{KB})) \subseteq \pi(\mathtt{KB})$. Given that $\overset{\mathrm{n}}{\models}$ exhibits (p4) CM and (p2) reflexivity, the beliefs$_{ac}$ of a maximally rational reasoner for $\overset{\mathrm{n}}{\models}$ does not decrease over any amount of a priori reasoning ($\pi(\mathtt{KB}) \subseteq \pi(\pi(\mathtt{KB}))$). Suppose that $\phi \in \pi(\mathtt{KB})$. Then $\mathtt{KB} \overset{\mathrm{n}}{\models} \phi$ (strictly ideal$_{ac}$). For all $\psi \in \pi(\mathtt{KB})$, $\mathtt{KB} \overset{\mathrm{n}}{\models} \psi$ (strictly ideal). Then $\mathtt{KB} \overset{\mathrm{n}}{\models} \pi(\mathtt{KB})$. $\mathtt{KB} \overset{\mathrm{n}}{\models} \pi(\mathtt{KB})$ and $\mathtt{KB} \overset{\mathrm{n}}{\models} \phi$, what entails $\mathtt{KB} \cup \pi(\mathtt{KB}) \overset{\mathrm{n}}{\models} \phi$ (CM). $\mathtt{KB} \cup \pi(\mathtt{KB}) \overset{\mathrm{n}}{\models} \phi$, then $\pi(\mathtt{KB}) \overset{\mathrm{n}}{\models} \phi$ (reflexivity) and $\phi \in \pi(\pi(\mathtt{KB}))$ (strictly ideal$_{ac}$). Then $\pi(\mathtt{KB}) \subseteq \pi(\pi(\mathtt{KB}))$.

A nonmonotonic logic $\overset{\mathrm{n}}{\models}$ with (p1)$-$(p4) yields a well-behaved strictly ideal$_{ac}$ reasoner, but, for several situations of interest, that strictly ideal$_{ac}$ reasoner for $\overset{\mathrm{n}}{\models}$ cannot be finite.

[31] Suppose that $\phi \in \pi(\mathtt{KB})$. Then $\phi \in \mathtt{KB}$ or $\phi \in \pi(\mathtt{KB})$. Then $\phi \in \mathtt{KB} \cup \pi(\mathtt{KB})$. Suppose that $\phi \in \mathtt{KB} \cup \pi(\mathtt{KB})$. Then $\phi \in \mathtt{KB}$ or $\phi \in \pi(\mathtt{KB})$. Suppose that $\phi \in \mathtt{KB}$. Then $\mathtt{KB} \overset{\mathrm{n}}{\models} \phi$ (reflexivity). Then $\phi \in \pi(\mathtt{KB})$ (strictly ideal$_{ac}$). Therefore $\pi(\mathtt{KB}) = \mathtt{KB} \cup \pi(\mathtt{KB})$.

(1) Suppose that $\mathcal{R} = \langle \pi, \texttt{KB} \rangle$ is strictly ideal$_{ac}$ for $\overset{\mathrm{n}}{\models}$.

(2) $(\forall \texttt{KB}, \phi, \psi)$ If $\texttt{KB} \overset{\mathrm{n}}{\models} \psi$ and $\texttt{KB}, \phi \overset{\mathrm{n}}{\not\models} \psi$, then $\texttt{KB} \overset{\mathrm{n}}{\not\models} \phi$ (CM).

Suppose that $\texttt{KB} \overset{\mathrm{n}}{\models} \psi$ and $\texttt{KB}, \phi \overset{\mathrm{n}}{\not\models} \psi$. Suppose that $\texttt{KB} \overset{\mathrm{n}}{\models} \phi$. Then $\texttt{KB} \overset{\mathrm{n}}{\models} \psi$ and $\texttt{KB} \overset{\mathrm{n}}{\models} \phi$. But if $\texttt{KB} \overset{\mathrm{n}}{\models} \psi$ and $\texttt{KB} \overset{\mathrm{n}}{\models} \phi$, then $\texttt{KB}, \phi \overset{\mathrm{n}}{\models} \psi$ (CM). Then $\texttt{KB}, \phi \overset{\mathrm{n}}{\models} \psi$, what contradicts the first supposition. Then $\texttt{KB} \overset{\mathrm{n}}{\not\models} \phi$.

(3) $(\exists \texttt{KB}, \phi, \psi)$ $\texttt{KB} \overset{\mathrm{n}}{\models} \psi$ and $\texttt{KB}, \phi \overset{\mathrm{n}}{\not\models} \psi$ (nonmonotony).

(4) $(\exists^{\infty} \texttt{KB}, \phi, \psi)$ $\texttt{KB} \overset{\mathrm{n}}{\models} \psi$ and $\texttt{KB}, \phi \overset{\mathrm{n}}{\not\models} \psi$.

$(\exists \texttt{KB}, \phi, \psi)$ $\texttt{KB} \overset{\mathrm{n}}{\models} \psi$ and $\texttt{KB}, \phi \overset{\mathrm{n}}{\not\models} \psi$ (3). But if ($\texttt{KB} \overset{\mathrm{n}}{\models} \psi$ and $\texttt{KB}, \phi \overset{\mathrm{n}}{\not\models} \psi$), then ($\texttt{KB}, \top \overset{\mathrm{n}}{\models} \psi \wedge \top$ and $\texttt{KB}, \phi, \top \overset{\mathrm{n}}{\not\models} \psi \wedge \top$), where $\top$ is an arbitrary classical tautology[32]. There exist infinitely many classical tautologies.

(5) $(\forall \texttt{KB}, \phi, \psi)$ If $\psi \in \pi(\texttt{KB})$ and $\psi \notin \pi(\texttt{KB} \cup \{\phi\})$, then $\texttt{KB} \overset{\mathrm{n}}{\not\models} \phi$ (1 and 2).

(6) $(\exists^{\infty} \texttt{KB}, \phi, \psi)$ $\psi \in \pi(\texttt{KB})$ and $\psi \notin \pi(\texttt{KB} \cup \{\phi\})$ (1 and 4).

$(\therefore)$ For infinitely many epistemic situations ($\texttt{KB}$s), $\mathcal{R}$ cannot be finite (6, see below).

If π is finite, then (for infinitely many $\texttt{KB}$, ϕ, and ψ) π is be able to compute in finitely many steps $\psi \in \pi(\texttt{KB})$ and $\psi \notin \pi(\texttt{KB} \cup \{\phi\})$ (6). This entails that (for the $\texttt{KB}$ and ϕ) π is be able to compute in finitely many steps $\texttt{KB} \overset{\mathrm{n}}{\not\models} \phi$ (5). However, there does not exist restrictions on which are the $\texttt{KB}$ and ϕ in question. At the limit, the $\texttt{KB}$ and ϕ could be all of those such that $\texttt{KB} \overset{\mathrm{n}}{\not\models} \phi$. In this case, π is able to compute in finitely many steps $\texttt{KB} \overset{\mathrm{n}}{\not\models} \phi$ (in general). But $\texttt{KB} \overset{\mathrm{n}}{\not\models} \phi$ entails $\texttt{KB} \overset{\mathrm{c}}{\not\models} \phi$ (supraclassicality). Then function π is able to compute in finitely many steps $\texttt{KB} \overset{\mathrm{c}}{\not\models} \phi$. But this is not possible because classical logic is not decidable (see Church, 1936a). An option would be to list the problem cases (the $\texttt{KB}$ and ϕ) in the definition of π, but these cases are infinitely many. Therefore, for infinitely many epistemic situations ($\texttt{KB}$), $\mathcal{R}$ cannot be finite.

[32] Suppose that $\texttt{KB} \overset{\mathrm{n}}{\models} \psi$. $\texttt{KB} \overset{\mathrm{n}}{\models} \top$ (tautology, supraclassicality). But if $\texttt{KB} \overset{\mathrm{n}}{\models} \psi$ and $\texttt{KB} \overset{\mathrm{n}}{\models} \top$, then $\texttt{KB}, \top \overset{\mathrm{n}}{\models} \psi$ (CM). Then $\texttt{KB}, \top \overset{\mathrm{n}}{\models} \psi$. But $\texttt{KB}, \top \overset{\mathrm{n}}{\models} \top$ (reflexivity). Then $\texttt{KB}, \top \overset{\mathrm{n}}{\models} \psi \wedge \top$ (supraclassicality). Suppose that $\texttt{KB}, \phi \overset{\mathrm{n}}{\not\models} \psi$. Suppose that $\texttt{KB}, \phi, \top \overset{\mathrm{n}}{\models} \psi \wedge \top$. Then $\texttt{KB}, \phi, \top \overset{\mathrm{n}}{\models} \psi$ (supraclassicality). Also, $\texttt{KB}, \phi \overset{\mathrm{n}}{\models} \top$ (tautology, supraclassicality). But if $\texttt{KB}, \phi \overset{\mathrm{n}}{\models} \top$ and $\texttt{KB}, \phi, \top \overset{\mathrm{n}}{\models} \psi$, then $\texttt{KB}, \phi \overset{\mathrm{n}}{\models} \psi$ (cut). Then $\texttt{KB}, \phi \overset{\mathrm{n}}{\models} \psi$, what is a contradiction.

This argument does not show that the class of finite reasoners which are strictly ideal$_{ac}$ for $\overset{n}{\models}$ is a priori empty. In fact, at least for the epistemic situations in which all consequences are classical, there exist finite reasoners that are strictly ideal$_{ac}$ for $\overset{n}{\models}$ because, in those situations, the finite reasoners which are strictly ideal$_{ac}$ for $\overset{c}{\models}$ are also strictly ideal$_{ac}$ for $\overset{n}{\models}$ (the same hold for the other two classes). Then the notion of an ideal$_{ac}$ reasoner for $\overset{n}{\models}$ fulfills goal (g1). But this notion has problems regarding goal (g2).

The aim of investigating strictly ideal$_{ac}$ reasoners for $\overset{n}{\models}$ was to investigate maximum rationality in situations of incomplete information. This investigation would include the construction of the optimal sets of nonmonotonic rules (sets of nonmonotonic rules which generate successful patterns of inference in situations of incomplete information). There is no indication that there exist *the* optimal set of nonmonotonic rules because which patterns of inference are successful in a situation with incomplete information depends on contingent facts about the environment. Then an investigation about optimal sets of nonmonotonic rules can hardly be carried out a priori using only analytic methods[33]. In this case, the investigation about optimal sets of nonmonotonic rules would need to be carried out numerically by observing the behavior of reasoners in a large set of situations. The amount of calculation involved in this kind of investigation might be too large and the investigator would benefit from the use of computer simulations. But the above argument may be interpreted as stating that there does not exist, for every situation, a computable function π which is strictly ideal$_{ac}$ for $\overset{n}{\models}$[34]. This is a shortcoming regarding goal (g2)[35].

There exist two strategies for dealing with this problem: (s1) restrict the investigation to situations in which there exist a computable function π which is strictly ideal$_{ac}$ for $\overset{n}{\models}$[36]; (s2) use a notion of beliefs which, for all situations, enables the existence of a computable function π which is strictly ideal for $\overset{n}{\models}$. In the following, I investigate strategy (s2).

[33] An evidence for this claim is that logicicians usually study the general (formal) properties of nonmonotonic formalisms, leaving aside the investigation about the optimal set of nonmonotonic rules.

[34] A function is said to be computable if some Turing machine will take in arguments of the function and, carrying out some finite number of basic operation, produce the corresponding value − and, moreover, will do thus no matter which argument of the function is presented.

[35] Appendix A contains a different line of reasoning for the conclusion that beliefs$_{ac}$ is not an adequate notion of beliefs for dealing with strictly ideal reasoners for a nonmonotonic logic with (p1)−(p4).

[36] In fact, there exist quite expressible fragments of classical logic that are decidable and function π may be generally computable for a nonmonotonic extension of these fragments (see Böerger et al., 1997).

1.4 Stable beliefs

In section 1.1, I have modeled a reasoner as a pair $\langle \mathtt{KB}, \pi \rangle$, where $\mathtt{KB}$ is a set of sentences in a formal language $\mathcal{L}$ which models the beliefs$_{ex}$ of the reasoner and $\pi : 2^{\mathcal{L}} \times \mathbb{Z}^{+} \rightarrow 2^{\mathcal{L}}$ is a function for updating $\mathtt{KB}$ which models the pattern of inference of the reasoner. A fact about the pattern of inference of a reasoner is that the reasoner may perform different inferences from the same premises. In the model, this fact is expressed using the numeric parameter of function π. In this context, $\pi(\mathtt{KB}, 1)$ models one inference from $\mathtt{KB}$, $\pi(\mathtt{KB}, 2)$ models another inference from $\mathtt{KB}$, etc. Then function π determines a reasoning sequence $\mathtt{KB}_0, \mathtt{KB}_1, \ldots, \mathtt{KB}_i, \ldots$, where $\mathtt{KB}_0$ is the initial $\mathtt{KB}$ of the reasoner and $\mathtt{KB}_{i+1} = \pi(\mathtt{KB}_i, i{+}1)$. Supposing that the numeric parameter models an order of preference, the reasoning sequence of a reasoner models how the reasoner would reason from the available information if it had enough cognitive resources (e.g. time for reasoning).

Informally, a reasoner has the stable belief that ϕ iff if the reasoner were to reason undefinitely from the available information, there would be a point of the reasoning such that the reasoner would believe$_{ex}$ ϕ at every point after that. Consider the notion of stable beliefs in the model:

Definition 1.4.1 (Stable beliefs (belief$_{\omega}$)). A reasoner $\mathcal{R} = \langle \mathtt{KB}, \pi \rangle$ believes$_{\omega}$ ϕ iff $\phi \in \mathtt{KB}_{\omega}$.

$\mathtt{KB}_{\omega}$, the set of beliefs$_{\omega}$ of a reasoner with reasoning sequence $\mathtt{KB}_0, \mathtt{KB}_1, ..., \mathtt{KB}_i, ...$, is composed of the sentences ϕ for which there is an i such that, for all $j \geq i$, $\phi \in \mathtt{KB}_j$. Formally, $\mathtt{KB}_{\omega} = \bigcup_i \bigcap_{j \geq i} \mathtt{KB}_j$[37].

This is the definition of an ideal reasoner for $\overset{x}{\models}$ in terms of beliefs$_{\omega}$:

Definition 1.4.2 (Ideal$_{\omega}$ reasoner for $\overset{x}{\models}$). A reasoner $\mathcal{R} = \langle \mathtt{KB}, \pi \rangle$ is ideal$_{\omega}$ for $\overset{x}{\models}$ iff

(i) $\mathcal{R}$ believes$_{\omega}$ all logical consequences of its epistemic situation ($\mathtt{KB} \overset{x}{\models} \phi \rightarrow \phi \in \mathtt{KB}_{\omega}$);

(ii) $\mathcal{R}$ has a nontrivial set of beliefs$_{\omega}$ ($\exists \phi (\mathtt{KB}_{\omega} \overset{x}{\not\models} \phi)$).

[37] The notion of beliefs$_{\omega}$ is related to the notions of defeasible enumeration (Pollock, 1995, p. 143), identification in the limit (Kelly, 1990), limiting recursion (Gold, 1965), and trial and error predicate (Putnam, 1965), and defeasible consequence (Antonelli, 2005, p. 87).

In the research program, using ideal$_\omega$ reasoners allows for the existence of finite reasoners which are maximally and minimally rational, and irrational for both $\overset{\text{c}}{\models}$ and $\overset{\text{n}}{\models}$. Consider the reasoner $\mathcal{R}_c$ in section 1.2. I addition to being strictly ideal$_{ac}$ for $\overset{\text{c}}{\models}$, $\mathcal{R}_c$ is also strictly ideal$_\omega$ for $\overset{\text{c}}{\models}$. This is the case because the function π of $\mathcal{R}_c$ was constructed in such a way that, for every i, $\pi(\texttt{KB}, i) \subseteq \pi(\texttt{KB}, i+1)$[38]. In this case, $\phi \in \texttt{KB}_\omega$ iff $\phi \in \pi(\texttt{KB})$. Then $\mathcal{R}_c$ is strictly ideal$_\omega$ for $\overset{\text{c}}{\models}$ for the same reasons that it is strictly ideal$_{ac}$ for $\overset{\text{c}}{\models}$.

From $\mathcal{R}_c$, it is possible to construct a reasoner $\mathcal{R}_n$ which is strictly ideal$_\omega$ for $\overset{\text{n}}{\models}$. Consider an axiomatization of $\overset{\text{n}}{\models}$ (e.g. Kraus et al., 1990). For every nonmonotonic rule $\frac{\gamma:\phi}{\psi}$ in the axiomatization consider a nonmonotonic operation $\frac{\gamma \in \texttt{KB}, \phi \notin \texttt{KB}}{\text{add } \psi \text{ to } \texttt{KB}}$[39]. Then build function π in such a way that, every time it adds ψ to KB using a nonmonotonic operation, it checks whether ϕ is in KB at every subsequent loop. If it the check results positive, it deletes ψ and its consequences from KB[40]. In this case, there exist two possibilities: either (i) ϕ is added to KB at some point and ψ and its consequences are deleted from KB or (ii) ϕ is never added to KB. In any case, $\psi \in \texttt{KB}_\omega$ iff $\texttt{KB} \overset{\text{n}}{\models} \psi$ and $\mathcal{R}_n$ strictly ideal$_\omega$ for $\overset{\text{n}}{\models}$[41].

Both $\mathcal{R}_c$ and $\mathcal{R}_n$ are finite. The limiting knowledge base $\texttt{KB}_\omega$ of a strictly ideal$_\omega$ reasoner is usually infinite, but $\texttt{KB}_\omega$ is not part of the reasoning sequence of the reasoner. In fact, a reasoner may have finitely many beliefs$_{ex}$ at every point of the reasoning sequence and, nevertheless, have an infinite $\texttt{KB}_\omega$. For example, if $\texttt{KB}_0 = \varnothing$ and each $\pi(\texttt{KB}_i, i+1)$ adds a sentence to $\texttt{KB}_i$, then every $\texttt{KB}_i$ in the reasoning sequence is finite but $\texttt{KB}_\omega$ is infinite. Then a strictly ideal$_\omega$ reasoner may have finite $\texttt{KB}_i$ at every point of the reasoning sequence. This is the case for both $\mathcal{R}_c$ and $\mathcal{R}_n$ because $\texttt{KB}_0 = \varnothing$, each execution of function π for a given i has finitely many loops, and each loop adds at most finitely many sentences to KB. The definition of function π is finite. Finally, the execution of function π is finite because, for each i, there exists a finite sequence of operations which computes $\pi(\texttt{KB}_i, i) = \texttt{KB}_{i+1}$. This is possible because function π does not need to check if $\phi \in \pi(\texttt{KB})$ before adding ψ to KB, but only to check if $\phi \in \texttt{KB}$ after adding ψ to KB (and the $\texttt{KB}_i$ are always finite).

[38] For every i, function π of $\mathcal{R}_c$ executes the operations of every $j < i$ in executing the operations of i and none of those operations delete sentences from KB.

[39] I have in mind a default rule: if γ is believed and ϕ cannot be inferred, then ψ may be inferred.

[40] The consequences of ψ are all sentences added to KB using an operation with ψ as premise.

[41] Properties in table 1.1 also hold for the $\texttt{KB}_\omega$ of a strictly ideal$_\omega$ reasoner for $\overset{\text{n}}{\models}$.

From $\mathcal{R}_c$ and $\mathcal{R}_n$, we can construct minimally rational$_\omega$ (not considering an operation) and irrational$_\omega$ finite reasoners (adding $\bot$ to $\mathtt{KB}$). Then the use of ideal$_\omega$ reasoners fulfills goal (g1) and has advantages regarding (g2) in relation to the notion of ideal$_{ac}$ reasoners because the argument in section 1.3 does not hold for strictly ideal$_\omega$ reasoners[42].

In this context, is it the case that the use of ideal$_\omega$ reasoners in the research program must always (i.e. for all $\overset{\mathrm{x}}{\models}$) be preferred over the use of ideal$_{ac}$ reasoners? Not necessarily. The issue is that the beliefs$_{ac}$ and the beliefs$_\omega$ of a given reasoner do not always coincide. This difference results from the fact that beliefs$_\omega$ considers the execution of function π in a given order whereas beliefs$_{ac}$ is independent of any order. Then it is easy to construct cases in which a reasoner believes$_{ac}$ ϕ, but do not believe$_\omega$ ϕ. For example, suppose that $\mathcal{R} = \langle \mathtt{KB}, \pi \rangle$ is such that $\phi \in \mathtt{KB}$ and π is such that $\pi(\mathtt{KB}, 1)$ deletes ϕ and adds ψ and $\pi(\mathtt{KB}, i)$ does nothing for every $i > 1$. In this case, $\mathcal{R}$ believes$_{ac}$ ϕ because $\phi \in \pi(\mathtt{KB}, 2)$ but $\mathcal{R}$ does not believe$_\omega$ ϕ because $\phi \notin \mathtt{KB}_i$ in the reasoning sequence for every $i > 0$[43].

As a consequence, using ideal$_{ac}$ reasoners or ideal$_\omega$ reasoners in the research program may yield different prescriptions regarding rationality and the material adequacy of these prescriptions must be considered. In this context, I am particularly interested in the fact that there exist more irrational$_{ac}$ than irrational$_\omega$ reasoners for a supraclassical $\overset{\mathrm{x}}{\models}$. For example, suppose in the situation in last paragraph that $\psi = \neg\phi$, then $\mathcal{R}$ is irrational$_{ac}$ for a supraclassical $\overset{\mathrm{x}}{\models}$ because $\phi \in \pi(\mathtt{KB})$ and $\neg\phi \in \pi(\mathtt{KB})$, but $\mathcal{R}$ is not irrational$_\omega$ for $\overset{\mathrm{n}}{\models}$ because $\neg\phi \in \mathtt{KB}_\omega$, but $\phi \notin \mathtt{KB}_\omega$. In cases like this, the result in term of beliefs$_\omega$ seems to be more reasonable. Suppose that ϕ is a contradiction. In this case, the reasoning sequence of $\mathcal{R}$ is to delete a contradiction, add its negation (a tautology), and them leave the tautology alone. This reasoning sequence does not seem to be irrational because $\mathcal{R}$ does not have a inconsistent $\mathtt{KB}$ at any step $i > 0$ of the reasoning sequence. In this context, it seems to be the case that the notion of an ideal$_\omega$ reasoner must indeed be preferred in the research program (at least for supraclassical $\overset{\mathrm{x}}{\models}$).

[42] Substituting $\pi(\mathtt{KB})$ for $\mathtt{KB}_\omega$, the argument goes through until step (6), but the conclusion does not follows. The point is that this function π does not need to compute a $\mathtt{KB}_\omega$ with the properties in (6): function π only needs to generate a reasoning sequence such that $\mathtt{KB}_\omega$ with those properties.

[43] This issue is discussed in Pollock (1995, p. 132−134).

1.5 Conclusions

In this chapter, I have investigated which notion of an ideal reasoner may be used in a research program about stating the maximum and minimum bounds of rationality for finite reasoners as to fulfill goals (g1) and (g2). In section 1.1, I have proposed a model of a reasoner which is used in most definitions in the chapter. I have investigated some notions of an ideal reasoner which are common in the literature, argued that these notions are not adequate for modeling maximum rationality for finite reasoners, and then proposed the related notion of a strictly ideal reasoner. In section 1.2, I have argued that the notion of a strictly ideal reasoner is relative to a logic and to a notion of beliefs. Then I have investigated a strictly ideal reasoner for classical logic defined in terms of explicit, implicit, and accessible beliefs. I have rejected the definitions in terms of explicit and implicit beliefs on the basis of goal (g1). In section 1.3, I have investigate a strictly ideal reasoner for a nonmonotonic logic with some essential properties defined in terms of accessible beliefs and rejected this definition on the basis of goal (g2). In section 1.4, I have introduced the notion of stable beliefs and used this notion in the definition of strictly ideal reasoners for both classical and nonmonotonic logic and argued that these definitions fulfill goals (g1) and (g2) and should be used in the research program.

In the research program, the choice of $\overset{\text{x}}{\models}$ provides the normative content for the notions of maximum and minimum rationality and irrationality. For example, the discussion about whether it is always irrational to believe a contradiction is equivalent to the question about whether the relevant $\overset{\text{x}}{\models}$ is paraconsistent. In this chapter, I was mainly interested in the formal features of a notion of an ideal reasoner which enable the use of the most common $\overset{\text{x}}{\models}$. But it is a major goal of the research program to select/construct the logic $\overset{\text{x}}{\models}$ which yields the most sensible description maximum and minimum rationality and irrationality for finite reasoners. There exist several issues to be dealt with in the fulfillment of this goal. For example, considering how much a strictly ideal reasoner for $\overset{\text{x}}{\models}$ maximizes the truth in its set of beliefs introduce whole new set of considerations for the choice of the $\overset{\text{x}}{\models}$. Also, if $\overset{\text{x}}{\models}$ has an operator for beliefs, the theory needs to deal with epistemic paradoxes and blindspot sentences. I will deal with some of these issues in the next chapters.

Chapter 2

Ideal reasoners don't believe in zombies

> Alice laughed. "There's no use trying", she said. "One can't believe impossible things". "I daresay you haven't had much practice", said the Queen. "When I was your age, I always did it for half-an-hour a day. Sometimes I've believed as many as six impossible things before breakfast".
> Lewis Carroll, *Through the Looking-Glass*.

Modal epistemology is the sub-area of epistemology which investigates how reasoners come to have nontrivial modal knowledge[1]. In the literature, the conceivability principle ('if ϕ is conceivable, then ϕ is possible') is often regarded as basing nontrivial modal knowledge that ϕ is possible (see Gendler and Hawthorne, 2002). In philosophy of mind, the conceivability principle is often employed in supporting the conclusion that zombies are possible and, therefore, that physicalism is false. The *zombie argument* runs as follows[2]:

(p1) $p \wedge \neg q$ is conceivable.
(p2) If ϕ is conceivable, then ϕ is possible.
(p3) If $p \wedge \neg q$ is possible, then physicalism is false.
$\therefore$ Physicalism is false.

In the zombie argument, p is the conjunction of the fundamental physical truths and laws and q is an arbitrary phenomenal truth[3]. If q expresses the truth that someone is

[1] Usually, 'nontrivial modal knowledge' refers either to knowledge that $\Box\phi$ or to knowledge that $\Diamond\phi$ which is not based on knowledge that ϕ (see Evnine, 2008, p. 665). For simplicity, I will focus on the case of knowledge that $\Diamond\phi$ which is not based on knowledge that ϕ because ϕ is not even believed.

[2] A zombie is a system which is physically identical to a conscious being but which lacks consciousness.

[3] In general, $\mathcal{L}$-truths are truths in $\mathcal{L}$. For example, physical truths are the true sentences in a physical language (a language in which the physics true of our world can be expressed) (Haugeland, 1982, p. 96).

conscious, then $\Diamond(p \land \neg q)$ asserts the possibility of a zombie world, i.e. a world physically identical to the actual world, but in which no one is conscious (Chalmers, 2002, 2010).

There exists an extensive literature about the zombie argument (see Kirk, 2012). Premise (p3) is relatively undisputed: it is often accepted that physicalism entails that the mental supervenes on the physical[4]. Premises (p1) and (p2) are disputed. Chalmers defends premises (p1) and (p2) by distinguishing different senses of 'conceivable' and 'possible' and by pointing senses in which $p \land \neg q$ would be conceivable and conceivability would entail possibility. Among other things, Chalmers defends the *negative zombie argument*:

(p1)* $p \land \neg q$ is ideally negatively conceivable.
(p2)* If ϕ is ideally negatively conceivable, then ϕ is possible.
(p3) If $p \land \neg q$ is possible, then physicalism is false.
$\therefore$ Physicalism is false.

The notion of ideal negative conceivability is sometimes explained as follows:

Definition 2.0.1 (Ideal negative conceivability 1). A sentence ϕ is ideally negatively conceivable iff the hypothesis expressed by ϕ cannot be ruled out a priori on ideal rational reflection (Chalmers, 2002, p. 143).

In this context, Chalmers defends the *negative zombie argument*:

(p1)* $p \land \neg q$ is ideally negatively conceivable.
(p2)* If ϕ is ideally negatively conceivable, then ϕ is possible.
(p3) If $p \land \neg q$ is possible, then physicalism is false.
$\therefore$ Physicalism is false.

In this chapter, I argue that premises (p1)* and (p2)* of the negative zombie argument are not both true for every choice of q. For example, (p1)* and (p2)* are not both true for $q =$'someone is conscious'. In section 2.1, I argue that (p1)* and (p2)* are both true iff there exists a strictly ideal reasoner for a logic with the relevant properties and that reasoner negatively conceiving ϕ is a (conclusive) reason for $\Diamond\phi$. In section 2.2, I investigate whether these conditions hold and conclude that, even if they do, the conclusive

[4] A set of sentences $\mathcal{L}$ (weakly) supervenes on a set of sentences $\mathcal{L}$' iff two possible worlds cannot be discernible with $\mathcal{L}$ without being discernible with $\mathcal{L}$', where two worlds are discernible with $\mathcal{L}$ when there exists a sentence in $\mathcal{L}$ which is true at one world and is not true at the other (Haugeland, 1982, p. 96).

reasoner would not be finite[5]. I argue that this fact undermines the use of the ideal negative conceivability principle in the negative zombie argument. In section 2.3, I argue that there exists a finite strictly ideal reasoner for the relevant logic, but that this reasoner negatively conceiving ϕ after any amount of reasoning would only be a defeasible reason for $\Diamond\phi$. In section 2.4, I argue that the defeasible reasoner does not negatively conceive $p \wedge \neg q$ for every choice of q. For example, this reasoner does not negatively conceive $p \wedge \neg q$ for $q =$'someone is conscious'. In the conclusions, I discuss the consequences of these results for the negative zombie argument and for the zombie arguments in general. The bottom line is that the negative zombie argument (and maybe the zombie arguments in general) is neither an a priori nor a conclusive argument against physicalism.

2.1 Ideal negative conceivability

The major difficulty in evaluating the negative zombie argument is that the notion of ideal negative conceivability used in (p1)* and (p2)* is not entirely clear. More specifically, the meaning of the phrase 'a priori ideal rational reflection' in that notion is not entirely clear. This problem could be avoided by using the notion of a strictly ideal reasoner proposed in chapter 1 (def. 1.1.2) as a model of ideal rational reflection. This reading is implicit in Chalmers (2010), but it is explicit in Menzies (1998):

> Under what circumstances do our corrective practices discount acts of conceiving as not being veridical indicators of possibility? The answer is simple: When they suffer from one kind of cognitive limitation or other. Let us call a subject who does not suffer any of those limitations an ideal conceiver. These reflections suggest a biconditional of the following kind for the concept of possibility: it is possible that ϕ iff an ideal conceiver could conceive that ϕ (Menzies, 1998, p. 268-269).

The notion of an ideal conceiver used in Menzies (1998) and Chalmers (2010, p. 143) is a cognitive notion of an ideal reasoner without cognitive limitations (infinite reasoner). However, a cognitive notion of an ideal reasoner cannot be used as a model of ideal rational reflection unless the notion specifies how that reasoner reasons. In the literature,

[5] Roughly, a finite reasoner is a reasoner with cognitive limitations such as finite memory and being able to execute only a finite number of inferential steps in a finite time interval (see chapter 1, def. 1.1.2).

there exists an epistemic notion of an ideal reasoner as a reasoner which (i) believes all the logical consequences of its epistemic situation and (ii) has a nontrivial set of beliefs, where the epistemic situation of a reasoner is the information available for reasoning[6]. The epistemic notion of an ideal reasoner is also not adequate for modeling ideal rational reflection because an ideal reasoner with features (i) and (ii) may have random beliefs[7].

In chapter 1, I have proposed the notion of a strictly ideal reasoner as a model of maximum rationality. This notion does not present the problem of random beliefs and may be used as a model of ideal rational reflection[8]. This is the notion of a strictly ideal reasoner presented in chapter 1:

Definition 2.1.1 (Strictly ideal reasoner). A reasoner $\mathcal{R}$ is strictly ideal iff:

(i) $\mathcal{R}$ believes all and only the logical consequences of its epistemic situation;

(ii) $\mathcal{R}$ has a nontrivial set of beliefs.

The notion of a (strictly) ideal reasoner is relative to a logic because the notions of logical consequence and triviality are relative to a logic[9]. Then I will always talk about a strictly ideal reasoner for $\overset{x}{\models}$, where $\overset{x}{\models}$ is a consequence relation.

In this context, I argue that (p1)* and (p2)* are both true only if the following is true:

(Z) There exists an a priori strictly ideal reasoner for a logic $\overset{x}{\models}$ with the relevant properties which believes $\Diamond(p \wedge \neg q)$ on the basis of not believing $p \rightarrow q$,

, where an a priori reasoner is a reasoner which has the empty set as its epistemic situation.

The sentence $p \wedge \neg q$ is ideally negatively conceivable iff $p \wedge \neg q$ cannot be ruled out a priori on ideal rational reflection (def. 2.0.1). I am supposing that the notion of a strictly

[6] The notion of an ideal reasoner is defined along these lines in Binkley (1968), Halpern and Moses (1985), Grim (1988), Duc (1995), Giunchiglia and Giunchiglia (2001), and Stalnaker (2006).

[7] As long as those beliefs do not contradict requirement (ii).

[8] The notion of ideal rational reflection is related to the notion of maximum rationality as following: both notions are related to reasoning which fulfills the epistemic norms (rational) and which is ideal in the sense of withdrawing all the logical consequences from the available information. In section 1.1, I argue that the notions of a strictly ideal and a maximally rational reasoner coincide in our framework.

[9] The notion of a strictly ideal reasoner is also relative to a notion of beliefs (explicit, implicit, etc), but this distinction is not relevant for now. I discuss this issue in section 2.3.

ideal reasoner for a relevant logic $\overset{x}{\models}$ is an adequate model of ideal rational reflection. Then $p \wedge \neg q$ is ideally negatively conceivable iff $p \wedge \neg q$ cannot be ruled out by an a priori strictly ideal reasoner for $\overset{x}{\models}$. A sentence ϕ is ruled out a priori iff $\neg\phi$ is a priori. Then (p1)* $p \wedge \neg q$ is ideally negatively conceivable iff an a priori strictly ideal reasoner for $\overset{x}{\models}$ does not believe $p \rightarrow q$, which is equivalent to $\neg(p \wedge \neg q)$. Finally, (p2)* ideal negative conceivability entails possibility for the case of $p \wedge \neg q$ only if a strictly ideal reasoner for $\overset{x}{\models}$ believes $\Diamond(p \wedge \neg q)$ on the basis of not believing $p \rightarrow q$[10].

The claim that (p1)* and (p2)* are both true only if (Z) is true depends on the properties of $\overset{x}{\models}$. As I see it, Chalmers is correct when he claims that the adequate $\overset{x}{\models}$ for evaluating the negative zombie argument is the logic of a priori (modal) knowledge[11]. According to Chalmers, ideal negative conceivability is a general principle for generating nontrivial modal knowledge in that logic. In this context, an adequate $\overset{x}{\models}$ for evaluating (Z) must have the following properties:

(i) $\Diamond(p \wedge \neg q)$ is expressible in the language of $\overset{x}{\models}$;

(ii) For some ϕ, $\overset{x}{\models} \Diamond\phi$ and $\overset{x}{\not\models} \phi$;

(iii) For all ϕ, if $\overset{x}{\not\models} \neg\phi$, then $\overset{x}{\models} \Diamond\phi$;

(iv) $\overset{x}{\models}$ tracks the metaphysical modalities correctly;

(v) For all truths ϕ that supervene on the physical (i.e. superphysical), $\overset{x}{\models} p \rightarrow \phi$.

Requirement (i) assures that a strictly ideal reasoner for $\overset{x}{\models}$ has the required concepts for its use in the negative zombie argument (the concepts of possibility and necessity, the concepts used in p, the concepts used in q, etc). If a strictly ideal reasoner for $\overset{x}{\models}$ is supposed to believe $\Diamond(p \wedge \neg q)$, then it must be able to believe $\Diamond(p \wedge \neg q)$. A strictly ideal reasoner for $\overset{x}{\models}$ is able to believe $\Diamond(p \wedge \neg q)$ iff $\Diamond(p \wedge \neg q)$ expressible in the language of $\overset{x}{\models}$ (def. 2.1.1, req. i). Then $\Diamond(p \wedge \neg q)$ must be expressible in the language of $\overset{x}{\models}$.

[10] In the following, 'strictly ideal reasoner' means a priori strictly ideal reasoner.

[11] Chalmers argues that the zombie argument denies an epistemic entailment from the physical to the phenomenal truths, where 'epistemic entailment' is explained as following: "On this notion [of epistemic entailment], ϕ implies ψ when the conditional 'If ϕ then ψ' is a priori – when a subject can know that if ϕ is the case, then ψ is the case with justification independent of experience" (Chalmers, 2010, p. 109).

Requirement (ii) assures that a strictly ideal reasoner for $\overset{x}{\models}$ is able to have nontrivial modal knowledge. A reasoner has nontrivial modal knowledge that $\Diamond\phi$ only if it believes $\Diamond\phi$ and does not believe ϕ (note 1, knowledge implies belief). A strictly ideal reasoner for $\overset{x}{\models}$ is able to, for some ϕ, believe $\Diamond\phi$ while not believing ϕ iff, for some ϕ, $\overset{x}{\models} \Diamond\phi$ and $\overset{x}{\not\models} \phi$ (def. 2.1.1, req. i). Requirement (iii) assures that a strictly ideal reasoner for $\overset{x}{\models}$ employs negative conceivability as a general principle in generating modal beliefs. A reasoner employs negative conceivability for generating the belief that $\Diamond\phi$ iff it believes $\Diamond\phi$ on the basis of not believing $\neg\phi$. A strictly ideal reasoner for $\overset{x}{\models}$ is able to employ negative conceivability as a general principle for generating modal beliefs iff, for all ϕ, if $\overset{x}{\not\models} \neg\phi$, then $\overset{x}{\models} \Diamond\phi$ (def. 2.1.1, req. i). Together, requirements (ii) and (iii) assure that a strictly ideal reasoner for $\overset{x}{\models}$ is able to employ negative conceivability as a general principle for generating nontrivial modal beliefs (and, possibly, nontrivial modal knowledge).

Requirement (iv) assures that a strictly ideal reasoner for $\overset{x}{\models}$ knows how to reason about the metaphysical modalities. A strictly ideal reasoner for $\overset{x}{\models}$ knows how to reason about the metaphysical modalities iff $\overset{x}{\models}$ tracks the metaphysical modalities correctly (def. 2.1.1, req. i). In the literature, it is usually accepted that the metaphysical modalities have the following two minimum features (see Salmon, 1989). First, all logical necessities are metaphysical necessities. Then $\overset{x}{\models}$ must be such that, for all ϕ, if $\overset{x}{\models} \phi$, then $\overset{x}{\models} \Box\phi$ (necessitation). Second, what is actual is possible (or, equivalently, what is necessary is actual). Then $\overset{x}{\models}$ must be such that, for all ϕ, $\overset{x}{\models} \Box\phi \rightarrow \phi$ (reflexivity).

Requirement (v) assures that a strictly ideal reasoner for $\overset{x}{\models}$ has the inferential capacities required for its use in the negative zombie argument. The negative zombie argument is such that, if $p \wedge \neg q$ is ideally negatively conceivable, then consciousness is not superphysical. In order to this argument to be sound, it must be the case that, for all superphysical truths ϕ, $p \wedge \neg\phi$ cannot be ideally negatively conceivable (contraposition). Then, for all superphysical truths ϕ, a strictly ideal reasoner must be able to rule out $p \wedge \neg\phi$ a priori. Then, for all superphysical truths ϕ, a strictly ideal reasoner for $\overset{x}{\models}$ must believe $p \rightarrow \phi$. A strictly ideal reasoner for $\overset{x}{\models}$ believes $p \rightarrow \phi$ for all superphysical truths ϕ iff, for all superphysical truths ϕ, $\overset{x}{\models} p \rightarrow \phi$ (def. 2.1.1, req. i). Then $\overset{x}{\models}$ must be such that, for all superphysical truths ϕ, $\overset{\text{x}}{\models} p \to \phi$.

2.2 Conclusive ideal conceiver

The first step in evaluating (Z) is to investigate the existence of a strictly ideal reasoner for a $\overset{\text{x}}{\models}$ with requirements (i)$-$(v). In this case, the easiest path is to look for an existing logic with requirements (i)$-$(v). If there exists a sound and complete axiomatization for that logic, then the construction of the strictly ideal reasoner would be facilitated[12].

Requirement (i) states that $\Diamond(p \wedge \neg q)$ must be expressible in the language of $\overset{\text{x}}{\models}$. Then requirement (i) has the consequence that the language of $\overset{\text{x}}{\models}$ must contain modal operators and quantifiers. The language of $\overset{\text{x}}{\models}$ must contain modal operators because the $\Diamond(p \wedge \neg q)$ contains modal operators ($\Diamond$). The language of $\overset{\text{x}}{\models}$ must contain quantifiers because physical laws are quantified sentences and p has the fundamental physical laws as conjuncts. For simplicity, I will suppose that if $\overset{\text{x}}{\models}$ is a modal extension of classical logic (first-order logic), then $\overset{\text{x}}{\models}$ fulfills requirement (i)[13]. I will suppose further that $\overset{\text{x}}{\models}$ must be a consistent modal extension of classical logic[14].

Requirement (ii) states that $\overset{\text{x}}{\models}$ must be such that, for some ϕ, $\overset{\text{x}}{\models} \Diamond\phi$ and $\overset{\text{x}}{\not\models} \phi$. Then requirement (ii) has the consequence that $\overset{\text{x}}{\models}$ cannot have a Kripke-like semantics (see Kripke, 1959). In those logics, call them $\overset{\text{k}}{\models}$ (from Saul Kripke), a frame is a pair $\langle W, R\rangle$, where W is a non-empty set of possible worlds and R is a binary (accessibility) relation on W. A model is a triple $\langle W, R, \Vdash\rangle$, where $\langle W, R\rangle$ is a frame and $\Vdash$ is a satisfaction relation between elements of W and sentences in the language. The satisfaction relation is such that $w \Vdash \Box\phi$ iff $w' \Vdash \phi$ for all w' such that wRw'; $w \Vdash \Diamond\phi$ iff $w' \Vdash \phi$ for some w' such that wRw'; and as usual for non-modal sentences. A sentence ϕ is valid in a model $\langle W, R, \Vdash\rangle$ iff $w \Vdash \phi$ for all $w \in W$; ϕ is valid in a frame $\langle W, R\rangle$ iff ϕ is valid in $\langle W, R, \Vdash\rangle$ for all choices of $\Vdash$; ϕ is valid in $\overset{\text{k}}{\models}$ ($\overset{\text{k}}{\models} \phi$) iff ϕ is valid in $\langle W, R\rangle$ for all possible choices of W and the relevant Rs (the relevant Rs are often restricted given some properties).

[12] The construction of a reasoner includes writing a function which outputs in some ordering all and only the beliefs of the reasoner given the input of its epistemic situation (see section 1.1, def. 1.1.1).

[13] The conjunction of the fundamental physical truths and laws probably contains higher-order quantification (see Baltag and Smets, 2005), but the following arguments also hold for higher-order quantification.

[14] The theorems of classical logic are often thought to be a priori (Field, 1998).

In this context, for all ϕ, if $\overset{\mathrm{k}}{\models} \Diamond\phi$, then $\overset{\mathrm{k}}{\models} \phi$, what contradicts requirement (ii)[15].

There exist modal extensions of classical logic which fulfill requirements (i) and (ii). In fact, there exist modal extensions of classical logic which fulfill requirements (i)$-$(iv)[16]. Consider $\overset{\mathrm{r}}{\models}$ (from Rudolf Carnap) to be a modal extension of classical logic with a Carnap-like semantics (see Carnap, 1946). The difference between $\overset{\mathrm{r}}{\models}$ and $\overset{\mathrm{k}}{\models}$ is that $\overset{\mathrm{r}}{\models}$ has a fixed frame $\langle W, R\rangle$ with a fixed model $\langle W, R, \Vdash\rangle$, where the worlds in W represent all possible assignments for the atomic sentences in the language and R contains all possible pairs of worlds in W (fully connected). In this case, $\overset{\mathrm{r}}{\models} \phi$ iff $w \Vdash \phi$ for all $w \in W$ and $\overset{\mathrm{r}}{\models}$ fulfills requirements (i)$-$(iv). First, $\overset{\mathrm{r}}{\models}$ is a modal extension of classical logic[17]. Then $\overset{\mathrm{r}}{\models}$ fulfills requirement (i). Second, for all atomic ϕ, $\overset{\mathrm{r}}{\models} \Diamond\phi$ and $\overset{\mathrm{r}}{\not\models} \phi$[18]. Then $\overset{\mathrm{r}}{\models}$ fulfills requirement (ii). Third, for all ϕ, if $\overset{\mathrm{r}}{\not\models} \neg\phi$, then $\overset{\mathrm{r}}{\models} \Diamond\phi$[19]. Then $\overset{\mathrm{r}}{\models}$ fulfills requirement (iii). Finally, $\overset{\mathrm{r}}{\models}$ exhibits both necessitation and reflexivity[20]. Then $\overset{\mathrm{r}}{\models}$ fulfills requirement (iv).

In the literature, there exists a sound and complete axiomatization for a propositional version of $\overset{\mathrm{r}}{\models}$ (see Schurz, 2000). I don't know whether there exists a sound and complete axiomatization for a first-order version of $\overset{\mathrm{r}}{\models}$. If this is the case, then it may be possible to construct a strictly ideal reasoner for $\overset{\mathrm{r}}{\models}$ which validates (Z). The point that I want to

[15] Suppose that $\overset{\mathrm{k}}{\models} \Diamond\phi$. Then $\Diamond\phi$ is valid in all frames $\langle W, R\rangle$. Suppose that the relation R is irreflexive. Then $\Diamond\phi$ is valid in all irreflexive frames. Specifically, $\Diamond\phi$ is valid in the irreflexive singleton frame $\langle\{w\}, \varnothing\rangle$. Then, for all choices of $\Vdash$, $w \Vdash \Diamond\phi$. This is a contradiction because $w \Vdash \Diamond\phi$ only if there exists a $w' \Vdash \phi$ such that wRw', but there does not exist a w' such that wRw'. Then the relevant R are reflexive. Then $\Diamond\phi$ is valid in all reflexive frames. Specifically, $\Diamond\phi$ is valid in the reflexive singleton frame $\langle\{w\}, \{\langle w, w\rangle\}\rangle$. Then, for an arbitrary choice of $\Vdash$, $w \Vdash \Diamond\phi$. Then, for an arbitrary choice of $\Vdash$, $w \Vdash \phi$ ($w \Vdash \Diamond\phi$ iff $w' \Vdash \phi$ for some w' such that wRw'). Then $w \Vdash \phi$ for arbitrary w and $\Vdash$, what means that ϕ is a tautology. Then $\overset{\mathrm{k}}{\models} \phi$. Therefore, if $\overset{\mathrm{k}}{\models} \Diamond\phi$, then $\overset{\mathrm{k}}{\models} \phi$ (see Cohnitz, 2012, p. 68).

[16] For now, let's set aside requirement (v).

[17] Suppose that $\overset{\mathrm{r}}{\models} \phi$. Then, for all w, $w \Vdash \phi$. Since all ws are consistent (possible worlds), it is not the case that, for all w, $w \Vdash \neg\phi$ (in fact, for all w, $w \nVdash \neg\phi$). Then $\overset{\mathrm{r}}{\not\models} \neg\phi$.

[18] Consider an arbitrary atomic sentence ϕ and a world $w \in W$. There exists a w' such that $w' \Vdash \phi$ because ϕ is atomic and W contains all assignments for atomic sentences. wRw' because w is fully connected. Then $w \Vdash \Diamond\phi$. Since w is arbitrary, $\overset{\mathrm{r}}{\models} \Diamond\phi$. Also, there exists a w such that $w \nVdash \phi$ because ϕ is atomic and W contains all assignments for atomic sentences. Then $\overset{\mathrm{r}}{\not\models} \phi$. Therefore, $\overset{\mathrm{r}}{\models} \Diamond\phi$ and $\overset{\mathrm{r}}{\not\models} \phi$.

[19] Suppose that $\overset{\mathrm{r}}{\not\models} \neg\phi$. Then there exists a w' such that $w' \nVdash \neg\phi$, what means that $w' \Vdash \phi$ (def. possible world). For all $w \in W$, wRw' because R is fully connected. Then, for all $w \in W$, $w \Vdash \Diamond\phi$ and $\overset{\mathrm{r}}{\models} \Diamond\phi$. Therefore, if $\overset{\mathrm{r}}{\not\models} \neg\phi$, then $\overset{\mathrm{r}}{\models} \Diamond\phi$.

[20] Suppose that $\overset{\mathrm{r}}{\models} \phi$. Then, for all $w \in W$, $w \Vdash \phi$. Then, for all $w \in W$, all the $w' \in W$ such that wRw' are such that $w' \Vdash \phi$. Then, for all $w \in W$, $w \Vdash \Box\phi$ and $\overset{\mathrm{r}}{\models} \Box\phi$. Therefore, if $\overset{\mathrm{r}}{\models} \phi$, then $\overset{\mathrm{r}}{\models} \Box\phi$ (necessitation). Consider an arbitrary w. Suppose that $w \Vdash \phi$. wRw (R is fully connected), then $w \Vdash \Diamond\phi$. Then, for all w, if $w \Vdash \phi$, then $w \Vdash \Diamond\phi$. Then $\overset{\mathrm{r}}{\models} \phi \to \Diamond\phi$ (reflexivity).

stress here is that a strictly ideal reasoner for $\overset{\mathrm{r}}{\models}$ is not finite. More generally, a strictly ideal reasoner for a logic $\overset{\mathrm{x}}{\models}$ with properties (i)−(iv) is not finite (Cohnitz, 2012, p. 70). A finite reasoner is a reasoner with cognitive limitations such as finite memory and being able to execute only a finite number of inferential steps in a finite time interval. That a finite reasoner is able to execute only a finite number of inferential steps in a finite time interval may be modeled as requiring the set of beliefs of a finite reasoner to be recursively enumerable (Boolos et al., 2007, p. 3)[21]. Consider the following argument:

(1) Suppose that $\mathcal{R}$ is a strictly ideal reasoner for a $\overset{\mathrm{x}}{\models}$ with requirements (i)−(iv).

(2) $(\forall\phi)$ $\overset{\mathrm{x}}{\models} \phi$ iff $\overset{\mathrm{x}}{\models} \Box\phi$ (see below).

If $\overset{\mathrm{x}}{\models} \phi$, then $\overset{\mathrm{x}}{\models} \Box\phi$ (req. iv, necessitation). Suppose that $\overset{\mathrm{x}}{\models} \Box\phi$. Then $\overset{\mathrm{x}}{\models} \phi$ (req. iv, reflexivity). Therefore, $\overset{\mathrm{x}}{\models} \phi$ iff $\overset{\mathrm{x}}{\models} \Box\phi$.

(3) $(\forall\phi)$ $\overset{\mathrm{x}}{\not\models} \phi$ iff $\overset{\mathrm{x}}{\models} \Diamond\neg\phi$ (see below).

If $\overset{\mathrm{x}}{\not\models} \phi$, then $\overset{\mathrm{x}}{\models} \Diamond\neg\phi$ (req. iii, $\phi/\neg\phi$). Suppose that $\overset{\mathrm{x}}{\models} \Diamond\neg\phi$. Then $\overset{\mathrm{x}}{\models} \neg\Box\phi$ (definition of $\Diamond$). Then $\overset{\mathrm{x}}{\not\models} \Box\phi$ (consistency)[22]. Then $\overset{\mathrm{x}}{\not\models} \phi$ (from 2). Therefore, $\overset{\mathrm{x}}{\not\models} \phi$ iff $\overset{\mathrm{x}}{\models} \Diamond\neg\phi$.

(∴) $\mathcal{R}$ is not finite (see below).

Suppose that $\mathcal{R}$ is finite. Then $\overset{\mathrm{x}}{\models}$ is recursively enumerable (1, def. finite reasoner). $(\exists^{\infty}\phi)$ $\overset{\mathrm{x}}{\models} \Diamond\neg\phi$ (req. ii, $\phi/\neg\phi$, see note)[23] and $\overset{\mathrm{x}}{\models} \Diamond\neg\phi$ iff $\overset{\mathrm{x}}{\not\models} \phi$ (3). Then $\overset{\mathrm{x}}{\models}$ is enumerable iff $\overset{\mathrm{x}}{\models}$ is decidable[24]. If $\overset{\mathrm{x}}{\models}$ is decidable, then classical logic is decidable (req. i, supraclassicality). But classical logic is not decidable (Church, 1936a), what is a contradiction. Therefore, $\mathcal{R}$ is not finite.

[21] A set is recursively enumerable iff there exists a function which outputs all and only the members of the set given some ordering and executes at most finitely many operations in generating each output. This models the fact that the formation of each belief of a finite reasoner may depend at most of finitely many inferential steps (see chapter 1, def. 1.1.2).

[22] I am supposing that $\overset{\mathrm{x}}{\models}$ is consistent (req. i), but the problem presented here also may be avoided using a paraconsistent $\overset{\mathrm{x}}{\models}$ (Cohnitz, 2012, p. 73).

[23] Consider $\top$ to be a tautology. Then $\overset{\mathrm{x}}{\models} \top$ and, for all $w \in W$, $w \Vdash \top$. $(\exists\psi)$ $\overset{\mathrm{x}}{\models} \Diamond\psi$ (req. ii). Then, for some $w \in W$, $w \Vdash \psi$ and $w \Vdash \top$. Then $w \Vdash \psi \wedge \top$ (def. conjunction) and $\overset{\mathrm{x}}{\models} \Diamond\phi$, where $\phi = (\psi \wedge \top)$. There are infinitely many tautologies. Therefore, $(\exists^{\infty}\psi)$ $\overset{\mathrm{x}}{\models} \Diamond\psi$. If $\phi = \neg\psi$, then $(\exists^{\infty}\phi)$ $\overset{\mathrm{x}}{\models} \Diamond\neg\phi$.

[24] A set is recursively decidable iff both the set and its complement are recursively enumerable. The set $\overset{\mathrm{x}}{\models}$ is enumerable iff it is decidable because enumerating the infinitely many members $\Diamond\neg\phi$ of this set is equivalent to enumerate the set $\overset{\mathrm{x}}{\not\models}$, which is the complement of $\overset{\mathrm{x}}{\models}$.

There exists a different reason for which a strictly ideal reasoner for $\overset{\text{x}}{\models}$ is not finite. Requirement (v) states that $\overset{\text{x}}{\models}$ must be such that, for all superphysical truths ϕ, $\overset{\text{x}}{\models} p \rightarrow \phi$, where p is the conjunction of the fundamental physical truths and laws. Quantum mechanics (QM) is usually thought to be the best candidate for the fundamental physical description of the world. But QM is usually though to have an indeterministic character in the sense that the laws of QM and the full description of a quantum system at t_1 do not determine the full state of the system at $t_2 > t_1$[25]. In this context, the relation between a p with finitely many conjuncts and the complete set of physical truths (which are superphysical) cannot be deductive. The relation between a p with finitely many conjuncts and any complete set of non-probabilistic physical truths must be, at most, nonmonotonic. For this reason, a $\overset{\text{x}}{\models}$ which fulfills requirement (v) must, at most, be nonmonotonic[26]. However, a strictly ideal reasoner for a nonmonotonic $\overset{\text{x}}{\models}$ with essential properties is not finite[27].

In considering a similar argument, Chalmers argues that the conclusion that $\overset{\text{x}}{\models}$ must be nonmonotonic may be avoided by adding to p what he calls 'interpretative principles':

> On the collapse interpretation [of QM], a natural interpretative strategy is to say that an entity is located in a certain region of three-dimensional space if a high enough proportion of the (squared) amplitude of its wave-function is concentrated within that region. ...An interpretative principle involving 'a high enough proportion' will then deliver classical truth at both the microscopic and macroscopic level (Chalmers, 2012, p. 294-295).

But Chalmers's interpretative principles are themselves nondeductive. In the example, from the fact that there exists a region where a high enough proportion of the (squared) amplitude of its wave-function is concentrated, it does not *follow* that the entity is located

[25] There are interpretations under which QM is deterministic (Bohm, 1952), but quantum indeterminacy does not need to be actual. It is enough that quantum indeterminacy is possible because $\overset{\text{x}}{\models} \phi$ iff $\overset{\text{x}}{\models} \Box\phi$.

[26] In this case, a priori knowledge would be fallible and defeasible, what is often accepted in the literature (see Russell, 2013). First, a priori knowledge would be fallible because we are a priori justified in believing both premises of sorites paradox but one of these premises is false (Bealer, 1998, p. 202). Second, a priori knowledge would be defeasible because we were, all things considered, a priori justified in believing that every event has a cause but, after developments in physics, we are not anymore (Russell, 2013, p. 4).

[27] In section 1.3, I have argued that a strictly ideal reasoner for a nonmonotonic logic which exhibits supraclassicality, reflexivity, cut, and cautious monotony is not finite for situations with nonmonotonic consequences. Then an (a priori) strictly ideal reasoner for a nonmonotonic $\overset{\text{x}}{\models}$ which fulfills requirement (i)$-$(v) would not be finite because the theorems of this logic (e.g. $p \rightarrow \phi$) would be nonmonotonic.

in that region (argument due to Parent, 2014, p. 3). Then a logic which enables using interpretative principles must be nonmonotonic and the conclusion is not avoided.

In other occasions, Chalmers argues that the conclusion may be avoided by adding more information to p:

> The apparent failure of determinism in quantum mechanics suggests that the [Laplace's] demon [i.e. the strictly ideal reasoner] could not predict the future just from facts about physical laws and about the present. ...To avoid these problems, however, we need only give Laplace's demon more information than Laplace allows. To accommodate nondeterminism, we might give the demon full information about the distribution of the fundamental physical entities throughout space and time (Chalmers, 2012, p. xiv).

But, if p contains 'the full information about the distribution of the fundamental physical entities through space and time', then p must have infinitely many conjuncts. This is the case because time is often modeled as dense (see Ray, 1991, p. 20) and p would need to contain information about the position of the fundamental physical entities for infinitely many moments between every t_1 and t_2. In this case, a strictly ideal reasoner for $\overset{x}{\models}$ would not be finite. A finite reasoner has finite memory, but p is an infinite conjunction. This means that a finite strictly ideal reasoner for $\overset{x}{\models}$ cannot believe $p \to \phi$[28].

Then requirement (v) has the consequence that either $\overset{x}{\models}$ must be a nonmonotonic logic or that p is infinite. In each case, a strictly ideal reasoner for a $\overset{x}{\models}$ with property (v) is not finite[29]. This conclusion seems to be accepted by Chalmers (2010) and Menzies (1998). After all, both Chalmers and Menzies presuppose an ideal conceiver without cognitive limitations in their notion of ideal (negative) conceivability.

If a strictly ideal reasoner for $\overset{x}{\models}$ cannot be finite, it is doubtful how to evaluate (Z) and, consequently, (p1)* and (p2)*. More specifically, it is doubtful that we have a clear grasp about whether $p \wedge \neg q$ is ideally negatively conceivable. First, it is dubious whether finite

[28] There exist different notions of beliefs and, for some of them, a finite reasoner may believe $p \to \phi$ (e.g. implicit beliefs). But a finite reasoner cannot believe $p \to \phi$ for all notions of beliefs which require the reasoner to be able to explicitly believe $p \to \phi$ (e.g. explicit, accessible, stable beliefs). See sections 1.2 and 1.4 for definitions of these notions and for why the notions in the second group are more relevant.

[29] The two reasons for $\overset{x}{\models}$ with properties (i)$-$(v) not being recursively enumerable are related. Negative conceivability is a nonmonotonic principle (a principle which relies on the nonderivability of sentences) and $\overset{r}{\models}$ is closely related to autoepistemic logic (Gottlob, 1994), which is a system of nonmonotonic logic.

reasoners (e.g. humans) can, even in principle, grasp the sentence p or implement the hypercomputational inferential procedures necessary for enumerating the theorems of $\overset{x}{\models}$ (see Lokhorst, 2000)[30]. Second, it is also doubtful whether a finite reasoner is able to have the slightest idea about how a infinite reasoner reasons. The issue is that the patterns of inference of a reasoner without cognitive limitations may be fundamentally different from those of a finite reasoner. For example, a reasoner which has full information about the distribution of the fundamental physical entities through space and time 'predicts' the position of a fundamental physical entity at t simply by reinstating what it already presupposes and without making use of the fundamental physical laws[31]. This pattern of inference is fundamentally different from how finite reasoners reason about physics. Finally, if the notion of ideal negative conceivability is supposed to have a normative role in modal epistemology, the former example shows that how infinite reasoners reason can hardly be a norm for how finite reasoners should reason. Then it is doubtful whether ideal negative conceivability could be model using an (infinite) strictly ideal reasoner for $\overset{x}{\models}$ with requirements (i)$-$(v).

In answering a similar objection[32], Chalmers argues the following:

> I think that there is little reason to accept this claim. Although we are non-ideal, we can know that it is not ideally conceivable that $0 = 1$ and that it is ideally conceivable that someone exists. We know that certain things about the world (say, that all philosophers are philosophers) are knowable a priori and that certain things about the world (say, that there is a table in this room) are not so knowable even by an ideal reasoner (Chalmers, 2010, p. 155).

These analogies are not good: '$1 = 0$' is easily shown to entail a contradiction and 'someone exists' is easily derived from everyday knowledge (e.g. that I exist). These are cases for which we *do not* need ideal negative conceivability because these sentences are easy to grasp and evaluate. This is not the case for $\Diamond(p \wedge \neg q)$ and, consequently, for (Z).

[30] It is argued, for example, that the Bekenstein bound (Bekenstein, 1981) states an upper limit on the information that can be contained in a given finite region of space and that this would render inferential capacities higher than computability or infinite memory physically impossible (Lokhorst, 2000).

[31] For another example, if there exists a procedure for checking guesses, a reasoner which is able to perform inferences instantaneously may solve any problem (instantaneously) simply by generating and checking (a large number of) random guesses.

[32] The objection is: "arguments from ideal conceivability are toothless since nonideal creatures such as ourselves cannot know whether or not a given statement is ideally conceivable" (Chalmers, 2010, p. 155).

2.3 Defeasible ideal conceiver

In the last sections, I have argued that premises (p1)* and (p2)* of the negative zombie argument are true iff there exists a strictly ideal reasoner for $\overset{x}{\models}$ which believes $p \wedge \neg q$ on the basis of not believing $p \to q$. I have shown that, if $\overset{x}{\models}$ is conclusive, such a reasoner is not finite. Then I have argued that this fact undermines the use of ideal negative conceivability in the negative zombie argument. In this section, I argue that there may exist a finite strictly ideal reasoner for a nonmonotonic version of $\overset{x}{\models}$. However, this reasoner negatively conceiving ϕ after any amount of reasoning would only be a defeasible reason for $\Diamond\phi$.

Occasionally, Chalmers proposes a different definition of ideal negative conceivability:

Definition 2.3.1 (Ideal negative conceivability 2). A sentence ϕ is ideally negatively conceivable iff ϕ is negatively conceivable with justification that is undefeatable by better reasoning (Chalmers, 2002, p. 172),

where ϕ is negatively conceivable iff $\neg\phi$ is not believed on a priori basis.

The notion of undefeatability by better (further) reasoning is often used in the literature about nonmonotonic reasoning. In the theory of defeasible reasoning (Pollock, 1990, 1995), for example, there exists a notion of warrant which is closely related to undefeatability by better reasoning. The theory describes reasoning as the procedure of adopting beliefs on the basis of reasons and retracting beliefs on the basis of defeaters. Reasons are sets of mental states that provide epistemic support for believing something. A reason may be conclusive or defeasible depending on the epistemic support that it provides. The epistemic support of conclusive reasons cannot be overridden by new information (defeaters). For example, believing ϕ is a conclusive reason for believing $\phi \vee \psi$. The epistemic support of defeasible reasons can be overridden by defeaters. For example, perceiving an object as red is defeasible reason for believing that the object is red[33].

[33] A defeater may be a rebutting or an undercutting defeater. Rebutting defeaters attack the conclusion of a reason. For example, if perceiving at distance what appears to be a sheep in the field is defeasible reason for believing there is a sheep in the field, then hearing from the shepherd that there are no sheep in the field is a rebutting defeater for that reason (Pollock, 1995, p. 85). Undercutting defeaters attack the connection between a reason and its conclusion. For example, if perceiving an object as red is defeasible

Let $\mathtt{KB}_i$ be a set of reasons of a reasoner and $\mathtt{KB}_0, \mathtt{KB}_1, ..., \mathtt{KB}_i, ...$ be the reasoning sequence of the reasoner (how the reasoner would reason from the available information if it had enough cognitive resources, e.g. time). Then the notion of warrant is the following:

Definition 2.3.2 (Warrant). A belief ϕ is warranted relatively to $\mathtt{KB}_0, \mathtt{KB}_1, ..., \mathtt{KB}_i, ...$ iff there is an i such that, for all $j \geq i$, ϕ is undefeated relatively to $\mathtt{KB}_j$ (Pollock, 1995, p.133).

In other words, the belief that ϕ is warranted iff there is a stage in the reasoning sequence such that the belief is undefeated at every subsequence stage. Then the belief that ϕ is undefeatable by better (further) reasoning iff that belief is warranted[34].

In chapter 1, I have presented a model of a reasoner which provides a clear definition for the notion of a reasoning sequence used in the notion of warrant (def. 1.1.1). In the model, an (a priori) reasoner is composed of a formal language $\mathcal{L}$, a set of sentences in $\mathcal{L}$ ($\mathtt{KB}$) which model the explicit beliefs of the reasoner, and a belief update function $\pi : 2^{\mathcal{L}} \times \mathbb{Z}^+ \rightarrow 2^{\mathcal{L}}$ which models the pattern of inference of the reasoner. A fact about the pattern of inference of a reasoner is that the reasoner may perform different inferences from the same premises. In the model, this fact is expressed using a function π with a numeric parameter (integer) in addition to the parameter for $\mathtt{KB}$. In this context, $\pi(\mathtt{KB}, 1)$ models an inference from $\mathtt{KB}$, $\pi(\mathtt{KB}, 2)$ models another inference from $\mathtt{KB}$, etc. Then function π determines a reasoning sequence $\mathtt{KB}_0, \mathtt{KB}_1, \ldots, \mathtt{KB}_i, \ldots$, where $\mathtt{KB}_0 = \mathtt{KB}$ is the initial set of explicit beliefs and $\mathtt{KB}_{i+1} = \pi(\mathtt{KB}_i, i + 1)$. Supposing that the integer parameter models some order of priority, the reasoning sequence of a reasoner models how the reasoner would reason from the available information if it had enough cognitive resources (e.g. time). In this context, let the explicit beliefes of a reasoner at $\mathtt{KB}_i$ (beliefs$_{ex}$) be the sentences at $\mathtt{KB}_i$ and the stable beliefs of a reasoner (beliefs$_\omega$) be the sentences in $KB_\omega = \bigcup_i \bigcap_{j \geq i} KB_j$.

A sentence is in $\mathtt{KB}_\omega$ iff there exists an i such that, for all $j \geq i$, $\phi \in \mathtt{KB}_j$. Then ϕ is warranted iff $\phi \in \mathtt{KB}_\omega$. But ϕ is undefeatable by better (further) reasoning iff ϕ is

reason for believing the object is red, then learning that the object is under red light is an undercutting defeater for that reason. Undercutting defeaters are reasons for believing that a reason does not support a conclusion (Pollock, 1995, p. 86).

[34] The phrase 'undefeatable by *better* reasoning' seem to have an evaluative character in the negative conceivability principle ('better reasoning' in the sense of more rational reasoning). I will contemplate this character dealing with strictly ideal reasoning sequences.

warranted. Then ϕ is undefeatable by better (further) reasoning iff $\phi \in \texttt{KB}_\omega$. The belief that ϕ is undefeatable by better (more rational) reasoning iff ϕ is in a strictly ideal $\texttt{KB}_\omega$ (the $\texttt{KB}_\omega$ of a strictly ideal$_\omega$ reasoner).

In this context, (Z) may be interpreted as following:

(Z)* There exists a strictly ideal$_\omega$ reasoner for a logic $\overset{x}{\models}$ with properties (i)$-$(v) which believes$_\omega$ $\Diamond(p \wedge \neg q)$ on the basis of not believing$_{ex}$ $p \rightarrow q$ after some amount of reasoning.

If (Z)* is a reasonable interpretation of (Z), then (p1)$*$ and (p2)$*$ are true only if (Z)* is true. The first step in evaluating (Z)* is to investigate how $\overset{x}{\models}$ looks like. The logic $\overset{r}{\models}$ fulfills requirements (i)$-$(iv) and $\overset{x}{\models}$ must be nonmonotonic in order to fulfill requirement (v). Then a natural choice is to make $\overset{x}{\models}$ a nonmonotonic extension of $\overset{r}{\models}$. However, the way $\overset{r}{\models}$ deals with requirement (iii) has the consequence of making a strictly ideal reasoner for $\overset{x}{\models}$ not finite[35]. Then $\overset{x}{\models}$ must deal with this requirement differently. More specifically, $\overset{x}{\models}$ must deal with requirement (iii) using nonmonotonic rules. For example, $\overset{x}{\models}$ may use the following defeasible rule for $\Diamond$ introduction and its defeater[36]:

Definition 2.3.3 (Defeasible $\Diamond$I). Not believing$_{ex}$ $\neg\phi$ on a priori basis is defeasible reason for believing$_{ex}$ $\Diamond\phi$.

Definition 2.3.4 (Defeater for defeasible $\Diamond$I). Believing$_{ex}$ $\neg\phi$ on a priori basis is a defeater for defeasible $\Diamond$I.

In order to deal with requirement (v), a reasoner may have the following defeasible rule and defeater, where a is an arbitrary individual and F and G are arbitrary properties:

[35] It is because $\overset{r}{\models} \Diamond\neg\phi$ iff $\overset{r}{\not\models} \phi$ that $\overset{r}{\models}$ is (recursively) enumerable iff $\overset{r}{\models}$ is decidable and a strictly ideal reasoner for $\overset{r}{\models}$ is not finite. More generally, the patterns of inference of a finite strictly ideal reasoner for $\overset{x}{\models}$ cannot depend on checking whether $\overset{x}{\not\models} \phi$ or $\phi \notin \texttt{KB}_\omega$.

[36] These are simplified glosses, omitting important qualifiers and details. For example, I am assuming that all defeasible reasons generate the same amount of epistemic support. For a discussion about reasons with different degrees of epistemic support, see Pollock (1995, p. 93).

Definition 2.3.5 (Statistical syllogism). If $r \geq .5$, then believing$_{ex}$ $pr(F|G) > r$ and Ga is defeasible reason for believing Fa (Pollock, 1995, p. 68)[37].

Definition 2.3.6 (Defeater for statistical syllogism). Believing Ha and believing$_{ex}$ $pr(F|G\&H) < pr(F|G)$ is a defeater for the statistical syllogism (Pollock, 1990, p. 9).

The reasoner may use statistical syllogism in, for example, inferring from the fact that a high enough proportion of the (squared) amplitude of its wave-function is concentrated within in a region that the electron is in that region (Chalmers's interpretative principles). In order to deal with the statistical syllogism, $\overset{x}{\models}$ must contain rules and axioms necessary for dealing with probabilities ($\overset{x}{\models}$ is probabilistic).

In this context, it may be possible to construct a finite strictly ideal$_{\omega}$ reasoner for $\overset{x}{\models}$ which verifies (Z)*[38]. Let $\overset{x}{\models}$ be a nonmonotonic extension of $\overset{r}{\models}$ with a nonmonotonic level corresponding to the rules of defeasible $\Diamond$I and statistical syllogism along with the respective defeaters. The reasoner would be constructed in such a way that, at each stage of the reasoning sequence, it attempts to apply every rule of $\overset{x}{\models}$ to all possible combinations of sentences that it believes$_{ex}$. The key point is to construct the reasoner in such a way that, every time it concludes that ϕ using a nonmonotonic rule, it checks, after each stage of the reasoner sequence, if a defeater was derived. If a defeater was derived, the reasoner withdraws the conclusion that ϕ (along with all conclusions derived from ϕ). There exist two possibilities: either (i) the defeater is derived and ϕ is deleted at some point of reasoning or (ii) the defeater is not derived and ϕ is not deleted. In any case, the reasoner would believe$_{\omega}$ ϕ iff $\overset{x}{\models} \phi$ and the reasoner would be strictly ideal$_{\omega}$ for $\overset{x}{\models}$. The reasoner would be finite because it does not need to generate $\texttt{KB}_{\omega}$ ($\texttt{KB}_{\omega}$ is not part of the reasoning sequence) and no rule in $\overset{x}{\models}$ depends on checking whether $\overset{x}{\not\models} \phi$ ($\phi \notin \texttt{KB}_{\omega}$), but only on checking whether $\phi \notin \texttt{KB}$.

[37] The probabilities involved in this principle are indefinite probabilities, where $pr(F|G)$ means 'the proportion of physically possible Gs that would be Fs', The value for r must be low enough for fulfilling requirement (v). I will suppose that $r = .5$ (the lowest reasonable value), but the argument in next section holds for values of r very close to 1.

[38] The possibility of constructing such a reasoner depends on, for example, the existence of an adequate semantics for $\overset{x}{\models}$ along with a sound and complete axiomatization. I will grant all these conditions in order to evaluate other problems for the negative zombie argument.

If the construction of a finite strictly ideal$_\omega$ reasoner for $\overset{\mathrm{x}}{\models}$ is indeed possible and that reasoner believes$_\omega$ $\Diamond p \wedge \neg q$ on the basis of not believing$_{ex}$ $p \rightarrow q$, then there exists good prospects for (p1)$*$ and (p2)$*$ being both true and the negative zombie argument being sound. However, the use of strictly ideal$_\omega$ reasoners as the model of ideal negative conceivability has some consequences for the negative zombie argument. The first consequence is that whereas the patterns of inference of a strictly ideal reasoner for $\overset{\mathrm{x}}{\models}$ are unreachable for a finite reasoner[39], a strictly ideal$_\omega$ reasoner for $\overset{\mathrm{x}}{\models}$ may be seen as an extrapolation of the patterns of inference of a finite reasoner. In this sense, ideal negative conceivability would be a more amenable principle of modal epistemology for finite reasoners. The second consequence is that the relation between ideal negative conceivability and possibility cannot be of entailment. A strictly ideal$_\omega$ reasoner for $\overset{\mathrm{x}}{\models}$ believing$_{ex}$ $\Diamond\phi$ at any stage of a reasoning sequence is, at most, a defeasible reason for $\Diamond\phi$ being true. In this context, ideal negative conceivability would be a defeasible principle of modal epistemology and the negative zombie argument would not be conclusive. The third consequence is that an ideal$_\omega$ reasoner for $\overset{\mathrm{x}}{\models}$ may not believe$_\omega$ $\Diamond(p \wedge \neg q)$ for every choice of q (see next section).

2.4 Someone is conscious!

The negative zombie argument is such that, for all superphysical ϕ, $\overset{\mathrm{x}}{\models} p \rightarrow \phi$. The exact nature of p is very difficult to grasp, but is much less controversial that some ϕ are superphysical. For example, it is uncontroversial that the truths of neuroscience n (e.g. ‘C-fiber x is stimulated’ or ‘neural activity on area x is enhanced’) are superphysical. What about truths of the form ‘the probability of q is x’ (i.e. $pr(q) = x$) and ‘n and q have a correlation of x’ (i.e. $corr(n, q) = x$), where n is a sentence of neuroscience, q is a phenomenal sentence, and x is a number? Are those truths superphysical? I think that this is the case if these sentences are interpreted using a notion of (natural or nomological) probabilities defined in terms of (proportions of) physically possible worlds (see Pollock, 1990). In this case, p contains the fundamental physical laws, which determine the physically possible worlds. Possible worlds are saturated entities. So, given any physically possible world w,

[39] The same for the patterns of inference of a strictly ideal$_{ac}$ reasoner for $\overset{\mathrm{x}}{\models}$ (see section 1.3).

either n is true at w or false at w (and the same holds for q). Then, for any $pr(q) = x$, the value of x is determined by p. And, for any $corr(n, q) = x$, the value x is determined by p. A change on these values would entail a change on the fundamental physical laws and, consequently, in p. Then $pr(q) = x$ and $corr(n, q) = x$ are superphysical in this interpretation. Chalmers accepts that the relevant interpretation of probabilities in those cases is that using physically possible worlds[40]. In the following, I will assume that all truths of the form n, $p(q) = x$, and $corr(n, q) = x$ are superphysical and, consequently, that $\overset{\mathrm{x}}{\models} p \rightarrow n$, $\overset{\mathrm{x}}{\models} p \rightarrow pr(q) = x$, and $\overset{\mathrm{x}}{\models} p \rightarrow corr(n, q) = x$.

There exist different correlation coefficients. The most common is Pearson correlation coefficient, which is sensitive to linear relationships between random variables. Pearson correlation coefficient between two variables X and Y ($corr(X, Y)$) is obtained by dividing the covariance of the variables by the product of their standard deviations:

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

’ The value of $corr(X, Y)$ is such that $-1 \leq corr(X, Y) \leq 1$. A positive value of $corr(X, Y)$ indicates that Y tends to increase when X does, a negative value indicates that Y tends to decrease when X increases and the value 0 means that X and Y are uncorrelated. In statistics, an indicator variable is a (discrete) variable X which may only take the values of 0 ($X = 0$) or 1 ($X = 1$). For indicator variables, the notion of correlation entails that[41]:

- $corr(X, Y) > 0$ iff $pr(Y = 1|X = 1) > pr(Y = 1)$.

[40] “The sort of possibility being considered here is natural or nomological possibility or possibility compatible with the laws of nature. If we required correlation across all logically possible cases, there might be no total NCC [neural correlate of consciousness] at all, as it is arguably logically possible or coherently conceivable to instantiate any physical process at all without consciousness. If we require correlation across naturally possible cases, the problem goes away, as these cases are probably not naturally possible” (Chalmers, 2010, p. 74).

[41] The correlation between two random variables X and Y is such that $corr(X, Y) = cov(X, Y)/(\sigma_X \sigma_Y)$ and $cov(X, Y) = E[XY] - E[X]E[Y]$, where $E[X]$ is the expected value of X. If X and Y are indicator variables, $E[X] = pr(X = 1)$, $E[Y] = pr(Y = 1)$, and $E[XY] = pr(X = 1 \wedge Y = 1)$. Then $corr(X, Y) = [pr(X = 1 \wedge Y = 1) - pr(X = 1)pr(Y = 1)]/(\sigma_X \sigma_Y)$. Since σ_X and σ_Y are always positive, $corr(X, Y) > 0$ iff $pr(X = 1 \wedge Y = 1) - pr(X = 1)pr(Y = 1) > 0$. But $pr(X = 1 \wedge Y = 1) - pr(X = 1)pr(Y = 1) = pr(X = 1)pr(Y = 1|X = 1) - pr(X = 1)pr(Y = 1) = pr(X = 1)[pr(Y = 1|X = 1) - pr(Y = 1)]$ and $pr(X = 1)$ is always positive. Then $corr(X, Y) > 0$ iff $pr(Y = 1|X = 1) - pr(Y = 1) > 0$. Therefore $corr(X, Y) > 0$ iff $pr(Y = 1|X = 1) > pr(Y = 1)$.

In a two-valued logic, sentences are indicator variables, where $\neg x$ is the value $X = 0$ and x is the value $X = 1$. Then, in terms of n and q the above theorem entails that:

- $corr(n, q) > 0$ iff $pr(q|n) > pr(q)$.

I intend to show that $p \wedge \neg q$ is not ideally negatively conceivable for every choice of q. In the (Z)* model, if $p \wedge \neg q$ is ideally negatively conceivable for every choice of q, then, for every choice of q, a strictly ideal$_\omega$ reasoner for $\overset{x}{\models}$ does not believe$_{ex}$ $p \rightarrow q$ after any amount of reasoning. I intend to show that, for some q, a strictly ideal$_\omega$ reasoner for $\overset{x}{\models}$ does believe$_{ex}$ $p \rightarrow q$ after some amount of reasoning. More specifically, I intend to show that a strictly ideal$_\omega$ reasoner for $\overset{x}{\models}$ believes$_\omega$ $p \rightarrow q$ for every q such that $pr(q) \geq .5$ and for which there exists a true n such that $corr(n, q) > 0$[42]. This is the case because n is a truth of neuroscience and $\overset{x}{\models} p \rightarrow n$ (n is superphysical). Given that $corr(n, q) > 0$, then $\overset{x}{\models} p \rightarrow corr(n, q) > 0$ ($corr(n, q) > 0$ is superphysical). But $corr(n, q) > 0$ iff $pr(q|n) > pr(q)$ and $\overset{x}{\models}$ is probabilistic, then $\overset{x}{\models} p \rightarrow pr(q|n) > pr(q)$. Given that $pr(q) \geq .5$ is superphysical, then $\overset{x}{\models} p \rightarrow pr(q|n) > .5$. Then $\overset{x}{\models} p \rightarrow (n \wedge pr(q|n) > .5)$. Then, applying the statistical syllogism, $\overset{x}{\models} p \rightarrow q$. Therefore, a strictly ideal$_\omega$ reasoner for $\overset{x}{\models}$ believes$_\omega$ $p \rightarrow q$ (def. 2.1.1, req. i).

Empirical data suggest that there exist truths n and q such that $corr(n, q) > 0$. There exists an extensive literature on neural correlates of conscious which supports this claim (see Tononi and Koch, 2008, for a review). I think that it is dubious how to establish $pr(q)$ for a given q, but it is very likely that $pr(q) \geq .5$ for many qs of interest. For example, it is very likely that $pr(q) \approx 1$ for $q =$'someone is conscious'. The number humans alive is approximately 7×10^9. Let $q_1 =$'human 1 is conscious' and, more generally, $q_i =$'human i is conscious'. Then $pr(q) = 1 - \prod_{i=1}^{7\times 10^9} pr(\neg q_i)$. Then $pr(q) \approx 1$ (.999) when the probability of each human being conscious is as low as 1×10^{-9}. And $pr(q) = .503$ (what is still greater than or equal to .5) when the probability of each human being conscious is 1×10^{-10}. Then the conditions of the argument in last paragraph are very likely fulfilled for some true n and q.

[42] I am supposing that the n and q such that $corr(n, q) = x$ are about the same individual. For example, $n = N(a)$ and $q = Q(a)$, where a is some individual.

Therefore, a strictly ideal$_\omega$ reasoner for $\overset{x}{\models}$ does believe$_{ex}$ $p \to q$ after some amount of a priori reasoning (because it believes$_\omega$ $p \to q$), which means that $p \wedge \neg q$ is not ideally negatively conceivable (in the (Z)* model) and that the negative zombie argument is not sound for $q =$'someone is conscious'. The same hold for all q such that $pr(q) \geq .5$ and for which there exists some truth n such that $corr(n, q) > 0$.

In commenting this argument, the first interesting point is that it is not affected Chalmers's structure and dynamics counterargument:

> First, physical descriptions of the world characterize the world in terms of structure and dynamics. Second, from truths about structure and dynamics, one can deduce only further truths about structure and dynamics. Third, truths about consciousness are not truths about structure and dynamics (Chalmers, 2010, p. 120).

Let's grant that (a) physical truths (e.g. the conjuncts in p) are about structure and dynamics, (b) that phenomenal truths aren't about structure and dynamics, (c) and that there's no deduction from sentences of one form to sentences of a different form. Even so, the structure and dynamics counterargument does not affect the argument in this section. Premise (c) may be true of deductions (which are valid by virtue of form), but Chalmers's own presuppositions entails that $\overset{x}{\models}$ (the logic of a priori knowledge) cannot be deductive and (c) is not true of inductions. Second, since all superphysical truths follows from p in $\overset{x}{\models}$, saying that q should not follow from p in $\overset{x}{\models}$ without an independent reason is begging the question on q not being superphysical and on physicalism being false.

There are other objection which may appear. First, it may be questioned whether the statistical syllogism is a rational rule of inference. This objection is not very strong. The statistical syllogism is successfully in our everyday and scientific reasoning. More acutely, it may be questioned whether the statistical syllogism is a rational rule of inference for a priori reasoning. There exists some discussion in the literature about whether the statistical syllogism is a rational rule for a priori reasoning (see Russell, 2013). But I do not want to discuss this issue here. My point is only that Chalmers's negative zombie argument requires something like the statistical syllogism to be a rational rule of inference for a priori reasoning. Second, it may be questioned whether the conclusion of the reasoning is an artifact of a too low value of r (i.e. $r = .5$). It may be the case that the specific line of reasoning presented above depends on this specific value of r, but it is possible to construct similar lines of reasoning for values of r very close to 1 because, for some qs of interest (e.g. $q =$'someone is conscious'), $pr(q) \approx 1$.

2.5 Conclusions

In this chapter, I have argued that premises (p1)* and (p2)* of the negative zombie argument are not both true for every choice of q. For example, (p1)* and (p2)* are not both true for $q =$'someone is conscious'. In section 2.1, I have argue that (p1)* and (p2)* are both true iff there exists a strictly ideal reasoner for a logic with some properties and that reasoner negatively conceiving ϕ is a (conclusive) reason for $\Diamond\phi$. In section 2.2, I have investigated whether these conditions hold and conclude that, even if this was the case, the conclusive reasoner would not be finite. I have argued that this fact undermines the use of the ideal negative conceivability principle in the negative zombie argument. In section 2.3, I have argued that there exists a finite strictly ideal reasoner for a related logic, but that this reasoner negatively conceiving ϕ after any amount of reasoning would only be a defeasible reason for $\Diamond\phi$. In section 2.4, I argue that the defeasible reasoner does not negatively conceive $p \wedge \neg q$ for every choice of q. This reasoner does not negatively conceive all q such that $pr(q)$ and $\exists n(n \wedge corr(n, q) > 0)$ (e.g. $q =$'someone is conscious').

In this context, the negative zombie argument may still be sound for some q, but it soundness would depend on which q is chosen. I will not answer the question about whether there exists a q which satisfies the negative zombie argument. My point is that this is an empirical question. In order to establish whether some specific truth q satisfies the negative zombie argument, one needs empirical information for supporting either that $pr(q) < .5$ or that, for all truths n, $corr(n, q) \leq 0$. But any empirical information for $pr(q) < .5$ or (more dramatically) for $\forall n(n \rightarrow corr(n, q) \leq 0)$ will not be conclusive. The bottom line is that the negative zombie argument is neither an a priori argument nor a conclusive argument against physicalism.

In several passages, Chalmers argues that ideal positive conceivability entails ideal negative conceivability. For example, consider the following (where S is a sentence):

> Ideal primary positive conceivability entails ideal primary negative conceivability: if S can be ruled out a priori, then no coherent imagined situation will verify S (Chalmers, 2010, p. 148).

If this is the case, $p \wedge \neg q$ not being ideally negatively conceivable for an arbitrary q has the consequence of $p \wedge \neg q$ not being ideally positively conceivable for an arbitrary q. In addition, if ideal negative conceivability and ideal positive conceivability are the only two relevant kinds of ideal conceivability, this has the consequence of $p \wedge \neg q$ not being ideally conceivable in general for an arbitrary q. Then, if the reasoning of Chalmers is correct, the conclusions presented here have consequences for zombie arguments in general.

Chapter 3

Achilles and the truth-goal

> Achilles was still seated on the back of the much-enduring Tortoise, and was writing in his notebook, which appeared to be nearly full. The Tortoise was saying, "Have you got that last step written down? Unless I've lost count, that makes a thousand and one. There are several millions more to come".
> Lewis Carroll, *What the Tortoise said to Achilles*.

In last chapters, I have been investigating the maximum and minimum bounds of rationality for finite reasoners without considering the truth-value of beliefs. In this chapter, I investigate the maximum and minimum bounds of *epistemic rationality* for finite reasoners[1]. The notion of epistemic rationality is usually understood as being related to the truth-value of beliefs.

> Theoretical reason [i.e. epistemic reason]... addresses the considerations that recommend accepting particular claims as to what is or is not the case. That is, it involves reflection with an eye to the truth of propositions, and the reasons for belief in which it deals are considerations that speak in favor of such propositions' being true, or worthy of acceptance (Wallace, 2014).

More specifically, the notion of epistemic rationality is usually understood as being related to the fulfillment of the truth-goal. Roughly, the truth-goal is the following goal:

Definition 3.0.1 (Truth-goal). The truth-goal is the goal of believing truths and not believing falsehoods.

[1] I focus on the maximum bounds of epistemic rationality for finite reasoners. Occasionally, I discuss the minimum bounds as well.

In fact, epistemologists usually argue that the truth-goal is *the* fundamental goal for epistemic rationality. Alston and BonJour, for example, are quite explicit on this matter:

> Epistemic evaluation is undertaken from what we might call 'the epistemic point of view'. That point of view is defined by the aim of maximizing truth and minimizing falsity in a large body of beliefs (Alston, 1985, p. 59).
>
> What makes us cognitive beings at all is our capacity for belief, and the goal of our distinctively cognitive endeavors is *truth*: we want our beliefs to correctly and accurately depict the world (BonJour, 1985, p. 7-8).

Similar claims are found in Lehrer (2000), Foley (1993), Plantinga (1993), Goldman (1986), Moser (1985), Sosa (1985), and Chisholm (1982)[2].

In this chapter, I investigate the maximum and minimum bounds of epistemic rationality for finite reasoners given a notion of epistemic rationality related to the fulfillment of the truth-goal. In chapter 1, I have investigated the notion of maximum rationality as you believing all and only what you ought to believe and believing only what you may believe[3]. Under the supposition that epistemic rationality is a species of rationality, I will now investigate the proposal that a maximally epistemically rational reasoner is a maximally rational reasoner which fulfills the truth-goal. The problem with this proposal is that the requirements for the fulfillment of the truth-goal are not clear in the literature and most of the existing models have problems with blindspots. In this chapter, I survey the existing models of the truth-goal, propose a model of the truth-goal which does not have problems with blindspots, and investigate the maximum and minimum bounds of epistemic rationality for finite reasoners given the proposed model of the truth-goal.

In section 3.1, I discuss some existing models of the fulfillment of the truth-goal and argue that these models have problems with blindspots. I then argue that the fulfillment of the truth-goal is better modeled as the maximization of a function g $-$ which accepts a set of beliefs as input $-$ and returns a numeric evaluation and discuss the properties

[2] List due to David (2001).

[3] In chapter 1, I have defined a maximally rational reasoner as believing all and only what it ought to believe and only what it may believe (def. 1.1.6), a rational reasoner as believing only what it may believe (def. 1.1.7), and an irrational reasoner as believing something that it may-not believe (def. 1.1.8). In our framework, this entails that a maximally rational reasoner believes all and only the logical consequences of its explicit beliefs, a rational reasoner has a non-trivial set of beliefs, and an irrational reasoner has a trivial set of beliefs (proof C.1).

that the function must have in order to model the fulfillment of the truth-goal. In section 3.2, I show how the function g model may be used in stating the maximum and minimum bounds of epistemic rationality. I investigate the details concerning the implementation of this model and argue that this is the best model for stating the maximum and minimum bounds of epistemic rationality for finite reasoners (among other things, because it deals well with blindspots). In the discussion, I compare the function g model with other quantitative models of rationality, as, for example, the group of models known as 'epistemic utility theory' (e.g. Fitelson and Easwaran, 2015; Pettigrew, 2015a; Joyce, 1998).

3.1 The truth-goal

The truth-goal is usually expressed using epistemic norms in the form of (bi)conditionals. The simplest version of these (bi)conditional norms would be the following:

(T) $\mathcal{R}$ ought to believe ϕ iff ϕ is true.

The problem with (T) is that it has the consequence of making maximum epistemic rationality impossible in many interesting situations. More specifically, (T) has the consequence of making the class of maximally epistemically rational reasoners a priori empty for languages which allow for the expression of blindspots (Bykvist and Hattiangadi, 2007). Informally, a blindspot for a reasoner is a sentence that may be true but cannot be both true and believed by the reasoner (Sorensen, 1988)[4]. The definition is the following:

Definition 3.1.1 (Blindspot). A sentence ϕ is a blindspot for a reasoner $\mathcal{R}$ iff it is possible that ϕ and it is not possible that (ϕ and $\mathcal{R}$ believes ϕ).

Consider the following two blindspots for an arbitrary reasoner $\mathcal{R}$:

(bs1) = '$\mathcal{R}$ does not believe (bs1)'.

(bs2) = $\phi \wedge \psi$, where ϕ is a mundane truth (e.g. 'snow is white') and ψ = '$\mathcal{R}$ does not believe ϕ'.

[4] Sorensen allows for a much wider application of the term 'blindspot'. He defines a blindspot for a propositional attitude A and a reasoner $\mathcal{R}$ as a sentence that is possibly true but cannot have attitude A taken towards it by $\mathcal{R}$ (Sorensen, 1988). I am dealing only with A =true-belief blindspots.

The sentence (bs1) is the core example of a blindspot. That sentence has the interesting property of being true iff $\mathcal{R}$ does not believe (bs1). Whether (bs2) is a blindspot depends on whether we accept that 'belief' distributes over conjunctions (if $\mathcal{R}$ believes $\phi \wedge \psi$, then $\mathcal{R}$ believes ϕ and $\mathcal{R}$ believes ψ)[5]. If we accept that 'belief' distributes over conjunctions, it is easy to see that (bs2) is a blindspot. In this case, if $\mathcal{R}$ believes (bs2), then $\mathcal{R}$ believes ϕ and (bs2) is false because its second conjunct is false. Furthermore, if (bs2) is true, then $\mathcal{R}$ does not believe (bs2) because, otherwise, the second conjunct of (bs2) and (bs2) itself would be false[6].

In this context, (T) has the consequence of making the class of maximally epistemically rational reasoners a priori empty for languages which allows for the expression of blindspots. Consider the following argument[7]:

(1) Let $\mathcal{R}$ be a maximally epistemically rational reasoner.

(2) $\mathcal{R}$ believes (bs1) iff $\mathcal{R}$ ought to believe (bs1) (from 1).

(3) $\mathcal{R}$ ought to believe (bs1) iff (bs1) is true (from (T)).

(4) (bs1) is true iff $\mathcal{R}$ does not believe (bs1) (definition of (bs1)).

(5) $\mathcal{R}$ believes (bs1) iff $\mathcal{R}$ does not believe (bs1) (from 1, 2, and 3).

(6) $\perp$ (from 5).

In order to deal with blindspots, epistemologists often propose modified versions of (T). Boghossian (2003), for example, attempts to avoid the problem by rejecting one direction of (T). Boghossian accepts:

(Ta) If $\mathcal{R}$ ought to believe ϕ, then ϕ is true.

[5] In our framework, this assumption is unnecessary because, under the right conditions, a maximally rational reasoner $\mathcal{R}$ believes the logical consequences of its beliefs. If ϕ and ψ are logical consequences of $\phi \wedge \psi$, the logic in question has some essential features (e.g. cut), and $\mathcal{R}$ believes $\phi \wedge \psi$, then $\mathcal{R}$ believes ϕ and $\mathcal{R}$ believes ψ.

[6] I am focusing on (bs1) and (bs2) because they are simple examples of blindspots. These two blindspots involve second-order beliefs, but there are interesting first-order blindspots (e.g. '$\mathcal{R}$ is dead').

[7] This argument may be stated in terms of (bs2) with minor changes.

But he rejects:

(Ta') If ϕ is true, then $\mathcal{R}$ ought to believe ϕ.

The retraction from (T) to (Ta) avoids the original problem. It does not follow from (Ta) that $\mathcal{R}$ ought to believe a blindspot. But (Ta) is too weak (Raleigh, 2013, p. 248). The problem with (Ta) is that, regardless of ϕ being true or false, it does not entail an obligation. If ϕ is true, (Ta) does not entail anything. If ϕ is false, (Ta) entails only that $\mathcal{R}$ does not ought to believe ϕ (not that $\mathcal{R}$ ought not to believe ϕ). As a consequence, (Ta) underdeterminates the description of a maximally epistemically rational reasoner: (Ta) says what the reasoner does not believe, but it says nothing about what it believes. This problem is revealed by the fact that (Ta) may be fulfilled when a reasoner does not have any beliefs at all.

Raleigh (2013, p. 249) and Whiting (2010) propose a different biconditional norm:

(Tb) $\mathcal{R}$ may believe ϕ iff ϕ is true.

This formulation also avoids the original problem. It does not follow from (Tb) that $\mathcal{R}$ ought to believe a blindspot. But (Tb) has a different problem with blindspots: it is a consequence of (Tb) that a reasoner is permitted to believe a currently true blindspot, and yet, if the reasoner were to believe the blindspot, (Tb) would forbid the belief. Raleigh (2013, p. 249) argues that this is a good feature of his proposal:

> It seems to me that as much as we may have an intuition that attempting to believe a blindspot that happens to be true should be discouraged, we also have a conflicting intuition that we *would like* to believe such a true proposition.... A feature of my proposal is that there is no truth-norm forbidding belief in a currently nonbelieved blindspot, but there is a norm that kicks in the moment that one ventures to believe the blindspot. I think this may be seen as a virtue rather than a vice, in that it does justice, as far as possible, to *both* of these conflicting intuitions (Raleigh, 2013, p. 252).

I don't think that appealing to those conflicting intuitions helps here because the instantiation of (Tb) for a blindspot contradicts the very definition of the epistemic 'may'. Suppose that $\mathcal{R}$ does not believe (bs1). In this case, (bs1) is true and (Tb) says that $\mathcal{R}$ may believe (bs1). If $\mathcal{R}$ may believe (bs1), then there must exist an epistemically permissible situation in which $\mathcal{R}$ believes (bs1). But there does not exist such an epistemically permissible situation according to (Tb): in all situations in which $\mathcal{R}$ believes (bs1), (bs1) is false, and (Tb) forbids the belief in (bs1). Therefore, (Tb) contradicts the definition of the epistemic 'may'. Another problem is that (Tb) may be fulfilled by not having beliefs.

Another attempt to deal with the problem is to add the following condition to (T):

(Tc) $\mathcal{R}$ ought to believe ϕ iff (ϕ is true and ϕ is truly believable by $\mathcal{R}$)[8].

The norm (Tc) partially avoids the original problem. Since a blindspot is not truly believable by $\mathcal{R}$, it does not follow from (Tc) alone that $\mathcal{R}$ ought to believe a blindspot. But it follows from (Tc) and some reasonable assumptions that $\mathcal{R}$ ought to believe a blindspot with the form of (bs2) (Bykvist and Hattiangadi, 2013, p. 110). If $\mathcal{R}$ does not believe (bs2), ϕ is true and truly believable by $\mathcal{R}$ and it follows from (Tc) that $\mathcal{R}$ ought to believe ϕ. Also, ψ is true and truly believable by $\mathcal{R}$ and it follows from (Tc) that $\mathcal{R}$ ought to believe ψ. Then $\mathcal{R}$ ought to believe ϕ and $\mathcal{R}$ ought to believe ψ. If 'ought' aggregates over conjunctions, then $\mathcal{R}$ ought to (believe ϕ and believe ψ). If 'belief' also aggregates over conjunctions, then $\mathcal{R}$ ought to believe (bs2). It is usually accepted that 'ought' and 'belief' aggregate over conjunctions[9]. Therefore, it follows from (Tc) and some reasonable premises that $\mathcal{R}$ ought to believe (bs2). In this case, we may reinstate the original argument about blindspots from (T) to (Tc) substituting (bs1) for (bs2).

Finally, the original problem may be avoided by modifying (T) as following:

(Td) $\mathcal{R}$ may believe ϕ iff (ϕ is true and ϕ is truly believable by $\mathcal{R}$).

According to (Td), it is not the case that $\mathcal{R}$ may believe either (bs1) or (bs2) because both are not truly believable by $\mathcal{R}$. Also, we cannot reinstate the argument from (Tc)

[8] This norm was suggested by Luis Rosa in personal conversation. By 'ϕ is truly believable by $\mathcal{R}$', I mean that it is possible that (ϕ is true and $\mathcal{R}$ believes ϕ).

[9] In the standard modal interpretation, 'ought' and 'belief' are box-like and the box aggregates over conjunctions. In our framework, under the condition that $\phi \wedge \psi$ is a logical consequence of ϕ and ψ and the logic in question has some essential features (e.g. cut), if a maximally rational reasoner $\mathcal{R}$ ought to believe (and believes) ϕ and ought to believe (and believes) ψ, then $\mathcal{R}$ also ought to believe (and believes) the logical consequence $\phi \wedge \psi$.

to (Td) because it is usually not accepted that 'may' aggregates over conjunctions. Then (Td) is the best (bi)conditional so far. The problem here is that (Td) underdetermines the description of a maximally epistemically rational reasoner: (Td) says what the reasoner may believe, but it does not say anything about what it ought to believe and, consequently, believes. This problem is revealed by the fact that (Td) may be fulfilled when a reasoner does not have any beliefs at all.

3.1.1 Truth-goal as the maximization of function g

In order to avoid these problems, the fulfillment of the truth-goal is better modeled as the maximization of a function g which evaluates sets of beliefs. In order to model the truth-goal, a function $g : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ must have the following features, where t and f are respectively the number of true and false beliefs of a reasoner[10]:

(i) $g(t, f)$ varies directly as t;

(ii) $g(t, f)$ varies inversely as f;

(iii) g has a supremum, but not a maximum[11];

(iv) g has a infimum, but not a minimum[12].

Requirement (i) models the positive part of the truth-goal: maximizing true beliefs. When I say that $g(t, f)$ varies directly as t, I mean that if $t' > t$, then $g(t', f) > g(t, f)$. Then this requirement states that having more true beliefs always increases the value of function g. Because of requirement (i), functions which measure the proportion of true beliefs, e.g. $g(t, f) = \frac{t}{t+f}$, are not adequate for modelling the truth-goal. For those functions, if $f = 0$, then $g(t, f) = 1$ for all $t > 0$. Other functions which do not fulfil this requirement are $g(t, f) = \frac{t}{f}$ and $g(t, f) = \frac{t-f}{t+f}$.

[10] Note that a reasoner $\mathcal{R}$ defines a set of beliefs and that a set of beliefs defines the values of t and f. Then I will freely use $g(\mathcal{R})$, $g(B)$ (where B is a set of beliefs), and $g(t, f)$.

[11] The supremum of a function g is the least upper bound of the image of g, defined as a quantity s such that no member in the image of g exceeds s, but if ϵ is any positive quantity there is a member in the image of g that exceeds $s - \epsilon$. The maximum of a function is the largest member in its image.

[12] The infimum of a function g is the greatest lower bound of the image of g, defined as a quantity s such that no member in the image of g is less than s, but if ϵ is any positive quantity there is a member in the image of g that is less than $s + \epsilon$. The minimum of a function is the smallest member in its image.

Requirement (ii) models the negative part of the truth-goal: minimizing false beliefs. When I say that $g(t, f)$ varies inversely as f, I mean that if $f' > f$, then $g(t, f') < g(t, f)$. Then this requirement states that having more false beliefs always decreases the value of function g. Because of requirement (ii), functions which measure the difference between the proportion of true and false beliefs, e.g. $g(t, f) = \frac{t-f}{t+f}$, are not adequate for modeling the truth-goal. For those functions, if $t = 0$, then $g(t, f) = -1$ for all $f > 0$. Other functions which do not fulfill this requirement are $g(t, f) = \frac{t}{f}$ and $g(t, f) = \frac{t}{t+f}$.

It follows from requirements (i) and (ii) that g is defined for all t and f. This is a good result. If the truth-goal is a goal for all reasoners, then, for every reasoner $\mathcal{R}$, we need to be able to evaluate $\mathcal{R}$ towards the truth-goal. But we are only able to use function g to evaluate every $\mathcal{R}$ towards the truth-goal if g is defined for every $\mathcal{R}$ (i.e. for all t and f).

Requirement (iii) models the fact that function g is to be maximized by some reasoners. In order to be maximized, function g must have an upper bound. But requirement (i) has the consequence that g cannot have a maximum value (if function g had a maximum value, it would not vary directly as t)[13]. Then the upper bound of function g must be a supremum, but not a maximum. Because of requirement (iii), functions which measure the difference between true and false beliefs, such as $g(t, f) = t - f$, are not adequate for modeling the truth-goal. Functions do not have an upper bound. Another function which does not have an upper bound is $g(t, f) = \frac{t}{f}$. Functions $g(t, f) = \frac{t}{t+f}$ and $g(t, f) = \frac{t-f}{t+f}$ have an upper bound, but it is a maximum.

Requirement (iv) is needed for stating the minimum bounds of epistemic rationality (see section 3.2). The idea is that a minimally epistemically rational reasoner does not minimize function g. In order to be minimized, function g must have a lower bound. The reason for the lower bound of function g being an infimum but not a minimum parallels those for its upper bound being a supremum but not a maximum: having a minimum contradicts requirement (ii). Function $g(t, f) = t - f$ does not have a lower bound. Functions $g(t, f) = \frac{t}{f}$, $g(t, f) = \frac{t}{t+f}$, and $g(t, f) = \frac{t-f}{t+f}$ have a lower bound, but it is a minimum.

[13] Suppose that $g(t, f) = max(g)$ for some t and f. Now consider $t' > t$. Then either $g(t', f) > max(g)$, what is a contradiction, or $g(t', f) \leq g(t, f)$ and g does not fulfill requirement (i).

Function	Req (i)	Req (ii)	Req (iii)	Req (iv)
$g(t,f) = t/f$				
$g(t,f) = t - f$	X	X		
$g(t,f) = t/(t+f)$				
$g(t,f) = (t-f)/(t+f)$				
$g(t,f) = (t-f)/(t+f+c)$	X	X	X	X

Table 3.1: The behaviour of some functions regarding requirements (i)−(iv). The 'X's indicate which requirements the functions fulfill. In the last function, $c > 0$ is a constant.

There exist several functions which fulfill the requirements (i)−(iv) and I don't see any reason for preferring one of those functions over the other. For simplicity, I will deal with the function $g(t,f) = \frac{t-f}{t+f+c}$ (where $c > 0$ is a constant)[14]. The function $g(t,f) = \frac{t-f}{t+f+c}$ fulfills requirement (i) because if $t' > t$, then $\frac{t'-f}{t'+f+c} > \frac{t-f}{t+f+c}$ (see proof C.3). This function fulfills requirement (ii) because if $f' > f$, then $\frac{t-f'}{t+f'+c} < \frac{t-f}{t+f+c}$ (see proof C.4). This function is defined for all values of t and f because a division is only undefined when the divisor is equal to zero, but, since $t \geq 0$, $f \geq 0$, and $c > 0$, this never happens for this function. Finally, this function also fulfills requirements (iii) and (iv): its supremum is equal to 1 and its infimum is equal to -1 (see figure 3.1). In the following, 'function g' refers to the function $g(t,f) = \frac{t-f}{t+f+c}$, but all the arguments should hold for all functions which fulfill requirements (i)−(iv). Also, when I talk about maximizing (minimizing) function g, I am always talking about supremum (infimum), not about maximum (minimum) value.

Function g may be used for two-valued or infinitely-many valued logics and for qualitative or quantitative notions of beliefs. Let B be the set of beliefs of a reasoner, $v(\phi)$ be a function which returns the truth-value of ϕ, and $b(\phi)$ be a function which returns the belief-value of ϕ. Then the values of t and f are the following[15]

[14] The constant c defines the 'sensitivity' of the function $g(t,f) = (t-f)/(t+f+c)$: the smaller the c, the faster the function changes. In other words, the smaller the c, the bigger the prize for having more true beliefs and the bigger the penalty for having more false beliefs. As I see, the optimal value for c is an empirical question. For simplicity, I will use a value of c which is arbitrarily close to 0.

[15] The result of equation 3.1 for $v(\phi) = 0$ and $b(\phi) = 0$ is $t = 1$, but some may think that this result should be $t = 0$. I think that the result $t = 1$ is reasonable for two reasons. First, this result conforms the norm '$\mathcal{R}$ ought not to believe a false ϕ'. The second reason has to do with accuracy: having $b(\phi) = .99$ when $v(\phi) = .99$ is the most possibly accurate degree of belief and should receive the maximum evaluation. This is also the case for $v(\phi) = 0$ and $b(\phi) = 0$. For similar reasons, I think that the result of equation 3.2 for $v(\phi) = 1$ and $b(\phi) = 0$ is reasoble.

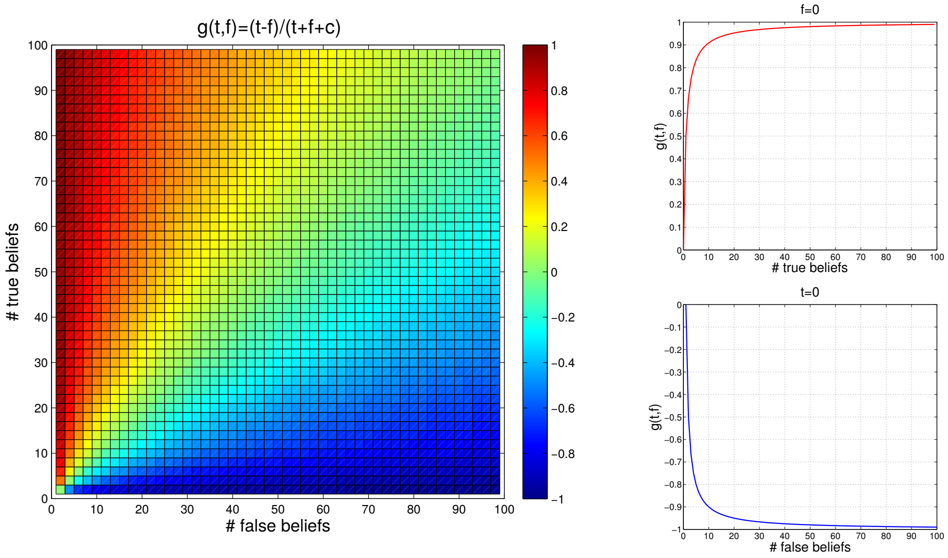


Figure 3.1: The value of $g(t, f) = \frac{t-f}{t+f+c}$ for a number of true/false belief from 0 to 100 (left). Red for $g(t, f) = 1$, green for $g(t, f) = 0$, and blue for $g(t, f) = -1$. In the right, the values of $g(t, f)$ for $f = 0$ (red) and $t = 0$ (blue). The value of $g(t, f)$ tends to 1 when t increases and tends to -1 when f increases.

$$t = \sum_{\phi \in B} 1 - |v(\phi) - b(\phi)| \tag{3.1}$$

$$f = \sum_{\phi \in B} |v(\phi) - b(\phi)| \tag{3.2}$$

In a two-valued logic, $v(\phi)$ has only two possible values: 0 when ϕ is false and 1 when ϕ is true. In an infinitely-many valued logic, $v(\phi)$ has continuously many values between 0 and 1: 0 when ϕ is absolutely false, 1 when ϕ is absolutely true, and the values between 0 and 1 meaning different degrees of truth. In a qualitative notion of beliefs, $b(\phi)$ has only two possible values: 0 when the reasoner doesn't believe ϕ and 1 when the reasoner believes ϕ. In a quantitative notion of beliefs, $b(\phi)$ has continuously many values between 0 and 1: 0 when the reasoner has absolute certainty that ϕ is false, 1 when the reasoner has absolute certainty that ϕ is true, and the values between 0 and 1 meaning degrees of certainty. In all cases, equations 3.1 and 3.2 return the values of t and f.

3.2 Maximum epistemic rationality

In chapter 1, I have defined a maximally rational reasoner as believing all and only what it ought to believe and only what it may believe (def. 1.1.6), a rational reasoner as believing only what it may believe (def. 1.1.7), and an irrational reasoner as believing something that it may-not believe (def. 1.1.8). In our framework, these definitions entail that a maximally rational reasoner believes all and only the logical consequences of its explicit beliefs, a rational reasoner has a non-trivial set of beliefs, and an irrational reasoner has a trivial set of beliefs (proof C.1). In this context, consider the following definitions:

Definition 3.2.1 (Maximum epistemic rationality). A reasoner $\mathcal{R}$ is a maximally epistemically rational reasoner iff:

(i) $\mathcal{R}$ is maximally rational;

(ii) $\mathcal{R}$ maximizes function g.

Similarly, a minimally epistemically rational reasoner is a rational reasoner which does not minimize function g and an epistemically irrational reasoner is either an irrational reasoner or minimizes function g.

There exist two putative issues in the implementation of these definitions. First, the number of (true, false) beliefs of a reasoner and, consequently, its value for function g vary depending on which notion of beliefs is used. For example, if a reasoner holds a certain number of occurrent beliefs, it probably hold much more dispositional beliefs. Second, function g is not maximized for any finite number of t and f and is not defined for infinite numbers of t or f. Luckily, these two issues are avoided by the same move.

In chapter 1, I have proposed a model of a reasoner which allow us to measure precisely the number of (true, false) beliefs given different notions of beliefs. In the model, a reasoner is composed of a language ($\mathcal{L}$), an input (`INPUT`), a knowledge base (`KB`), and a pattern of inference (π), where `INPUT` is a set of sentences in $\mathcal{L}$ which models the input of information for the reasoner, `KB` is a set of sentences in $\mathcal{L}$ which models the memory of the reasoner, and π is a function for updating `KB` which models the pattern of inference of the reasoner (def. 1.1.1). Consider the notion of an explicit belief. Informally, a reasoner explicitly believes the sentences stored in its memory. In the model, a reasoner explicitly believes the sentences in KB (def. 1.2.3). Consider the notion of an implicit belief. Informally, a reasoner implicitly believes the logical consequences of its explicit beliefs. In the model, a reasoner implicitly believes the logical consequence of KB (def. 1.2.1). Consider the notion of an accessible belief. Informally, a reasoner has the accessible belief that ϕ iff the reasoner may explicitly believe ϕ after some amount of reasoning. In the model, the accessible beliefs of a reasoner are the sentences in the output of π for some input i from INPUT and KB (def. 1.2.5).

The issue is that, for all of those notions of beliefs, the number of (true, false) beliefs of a reasoner is either finite or infinite, but function g is not maximized for finite numbers of t and f and is not defined for infinite numbers of t or f. Function g is not maximized for finite numbers of t and f because, for all sets of beliefs B with finite t and f, there exists a set of beliefs B' with finite $t+1$ and f, such that $g(B') > g(B)$. This is the case because function g varies directly as t (see note 13). On the other hand, function g is not defined for infinite t or f. Then it seems that function g cannot be maximized by any reasoner, what would defeat the purpose of the model. We may try to avoid this problem by adding some conditions to function g: for example, the condition 'if t is infinite and f is finite, then $g(t, f) = sup(g) = 1$'. This would partially fix the problem, but, in addition to its ad hoc flavor, it is unclear what to do when both t and f are infinite. We may choose a value (for example, $g(t, f) = 0$), but any choice is inadequate. Suppose that two reasoners $\mathcal{R}$ and $\mathcal{R}'$ have infinitely many true and false beliefs, that they have the same false beliefs and that, for every true belief that $\mathcal{R}'$ has, $\mathcal{R}$ has the same true belief and has an additional true belief. In this case, it is inadequate to say that $g(\mathcal{R}) = g(\mathcal{R}') = x$ because $\mathcal{R}$ seems to be in a better epistemic position than $\mathcal{R}'$.

In chapters 1, I have proposed the notion of stable beliefs. The pattern of inference of a reasoner is modeled as a function π which updates the set of (explicit) beliefs of the reasoner. A fact about the pattern of inference of a reasoner is that the reasoner may perform different inferences from the same premises. In the model, this fact is expressed using a function π which has a numeric parameter (integer) in addition to the parameters for $\texttt{INPUT}$ and $\texttt{KB}$. In this context, $\pi(\texttt{INPUT}, \texttt{KB}, 1)$ models an inference from $\texttt{INPUT}$ and $\texttt{KB}$, $\pi(\texttt{INPUT}, \texttt{KB}, 2)$ models another inference from $\texttt{INPUT}$ and $\texttt{KB}$, etc. Then function π determines a reasoning sequence $\texttt{KB}_0, \texttt{KB}_1, \ldots, \texttt{KB}_i, \ldots$, where $\texttt{KB}_0$ is the initial set of explicit beliefs and $\texttt{KB}_{i+1} = \pi(\texttt{INPUT}, \texttt{KB}_i, i+1)$. Supposing that the numeric parameter models some order of preference, the reasoning sequence of a reasoner models how the reasoner would reason from the available information if it had enough cognitive resources (e.g. time for reasoning). The stable beliefs of a reasoner are the beliefs in the set $\texttt{KB}_\omega = \bigcup_i \bigcap_{j \geq i} \texttt{KB}_j$ (the explicit beliefs at the limit of the reasoning sequence).

The notion of a reasoning sequence also allows us to make a measure for the value of function g of a reasoner $\mathcal{R}$ with reasoning sequence $\texttt{KB}_0, \texttt{KB}_1, \ldots, \texttt{KB}_i, \ldots$ as the following:

$$g(\mathcal{R}) = \lim_{i \to \infty} g(\texttt{KB}_i) \tag{3.3}$$

Measuring function g as the limit of its value when a reasoner advances in a reasoning sequence avoids the problem because this value may be maximized (and minimized)[16]. First, the value of function g is defined when both t and f approaches a finite number. In this case, the value of function g is obtained by simply substituting these values in the equation. For example, if, when i approaches infinity, both t and f approaches the same x, $\lim_{i \to \infty} g(\texttt{KB}_i) = 0$. Second, the value of function g is defined when the value of either t or f, but not both, approaches infinity. If t approaches infinity while f approaches a finite number, $\lim_{i \to \infty} g(\texttt{KB}_i) = 1$. In this case, function g is maximized. If f approaches infinity while t approaches a finite number, $\lim_{i \to \infty} g(\texttt{KB}_i) = -1$. In this case, function g is minimized. The value of $\lim_{i \to \infty} g(\texttt{KB}_i)$ may also be defined when both t and f approach infinity. In this case, the value of function g depends on how t and f grow.

In the following, 'function g model' refers to the interpretation of equation 3.3. In the

[16] Formally, I am defining another function g' from g. I am calling this function g for simplicity. There may exist some problematic cases for this interpretation. For example, suppose that a reasoner, at each stage of a reasoning sequence, adds 2 true beliefs and deletes one of the oldest belief in the set. In this case, it is unclear to me whether the reasoner has infinitely many or zero true beliefs at the limit of reasoning. To fix to this case, the number of t and f in equations 3.1 and 3.2 must be, respectively, the number of stable true and false beliefs at that stage of reasoning (in the example, 0). In other words, $t = \sum_{\phi \in \texttt{KB}_i \cap \texttt{KB}_\omega} 1 - |v(\phi) - b(\phi)|$ and $f = \sum_{\phi \in \texttt{KB}_i \cap \texttt{KB}_\omega} |v(\phi) - b(\phi)|$. This problem and its solutions were both suggested by Bernard Molyneux in personal conversation.

function g model, our situation in relation to the truth-goal parallels that of Achilles in relation to the tortoise in Zeno's paradox: no matter how fast we advance, the goal is always further ahead (the truth is out there).

3.2.1 Features of the function g model

The function g model is the best model for stating the maximum and minimum bounds of epistemic rationality for finite reasoners because this model:

1. explains how finite reasoners may achieve maximum epistemic rationality;

2. deals well with infinite sets of beliefs;

3. does not have problems with blindspots.

Informally, a finite reasoner is a reasoner with cognitive limitations, such as being able to store only a finite amount of information in its memory. In the model, a finite reasoner is defined as, among other things, holding at most finitely many explicit beliefs (def. 1.1.2). A finite reasoner may hold finitely many explicit beliefs at every stage of a reasoning sequence and, nevertheless, hold infinitely many stable beliefs. For example, if a finite reasoner starts reasoning from the unique explicit belief that the number of planets is equal to 8, and conclude that that this number is smaller than 9, and then conclude that this number is smaller than 10, and so on, the reasoner would hold infinitely many (true) stable beliefs while holding finitely many explicit beliefs at every stage of the reasoning sequence. In this case, the finite reasoner maximizes function g. In chapter 1, I have shown that a finite reasoner may be maximally rational (section 1.4). Therefore, the model explains how a finite reasoner may achieve maximum epistemic rationality[17].

In the literature, it is often cited the problem of comparing sets of beliefs with infinitely many true and false beliefs (see Treanor, 2013, p. 580). But the function g model deals with this comparison nicely. For example, if t and f grow at the same rate, then $\lim_{i\to\infty} g(\mathtt{KB}_i) = 0$. If t grows twice as fast as f, then $\lim_{i\to\infty} g(\mathtt{KB}_i) = \frac{1}{3}$. If f grows twice as

[17] The reasoner in the example would not be maximally epistemically rational because it would not be maximally rational. A maximally rational reasoner believes all the logical consequences of its explicit beliefs, what is not the case in the example.

fast as t, then $\lim_{i\to\infty} g(\mathtt{KB}_i) = -\frac{1}{3}$. There exist cases in which, when both t and f go to infinity, function g is maximized (or minimized). For example, consider a reasoner $\mathcal{R}$ with reasoning sequence such that, in the first step, the value of f is 1 and the value of t is such that the value of g is $1 - \frac{1}{2}$. In each subsequent step, the value of f increases by 1 and the value of t is such that the value of g is $1 - \frac{1}{4}$, $1 - \frac{1}{8}$, and so on. Then both t and f approaches infinity and $g(\mathcal{R}) = 1$[18]. Similar example may be constructed for $g(\mathcal{R}) = -1$. There also exist cases in which the value of g oscilates not converging to any value and $g(\mathcal{R})$ is undefined. In those cases, the reasoner is not maximally epistemically rational because it does not maximize function g (it may be minimally epistemically rational or epistemically irrational depending on whether it is rational or not), which seems to be the right result. Therefore, the function g model deals well with infinite sets of belief.

In the function g model, a reasoner ought to maximize function g in order to be maximally epistemically rational. The obligation of maximizing function g, however, does not generate obligations of believing or disbelieving individual sentences (e.g. blindspots). A reasoner may maximize function g at the limit of a reasoning sequence by lacking beliefs about blindspots – i.e. believing neither the blindspot or its negation. As a consequence, a reasoner may be maximally epistemically rational by lacking beliefs about blindspots. A reasoner may maximize function g at the limit of a reasoning sequence by having beliefs about blindspots – either believing the blindspot or its negation. As a consequence, a reasoner may be maximally epistemically rational while having beliefs about blindspots[19]. In this context, the problem of blidspots does not even arise in this this framework.

The case of (bs2) is not a case of denying that 'ought' aggregates over conjunctions, for two reasons. First, depending on how $\mathcal{R}$ updates its beliefs, it will not be the case that $\mathcal{R}$ ought* to believe ϕ and $\mathcal{R}$ ought* to believe ψ 'at the same time'. If $\mathcal{R}$ starts believing ϕ, then ψ is false, and $\mathcal{R}$ ought* not to believe ψ. If $\mathcal{R}$ starts believing ψ, ϕ is still true, but believing ϕ falsifies the belief that ψ and $\mathcal{R}$ ought* not to believe ϕ while believing ψ ($g(B + \phi + \psi) < g(B + \psi)$).

[18] This example was suggested by Hanti Lin in personal conversation.

[19] Although those beliefs may decrease the value of g, this decrease may not have an effect at the limit of the reasoning sequence.

3.3 Discussion

Following Joyce (1998) and Oddie (1997), a number of philosophers have proposed arguments which use principles of utility theory for motivating requirements of (epistemic) rationality (see Fitelson and Easwaran, 2015; Pettigrew, 2015a,b; Easwaran, 2013; Leitgeb and Pettigrew, 2010a,b; Greaves and Wallace, 2006). The resulting field has come to be known as 'epistemic utility theory' (EUT). The investigations in EUT usually employ a strategy which may be summarized in three steps (Fitelson and Easwaran, 2015, p. 11):

1. The definition of the ideal, perfect, or vindicated set of beliefs.
2. The definition of a measure of epistemic disutility as the inaccuracy ('distance') of a set of beliefs in relation to the ideal, perfect, or vindicated set of beliefs.
3. The use of the measure of epistemic disutility and some principles of utility theory in an argument for motivating requirements of (epistemic) rationality.

The first step is to define the ideal, perfect, or vindicated set of beliefs for a given situation. In EUT, the definition of such a set is usually taken as straightforward and does not receive much attention. For example, Fitelson and Easwaran (2015, p. 12) state:

> Step 1 is straightforward. It is clear what it means for a set $\mathbf{B}$ of this type to be perfectly accurate/vindicated at a world w. The vindicated set $\mathbf{B}_w$ is given by:
>
> $$\mathbf{B}_w \text{ contains } B(p)\ [D(p)] \text{ just in case } p \text{ is true [false] at } w,$$

where $B(p)$ means 'believe p' and $D(p)$ means 'disbelieve p' or 'believe not-p'. Similar definition is found in Pettigrew (2015a):

> If a proposition is true in a situation, I claim, the ideal credence for an agent in that situation is the maximal credence, which is represented as 1. On the other hand, if a proposition is false, the ideal credence is the minimal credence, which is represented as 0 (Pettigrew, 2015a, p. 3).

Then the ideal, perfect, or vindicated set of beliefs is usually defined as containing a belief of value 1 ([full] belief) for all and only the true sentences and a belief of value 0 ([full] disbelief) for all and only the false sentences.

The second step is to define a measure of epistemic disutility (inaccuracy). In EUT, inaccuracy is usually measured as the 'distance' between a set of beliefs and the ideal, perfect, or vindicated set of beliefs[20]. There exist different ways one could measure the 'distance' between two sets of beliefs. A naïve measure of the 'distance' between two sets B and B' is the number of beliefs they disagree (Fitelson and Easwaran, 2015, p. 12):

$$\sum |b(\phi) - b'(\phi)|,$$

where $b(\phi)$ is the value of the belief on ϕ (e.g. 1 for [full] belief and 0 for [full] disbelief).

In EUT, philosophers usually attempt to identify the essential properties of an adequate measure of epistemic disutility (see Joyce, 1998; Maher, 2002; Leitgeb and Pettigrew, 2010a; Pettigrew, 2015a). The naïve measure above usually does not fulfill some of the essential properties, as, for example, weak convexity (the idea that if two different sets of beliefs are equally inaccurate, their equal mixture must be less inaccurate than either). Usually, the essential properties are fulfilled when the measure has the form of a Brier Score, as, for example (see Pettigrew, 2015a, p. 33):

$$\sum |b(\phi) - b'(\phi)|^2.$$

The third step is to use the measure of epistemic disutility and some principles of utility theory in an argument for supporting requirements of (epistemic) rationality. Which principles of utility theory are used in the argument depends on the exact nature of the argument. Joyce (1998), for example, argues that a rational set of belief must fulfill Probabilism because sets of beliefs which do not fulfill this requirement are always dominated by sets of belief which fulfill this requirement. Easwaran (2013) argues that a procedure for updating beliefs must fulfill Conditionalization because sets of belief resulting from procedures which fulfill this requirement minimize expected epistemic disutility in comparision to other update procedures. Pettigrew (2015a) argues for Probabilism and Conditional-

[20] Not all arguments in EUT use the ideal, perfect, or vindicated set of beliefs as the parameter of (in)accuracy. Some arguments measure inaccuracy directly in relation to the truth-value of beliefs (Pettigrew, 2015a; Easwaran, 2013; Leitgeb and Pettigrew, 2010a,b; Greaves and Wallace, 2006; Joyce, 1998).

ization from the fact that resulting sets of beliefs which do not fulfill these requirements are always dominated by resulting sets of belief which fulfill those requirements[21].

The fact that most arguments in EUT are intended to support Bayesian requirements of (epistemic) rationality (e.g. Probabilism and Conditionalization) has consequences. These arguments use a notion of degrees of beliefs (credences) instead of plain full beliefs. These arguments also use principles of utility theory which consider situations different from the actual (e.g. dominance, minimization of *expected* inaccuracy, etc).Another consequence is the use of subjective notions of both possibility and probability.

EUT has several intersections with the function g model. However, these frameworks have formal differences and, more importantly, differ in their goals. Among the formal differences, I would list:

1. While EUT usually measures inaccuracy in relation to the ideal, perfect, or vindicated set of beliefs, the function g model measures inaccuracy directly in relation to the truth-value of beliefs (the value of f);

2. In addition to measuring the correctness of a set of beliefs (inaccuracy, the value of f), the function g model measures its completeness (the value of t);

3. The function g model, but not EUT, applies a normalizer to the overall results of correctness and completeness (the function g itself);

4. While EUT considers the epistemic disutility of sets of beliefs in different situations, the function g model only considers the actual situation.

As I see, these formal differences are direct consequences of EUT and the function g model having different goals. The difference in goals may be expressed as following:

- Whereas EUT is concerned with (epistemic) rationality, the function g model is concerned with the maximum and minimum bounds of epistemic rationality.

[21] Probabilism and Conditionalization are Bayesian requirements of (epistemic) rationality (see ch. 5).

The first formal difference is that the function g model measures inaccuracy directly in relation to the truth-value of beliefs instead of using the ideal, perfect, or vindicated set of beliefs. This must be the case because the ideal, perfect, or vindicated set of beliefs plays the role of maximum (epistemic) rationality in EUT. This set is always maximally (epistemically) rational in the sense that it is always evaluated positively no matter which measure of inaccuracy and principle of utility theory are used. In this context, it would be circular to measure inaccuracy in relation to the ideal, perfect, or vindicated set of beliefs when the goal is to establish the maximum bounds of epistemic rationality.

I think that there exist advantages in measuring inaccuracy directly in relation to the truth-value of beliefs. Using the ideal, perfect, or vindicated set of beliefs as the parameter for measuring the inaccuracy presupposes the existence of such a set, but this may not be the case in situations with blindspots. If the ideal, perfect, or vindicated set of beliefs contains a belief of value 1 for all and only the true sentences and a belief of value 0 for all and only the false sentences, the value of the belief in the sentence $\phi =$ 'the ideal, perfect, or vindicated set of beliefs does not contain a belief of value 1 for ϕ' is not defined[22]. This problem resembles the problem with norm (T) discussed in section 3.1.

In fact, some arguments in EUT measure inaccuracy directly in relation to the truth-value of beliefs (e.g. Pettigrew, 2015a; Easwaran, 2013; Leitgeb and Pettigrew, 2010a,b; Greaves and Wallace, 2006; Joyce, 1998). As I see, this strategy has been abandoned in more recent papers (e.g. Fitelson and Easwaran, 2015; Pettigrew, 2015a) because of the first-person nature of the notion of (epistemic) rationality. The notion of (epistemic) rationality is sensible to first-person features such as, for example, which situations are possible given the reasoner's current information. Measuring epistemic utility directly to the truth-value of beliefs usually disregards these features. For example, the actual situation may be impossible given current (misleading) information. The notion of maximum epistemic rationality, on the other hand, is related to the best-case epistemic situation for

[22] The idea is that, in situations of blindspot, the ideal, perfect, or vindicated set of beliefs may not exist and, as a consequence, cannot be used in evaluating the innacuracy of nonideal reasoners. Using truth-values in evaluating the innacuracy of nonideal reasoners ensure that the evaluation may always be done. For this reason, Pettigrew (2015a, p. 4) restricts his framework to situations without blindspots.

reasoners in general and is not sensible to first-person features of individual reasoners[23].

The second formal difference is a consequence of the first. Since the ideal, perfect, or vindicated set of beliefs is 'complete' in the sense that, for every relevant sentence, that set contains a degree of belief for that sentence, the 'distance' between a set of beliefs and the ideal, perfect, or vindicated set of beliefs always increases when the non-ideal set lacks a value for the belief in some sentence[24]. When we measure epistemic utility in relation to the truth-value of beliefs, on the other hand, we need to consider not only the correctness of the set of beliefs (the value of f), but also its completeness (the value of t). If this is not done, a reasoner would be able to minimize epistemic disutility simply by not having any beliefs. This is the same problem with norms (Ta), (Tb), and (Td) discussed in section 3.1.

The third formal difference (lack of normalizer) renders EUT two limitations in investigating the maximum and minimum bounds of (epistemic) rationality[25]

(a) EUT is able to (easily) distinguish between (epistemic) rationality and irrationality, but not between maximum and minimum (epistemic) rationality, and irrationality.

(b) EUT has problems in dealing with infinite sets of belief.

The third formal difference is a consequence of the different goals in the sense that it may be thought that (a) and (b) are limitations for dealing with maximum epistemic rationality, but not for dealing with regular cases of (epistemic) rationality (i.e. (epistemic) rationality for finite reasoners).

About (a), the lack of a normalizer for the inaccuracy measures renders EUT unable to (easily) provide fixed notions of maximum and minimum (epistemic) rationality, and irrationality[26]. In EUT, the global result for the inaccuracy of a set of beliefs is in

[23] A best-case epistemic situation may be, for example, one in which the reasoner has all the relevant information and no misleading information.

[24] It is usually supposed that the reasoners to be evaluated are opinionated, in the sense of having (degrees of) beliefs for every sentence/proposition considered.

[25] Here, 'normalizer' means a function which rescale values in a larger (often unbounded) interval to values in a smaller and more tractable (bounded) interval.

[26] There may exist clever ways of provide these notions. Here, I mean that it is not possible to do it in the most natural way: maximum and minimum rationality being related with the maximum and minimum values (not necessarily respectively) in the interval.

$[0, +\infty)$ and, therefore, has only one defined bound (0). As a consequence, EUT is able to distinguish between the ideal, perfect, or vindicated set of beliefs (degree 0 of inaccuracy, maximum epistemic rationality) and the other sets of beliefs (degree > 0). EUT is also able to compare two sets of beliefs and conclude that one is (epistemically) rational whereas the other is irrational (e.g. that one is less inaccurate than the other)[27]. On the other hand, the result of function g (in the limit interpretation) is between two integers $[x, y]$ (e.g. $-1, 1$) and, therefore, has two defined bounds. In this context, it is possible to define maximum epistemic rationality using the higher bound (e.g. 1), irrationality using the lower bound (e.g. -1), and minimum epistemic rationality using the rest of the interval (e.g. the interval $(-1, 1)$).

About (b), the lack of a normalizer for the inaccuracy of a set of beliefs renders EUT unable to deal properly with infinite sets of beliefs. This is the case because the global inaccuracy of a set of beliefs is in $[0, +\infty)$, but the result for infinite sets of beliefs often collapse to $+\infty$. The restriction to finite sets of beliefs is stressed in Leitgeb and Pettigrew (2010a), Pettigrew (2015a), and (partially) in Easwaran (2015). For example:

> Indeed, the only restriction we impose is that an agent's opinion set is finite (Pettigrew, 2015a, p. 14).

This restriction usually does not receive an epistemic interpretation:

> There is nothing particularly philosophical about our decision to stick to the case of finitely many worlds in this article; we simply assume this to be so and postpone the discussion of the infinite case to another time (Leitgeb and Pettigrew, 2010a, p. 209)[28].

The restriction to finite sets of belief is a problem in investigating maximum epistemic rationality because a maximally (epistemically) rational reasoner has infinitely many beliefs. However, I think that this may also be a problem in dealing with regular cases of

[27] The basis for comparison may change depending on which principle of utility theory is used in the argument (e.g. having smaller expected inaccuracy, having smaller inaccuracy in all relevant situation, etc). In principle, we could describe these comparisons as stating that one set of beliefs is more epistemically rational than another, but these comparisons are usually described in the literature as stating that one set is epistemically rational whereas the other is epistemically irrational (see Pettigrew, 2015b).

[28] For another example, Greaves and Wallace state that: "We will assume throughout that $\mathcal{S}$ [the set of states] is finite. This is merely for simplicity of exposition" (Greaves and Wallace, 2006, p. 611, n.1).

(epistemic) rationality. Easwaran points in that direction:

> I assume that the set of situations for each agent is finite. One might try to justify this assumption by arguing that actual agents are finite beings that only have the capacity or desire to make finitely many distinctions among ways the world could be. I happen to think that this sort of argument won't work, and that in fact the set of situations for a given agent is ordinarily infinite, but for the purposes of this paper I will restrict things to the finite case (Easwaran, 2015, p. 4).

In fact, depending of the notion of beliefs, finite reasoners may have infinitely many beliefs. This has the consequence of rendering EUT unable to deal properly not only with maximum (epistemic) rationality, but also with regular cases of (epistemic) rationality. These issues are avoided using the function g model because, in this model, the measures for a set of beliefs is between two defined boundaries (e.g. [-1,1]) for reasoners with either finite or infinite sets of beliefs.

Finally, the fourth formal difference is also a consequence of the different goals of EUT and the function g model: whereas EUT considers the inaccuracy of a set of beliefs in different situations, the function g model considers only the actual situation. Again, EUT is concerned with the first-person nature of the notion of (epistemic) rationality. In this context, an evaluation of (epistemic) rationality must take into account the situations which the reasoner cannot distinguish given the available information. The function g model is used in the investigation of the maximum bounds of epistemic rationality. The requirements for maximizing function g regarding only the actual situation are the same as (or smaller than) they would be regarding all epistemically possible situations. In principle, however, the function g model may be adapted for considering situations different from the actual. The restriction of the function g model to the actual situation is more a matter of simplicity than a bold claim in epistemology.

The conclusion is that there does not exist real discordance between EUT and the function g model. These frameworks have different goals and, consequently, use different tools. However, this fact should not preclude the cooperation between these frameworks. Some results within EUT may be implemented in the function g model. For example, through this chapter I have been using naïve measures of completeness and correctness (i.e. the number of true and false beliefs, see equations 3.1 and 3.2). However, the results of the discussion about the essential properties of this kind of measure may be applied to the function g model. For example, the function g model works perfectly using a measure with the form of a Brier Score ($t = \sum_{\phi \in B} 1 - |v(\phi) - b(\phi)|^2$, $f = \sum_{\phi \in B} |v(\phi) - b(\phi)|^2$).

Also, some features of the function g model may be implemented within of EUT. The function g model may be used (i) in defining the ideal, perfect, or vindicated set of beliefs or (ii) as providing a normalizer for the measures of inaccuracy. In the first case, the function g model would be used in defining the ideal, perfect, or vindicated set of beliefs, which may be used as a parameter for measuring inaccuracy. The advantage of this use is that the function g model does not have problems with blindspots. The disadvantages are the more complicated calculations and the fact that the model may define more than one ideal, perfect, or vindicated set of beliefs in some situations. In the second case, function g itself would be used as a normalizer for measures of inaccuracy. The advantage of this use is enabling comparisons between infinite sets of beliefs. One disadvantage would be that a reasoner needs now to be modeled not only as a set of beliefs, but also as having a pattern of inference. This renders more complicated calculations, but I think that this model is more sensible to the first-person features of the notion of (epistemic) rationality used in EUT because patterns of inference are internal features of a reasoner.

It is interesting that, when philosophers attempt to overcome some limitations of EUT, they end up proposing characteristics akin to those of the function g model. For example, Easwaran (2013) provides a framework which renders EUT able to deal with infinite sets of beliefs with three features:

1. Instead of evaluating a set of beliefs, evaluating a process of construction/updating of this set;

2. Considering the final evaluation as the limit of a succession of evaluation of the process of construction/updating of this set;

3. Using a measure which has a maximum value.

Instead of evaluating sets of beliefs, Easwaran evaluates plans for updating sets of belief given possible incoming information. The expected inaccuracy of an update is measured as the limit of the sum of the inaccuracy of the updates for each possible outcome weighted by the probability of that outcome. This measure is in the interval $[0, 1]$ because the inaccuracy of each possible update is in $[0, 1]$ and the probability distribution sums 1. The argument of Easwaran (2013) is that plans with smaller expected inaccuracy are those which may be described as applying conditionalization. The important point is that these features are exactly the features of the function g model.

Chapter 4

Computational epistemology

> The 9000 series is the most reliable computer ever made. No 9000 computer has ever made a mistake or distorted information. We are all, by any practical definition of the words, foolproof and incapable of error.
> HAL 9000, *2001: A Space Odyssey*.

In the last chapters, I have been investigating the maximum and minimum bounds of (epistemic) rationality for finite reasoners. Informally, a finite reasoner (e.g. a human) is a reasoner with cognitive limitations, such as finite perceptual input, finite memory, and finite computational capacities. The investigation has been carried out using a model of a reasoner which provides clear definitions for most notions of interest. In the model, a reasoner is composed of a language ($\mathcal{L}$), an input (`INPUT`), a knowledge base (`KB`), and a pattern of inference (π), where `INPUT` is a set of sentences in $\mathcal{L}$ which models the perceptual input of the reasoner, `KB` is a set of sentences in $\mathcal{L}$ which models the explicit beliefs of the reasoner, and π is a function for updating `KB` which models the pattern of inference of the reasoner (def. 1.1.1). The function π accepts `INPUT`, `KB`, and an integer as inputs and returns an updated `KB`. In the model, a finite reasoner is a reasoner with learnable $\mathcal{L}$, finite `INPUT`, finite `KB`, and computable π (def. 1.1.2)[1].

In chapter 1, I have investigated the maximum and minimum bounds of rationality for finite reasoners. In the literature, an ideal reasoner is often defined as a reasoner

[1] That π must be computable is a consequence of an argument in section 1.3. A function is computable if there exists an effective procedure such that for any input, it produces the value of the function in a finite time interval. For reasons expressed in chapter 1, note 8, I will leave aside considerations about $\mathcal{L}$.

which (i) believes all the logical consequences of its epistemic situation and (ii) has a nontrivial set of beliefs (def. 1.1.3), where the epistemic situation of a reasoner is the information available to the reasoner (`INPUT` and `KB`)[2]. I argue that this notion is not satisfactory for modeling maximum rationality because an ideal reasoner with features (i) and (ii) may still have all sorts of random beliefs[3]. I then propose the notion of a strictly ideal reasoner, i.e. an ideal reasoner which believes only the logical consequences of its epistemic situation. In our framework, the notion of a strictly ideal reasoner coincides with the notion of a maximally rational reasoner which believes all and only what it ought to believe and only what it may believe (proof C.1)[4]. I then investigate under which notions of beliefs a finite reasoner may be maximally rational (strictly ideal). I argue that there exist infinitely many logical theorems but a finite reasoner can only have finitely many explicit beliefs (sentences in `KB`). I argue that, for many relevant logics, there does not exist a computable function π which outputs all and only the logical consequences of an epistemic situation (accessible beliefs). I argue that, for many of those logics, there exists a computable function π which outputs all and only the logical consequences of an epistemic situation at the limit of a sequence of iterations of the function (a reasoning sequence, see section 4.1). I conclude that maximum rationality for finite reasoners must be understood in terms of the state of the set beliefs of the reasoner as it approaches the limit of its reasoning sequence (stable beliefs).

In chapter 3, I have investigated the maximum and minimum bounds of *epistemic* rationality for finite reasoners. In the literature, the notion of epistemic rationality is often understood as being related to the fulfilment of the truth-goal (the goal of believing truths and not believing falsehoods)[5]. Under the supposition that epistemic rationality

[2] Similar definitions are found in Stalnaker (2006), Giunchiglia and Giunchiglia (2001), Duc (1995), Grim (1988), Halpern and Moses (1985), and Binkley (1968). This notion is relative to a logic.

[3] As long as those beliefs do not contradict requirement (ii). In other words, the notion of an ideal reasoner has a very strong completeness requirement (closure), but only a very weak soundness requirement (nontriviality). This explanation was suggested by an anonymous reviewer.

[4] In our framework, the epistemic ought and may are defined in terms of ideal reasoners. A reasoner ought to believe ϕ iff all ideal reasoners in the same epistemic situation believe ϕ and a reasoner may believe ϕ iff there exist an ideal reasoner in the same situation which believes ϕ (def. 1.1.4 and 1.1.5).

[5] Similar claims are found in Lehrer (2000), Foley (1993), Plantinga (1993), Goldman (1986), Moser (1985), Alston (1985), BonJour (1985) Sosa (1985), and Chisholm (1982). List due to David (2001).

is a species of rationality, I investigate the proposal of maximum epistemic rationality being maximum rationality plus the fulfillment of the truth-goal. The problem with this proposal is that the conditions for the fulfillment of the truth-goal are not clear in the literature and most of the existing models have problems with blindspots[6]. I argue that the fulfilment of the truth-goal is better understood as the maximization of function g, which accepts a set of beliefs as input and returns a numeric evaluation, and discuss the properties of such a function[7]. I argue that this model does not have problems with blindspots and that a finite reasoner is only able to maximize function g at the limit of a reasoning sequence. I conclude that the notion of maximum epistemic rationality for finite reasoners must be understood in terms of the value of function g for a reasoner as it approaches the limit of a reasoning sequence (i.e. whether it reaches its supremum).

In this chapter, I argue that, if finite reasoners are only able to approach maximum (epistemic) rationality at the limit of a reasoning sequence, then considerations about the efficiency of patterns of inference are relevant to epistemology. In section 4.1, I present an argument for this conclusion. The argument exploits the facts that (i) a finite reasoner cannot reach actual maximum (epistemic) rationality and that (ii) there exists a relevant difference on how a finite reasoner may approach maximum (epistemic) rationality depending of how efficiently it reasons. I discuss some criticisms to the presented notion of efficiency and some problems to the general framework. Among the problems, I discuss the problem of interruptibility (Pollock, 1995). In section 4.2, I present more formally the research program of computational epistemology, a research program which investigates the maximum and minimum bounds of (epistemic) rationality for finite reasoners using computer simulations, takes into account differences in the efficiency of patterns of inference, and deals (indirectly) with the problem of interruptibility.

[6] A blindspot for a reasoner is a sentence which may be true but that cannot be both true and believed by the reasoner (Sorensen, 1988).

[7] Function g must vary directly as the number of true beliefs (t) and inversely as the number of false beliefs (f), what entails that it is defined for all values of t and f. Also, it must have an upper [lower] bound in the form of a supremum [infimum] but not a maximum [minimum] (see section 3.1.1). For simplicity, I work with the function $g = (t - f)/(t + f + c)$, where $c > 0$ is a constant.

4.1 Efficient reasoning

In chapter 1, I have presented a model of a reasoner which provides clear definitions for a reasoning sequence, the beliefs that a reasoner holds at the limit of a reasoning sequence (stable beliefs), and the value of function g at the limit of its reasoning sequence. The pattern of inference of a reasoner is modeled as a function π which updates the set of (explicit) beliefs of the reasoner. A fact about the pattern of inference of a reasoner is that the reasoner may perform different inferences from the same premises. In the model, this fact is expressed using a function π with a numeric parameter (integer) in addition to the parameters for INPUT and KB. In this context, $\pi(\texttt{INPUT}, \texttt{KB}, 1)$ models an inference from INPUT and KB, $\pi(\texttt{INPUT}, \texttt{KB}, 2)$ models another inference from INPUT and KB, etc. Then function π determines a reasoning sequence $\texttt{KB}_0, \texttt{KB}_1, \ldots, \texttt{KB}_i, \ldots$, where $\texttt{KB}_0 = \texttt{KB}$ is the initial set of explicit beliefs and $\texttt{KB}_{i+1} = \pi(\texttt{INPUT}, \texttt{KB}_i, i+1)$. Supposing that the integer parameter models some order of priority, the reasoning sequence of a reasoner models how the reasoner would reason from the available information if it had enough cognitive resources (e.g. time). In this context, the beliefs that a reasoner holds at the limit of its reasoning sequence (stable beliefs) are the sentences in $\texttt{KB}_\omega = \bigcup_i \bigcap_{j \geq i} \texttt{KB}_j$. The value of function g at the limit of a reasoning sequence is $\lim_{i \to \infty} g(\texttt{KB}_i)$.

The result of chapters 1 and 3 was that the notion of maximum epistemic rationality should be defined in terms of the limit of its reasoning sequence (stable beliefs):

Definition 4.1.1 (Maximum epistemic rationality). A reasoner $\mathcal{R}$ is maximally epistemically rational for a logic $\overset{\text{x}}{\models}$ only if:

(i) $\mathcal{R}$ believes all and only the logical consequences of its epistemic situation
($\texttt{INPUT} \cup \texttt{KB} \overset{\text{x}}{\models} \phi$ iff $\phi \in \texttt{KB}_\omega$);

(ii) $\mathcal{R}$ has a nontrivial set of beliefs ($\exists \phi (\phi \notin \texttt{KB}_\omega)$);

(iii) $\mathcal{R}$ maximizes function g ($\lim_{i \to \infty} g(\texttt{KB}_i) = sup(g)$).

In chapter 1 and 3, I have shown that, for many relevant logics, finite reasoners may be maximally (epistemically) rational in the sense of having a reasoning sequence $\texttt{KB}_0, \texttt{KB}_1, ..., \texttt{KB}_i, ...$ with properties (i)$-$(iii) (see sections 1.4 and 3.2). However, no $\texttt{KB}_i$

in a reasoning sequence of a finite reasoner have properties (i) or (iii). About (i), there exist infinitely many logical theorems, but the $\mathtt{KB}_i$ of a finite reasoner must be finite (see section 1.2). About (iii), function g is not maximized by a finite set of explicit beliefs, but the $\mathtt{KB}_i$ of a finite reasoner are finite (see section 3.2)[8]. This fact may be expressed as a finite reasoner being able to approach maximum (epistemic) rationality at the limit of a reasoning sequence, but not being able to reach *actual* maximum (epistemic) rationality[9].

There exist conditions which enable a finite reasoner to go further in its reasoning sequence. For example, there exist purely cognitive conditions, such as having more cognitive resources (e.g. time, memory, etc), which enable a finite reasoner to go further in a reasoning sequence. Purely cognitive conditions, however, are not directly relevant to epistemology. In the following, I argue that there exist features which enable a finite reasoner to go further in its reasoning sequence or even to get 'closer' to actual maximum (epistemic) rationality and which are relevant to epistemology. I argue that how far finite reasoners can go in a reasoning sequence depends on the efficiency of their patterns of inference, that finite reasoners with increasingly more cognitive resources can get infinitely 'closer' to actual maximum (epistemic) rationality having a polynomial pattern of inference in comparison to an exponential pattern of inference, and that this distinction is relevant to the investigation of maximum (epistemic) rationality. This is the argument:

(1) How far finite reasoners can go in a reasoning sequence depends on the efficiency of their patterns of inference.

(2) Finite reasoners with increasingly more cognitive resources can go infinitely further in a reasoning sequence having a polynomial pattern of inference in comparison to an exponential pattern of inference.

(3) Finite reasoners with increasingly more cognitive resources can get infinitely 'closer' to actual maximum (epistemic) rationality having a polynomial pattern of inference in comparison to an exponential pattern of inference.

[8] The value of $g(\mathtt{KB})$ varies directly as the number of true beliefs in $\mathtt{KB}$, then, for all finite $\mathtt{KB}_i$, there exist a $\mathtt{KB}_j$ such that $\mathtt{KB}_j$ is just like $\mathtt{KB}_i$, but $\mathtt{KB}_j$ has one more true belief. In this case, $g(\mathtt{KB}_j) > g(\mathtt{KB}_i)$.

[9] The notion of a reasoning sequence was introduced using a counterfactual phrase: 'how the reasoner would reason from its epistemic situation *if it had enough cognitive resources*'.

(4) This distinction is relevant to epistemology.

($\therefore$) The patterns of inference of maximally (epistemically) rational finite reasoners must be polynomial if this is possible.

In (1), the claim is that how far finite reasoners can go in a reasoning sequence depends on the efficiency of their patterns of inference. The point is that reasoning requires certain resources which a finite reasoner has only a finite amount of (e.g. time, memory, etc) and a finite reasoner will stop reasoning at some point of a reasoning sequence because it lacks the appropriate resources to continue. How far a finite reasoner with fixed cognitive resources can go in a reasoning sequence depends on the following features of the sequence:

(a) The relative number of inferential steps used in generating each stage of the sequence;

(b) The relative number of explicit beliefs in each stage of the sequence;

(c) The relative number, in each step of the sequence, of conclusions which are retracted in later stages of the reasoning sequence (Kelly, 1988).

Case (a) is related to time, (b) is related to memory, and (c) is related to both. I will focus on the relative number of inferential steps, which is the easiest case. An inference is a sequence of inferential steps, where an inferential step is the execution of an inference rule for a group of sentences, as, for example, concluding that q from $p \to q$ and p using modus ponens. Supposing that each inferential step takes time and that every finite reasoner has an upper bound for time (e.g. life span), executing more direct inferences allows a reasoner to go further in a reasoning sequence because otherwise it would reach its upper bound sooner[10]. Similar arguments may be done regarding cases (b) and (c).

In (2), the claim is that finite reasoners with increasingly more cognitive resources can go infinitely further in a reasoning sequence having a polynomial pattern of inference in comparison to an exponential pattern of inference. Consider a resource function $r(i)$, which measures the amount of some cognitive resource (e.g. time) which is required for

[10] For example, concluding q from $p \to q$ and p using modus ponens (one step) is a more direct inference than concluding q from a reduction to absurdity of the supposition that $\neg q$ (four).

a reasoner to reach the stage i of a reasoning sequence. Call this function $poln(i)$ if it is polynomial or $exp(i)$ if it is exponential[11]. In the following, 'polynomial pattern of inference' denotes a pattern of inference with polynomial resource function and 'exponential pattern of inference' denotes a pattern of inference with exponential resource function.

Consider the following theorem (see proof C.5):

$$\lim_{i\to\infty} \frac{exp(i)}{poln(i)} = \infty \tag{4.1}$$

Theorem 4.1 states that as one advances in a reasoning sequence, the difference in the amount of cognitive resources that $exp(i)$ and $poln(i)$ require approaches infinity. Consider a finite reasoner $\mathcal{R}$ with resource function $r(i)$ and upper bound u for some cognitive resource (e.g. time). Then $\mathcal{R}$ is able to reach stage i of a reasoning sequence iff $r(i) \leq u$. There does not exist a priori limitations of considering finite reasoners with increasingly larger upper bounds for a cognitive resource (as long as it remains finite). For any u, there exists a finite reasoner with upper bound u' such that $u' > u$. Then theorem 4.1 has the consequence that, as we consider finite reasoners with increasingly more cognitive resources, they can go infinitely further in a reasoning sequence having a polynomial pattern of inference in comparison to an exponential pattern of inference[12]:

$$\lim_{u\to\infty} \frac{max(\{i|poln(i) \leq u\})}{max(\{i|exp(i) \leq u\})} = \infty, \tag{4.2}$$

where u is the upper bound for a cognitive resource and i is a stage in a reasoning sequence.

In (3), the claim is that finite reasoners with increasingly more cognitive resources can get infinitely 'closer' to actual maximum (epistemic) rationality having a polynomial pattern of inference in comparison to an exponential pattern of inference. The conclusion of the previous paragraph is that finite reasoners with increasingly more cognitive resources can go infinitely further in its reasoning sequence having a polynomial pattern of infer-

[11] The function $r(i)$ is polynomial iff it may be represented as an expression of the kind $O(i^c)$, where c is a constant. The function is exponential iff it must be represented as an expression of the kind $O(2^{c^i})$, where $c > 0$ is a constant (see appendix B.1.1 for a survey on computational complexity theory).

[12] The expression $max(\{i|poln(i) \leq u\})$ is $poln(u)$ and $max(\{i|exp(i) \leq u\})$ is $log(u)$. Then it is the case that $\lim_{i\to\infty} \frac{poln(u)}{log(u)} = \infty$, where $log(u)$ is a function of the kind $O(log(u))$ (see proof C.6).

ence in comparison to an exponential pattern of inference. As a consequence, we may say that finite reasoners with increasingly more cognitive resources can get infinitely 'closer' to actual maximum (epistemic) rationality having a polynomial pattern of inference in comparison to an exponential pattern of inference. The idea of being 'closer' to the limit of a infinite sequence does not make sense for any finite difference of positions in the sequence, but the difference that we are considering is infinite at the limit. In this case, I think that it makes sense to say that finite reasoners with increasingly more cognitive resources can get infinitely 'closer' to actual maximum (epistemic) rationality having a polynomial pattern of inference in comparison to an exponential pattern of inference.

In (4), the claim is that this distinction is relevant to epistemology. This is the case for several reasons. The first reason is that this distinction does not seems to be merely quantitative (is infinite a number?) and, even if it was the case, it would be 'large enough' to be given special attention. The second reason is that, even if maximum (epistemic) rationality is defined using the limit of a reasoning sequence, actual maximum (epistemic) rationality is of interest to epistemology. The third and more important reason is that having more cognitive resources often enables one to be in a better epistemic position, but it is exactly when we consider finite reasoners with increasingly more cognitive resources that this distinction arises. Finally, exponential patterns of inference are often correlated with brute-force search and lack of deep understanding whereas polynomial patterns of inference are often correlated with deep understanding:

> The motivation for accepting this requirement is that exponential algorithms typically arise when we solve problems by exhaustively searching through a space of solutions, what is often called a brute-force search. Sometimes brute-force search may be avoided through a deeper understanding of a problem, which may reveal polynomial algorithms of greater utility (Sipser, 2012, p.285).

But, if this correlation is correct, to require reasoners to approach maximum (epistemic) rationality through deep understanding (if possible) is to require maximally (epistemically) rational finite reasoner to have polynomial patterns of inference (if possible).

In this context, maximum (epistemic) rationality must have the following requirement:

(iv) $\mathcal{R}$'s pattern of inference is polynomial if possible.

Requirement (iv) does not deal with all differences in efficiency, but only with the difference between having a polynomial and having an exponential pattern of inference. This is the case because the difference between polynomial and exponential patterns of inference is qualitative whereas the difference between more and less efficient patterns of inference in general may be merely quantitative[13]. The clause 'if possible' is in place because, in some cases, it is not possible for a finite reasoner with polynomial pattern of inference to approach maximum (epistemic) rationality (see appendix B.1.2).

The inclusion of requirement (iv) in the definition of maximum (epistemic) rationality may be criticized. The first objection would be that the distinction between polynomial and exponential patterns of inference does not track efficiency correctly. The problem would be that some exponential patterns of inference are, for all practical purposes, more efficient than some polynomial patterns of inference. For example, for all practical purposes, an exponential pattern of inference of the kind $O(1.00001^i)$ is (much) more efficient than a polynomial pattern of inference of the kind $O(i^{10000})$. This is true, but, as much as practical considerations are of interest of epistemology, the main concern of our investigation is the maximum bounds of (epistemic) rationality for finite reasoners. A finite reasoner may only approach maximum (epistemic) rationality at the limit of a reasoning sequence and, in this case, a polynomial pattern of inference is always more efficient than an exponential pattern of inference. Then, for our concerns, the distinction between polynomial and exponential patterns of inference tracks efficiency correctly.

Another objection would be that epistemic reasoning is supposed to assist practical reasoning, but approaching maximum (epistemic) rationality at the limit of a reasoning sequence may not assist practical reasoning at any point of the sequence (Kitcher, 1993). In other words, the objection would be that maximum (epistemic) rationality in this framework is not *interruptible* in the sense that if, at some stage of a sequence, a maximally (epistemically) rational reasoner must stop reasoning and act, it may not be reasonable to act on the (explicit) beliefs the reasoner holds at that stage (Pollock, 1995, p. 272).

[13] In the literature, this fact is often expressed as general efficiency being relative to implementation whereas the distinction between polynomial and exponential being independent of almost all choices of implementation (see appendix B.1.1).

I think that this objection is correct: the notion of maximum (epistemic) rationality defined in terms of the limit of a reasoning sequence is, in fact, insensible to considerations about interruptibility. Consider the following example:

> Let $\mathcal{R}_1$ be a reasoner that is both interruptible and d.e.-adequate [a notion 'close' to maximum rationality for a nonmonotonic logic]. Let $\mathcal{R}_2$ be just like $\mathcal{R}_1$ except that for the first million [or googolplex, etc] steps it draws conclusions purely at random, and then, after 1 million [or googolplex, etc] steps it withdraws all those randomly drawn conclusions and begins reasoning as $\mathcal{R}_1$ (Pollock, 1995, p. 146).

In the example, $\mathcal{R}_1$ and $\mathcal{R}_2$ both approach maximum (epistemic) rationality at the limit of their reasoning sequence. Nevertheless, the pattern of inference of $\mathcal{R}_2$ is not interruptible: it is not reasonable to act according to the (explicit) beliefs of $\mathcal{R}_2$ in any practical situation (because they are random). Then, if maximum (epistemic) rationality must be interruptible (as it seems to be the case), interruptibility is indeed a problem for understanding this notion in terms of the limit of a reasoning sequence.

In this context, maximum (epistemic) rationality must have the following requirement:

(v) $\mathcal{R}$'s patterns of inference are interruptible,

Requirement (v) is not satisfactory (ad hoc?) because it is very difficult to provide precise conditions of interruptibility. Pollock's use of this notion suggests that a pattern of inference is interruptible iff the *relevant* inferences are performed *quite early* in the reasoning sequence, but this is quite vague. I do not have a good theory of interruptibility and I don't think that this theory is possible with the tools that we have introduced so far. First, the problem of interruptibility is not banal. In fact, the problem of interruptibility is closely related to the frame problem, which is the problem of determining which inferences must be performed/ignored in the process of making some decision (Dennett, 1984). Second, the very notion of interruptibily seems to be relative to a class of problems because which are the relevant inferences depends on which problem we are trying to solve. In this context, instead of trying to deal with interruptibility using formal requirements about patterns of inferences, I will attempt to deal with this problem indirectly: by observing and comparing how different reasoners deal with some class of problems.

4.2 Computational epistemology

In this section, I present more formally the research program of computational epistemology (CE). CE investigates maximum (epistemic) rationality for finite reasoners taking into account both the requirements for (iv) efficiency and (v) interruptibility. In CE, the requirements for efficiency and interruptibility are dealt with by considering reasoners as agents dealing with some relevant problems. Informally, an agent is a reasoner with two extra components: a set of goals and a function which has a set of explicit beliefs (`KB`) and a set of goals as inputs and which returns an action. Which are the actions that this function may return depends on the environment in which the agents are in. An environment is composed of an initial state and a function which updates the current state of the environment given actions. A problem is an environment with two extra components: a goal-state (a state in which the problem is solved) and a function which assigns costs to solutions (e.g. time). I will explain these notions more formally in the next sections.

Roughly, an investigation in CE is carried out by choosing and describing a relevant class of problems; choosing and describing hypotheses of agents which would exhibit maximum (epistemic) rationality in dealing with problems in that class; designing and implementing computer simulations of the agents solving problems in that class; and by analyzing the data from the simulation given some parameters related with requirements (i)−(v) for maximum (epistemic) rationality. The analysis enables the compare the agents regarding maximum (epistemic) rationality.

An investigation in CE is composed of four stages:

1. The choice and formalization of a relevant class of problems.
2. The choice and formalization of hypotheses of agents which would exhibit maximum (epistemic) rationality problems in that class.
3. The designing and implementation of computer simulations of the agents solving problems in that class.
4. The analysis of the data from the simulation given parameters related with requirements (i)−(v) for maximum (epistemic) rationality.

4.2.1 Class of problems

Using the methods of CE for investigating maximum (epistemic) rationality regarding all possible problems is a strenuous or even an impossible task[14]. On the other hand, the notion of maximum (epistemic) rationality cannot be tied to few and specific problems[15]. Then an investigation in CE should be carried out in terms of a relevant class of problems. In this case, we talk about maximum (epistemic) rationality *for a class of problems*.

The choice of a class of problems follows two guidelines. First, the class of problems must be as abstract as possible in the sense of having states with as few features as possible. This facilitates the design and implementation of the problem in a computer simulation and the analysis of the data from the simulation. Second, the class of problems must have the essential features of a relevant class of real-world problems (e.g. reasoning under uncertainty). This would make results about a class of problems relevant to maximum (epistemic) rationality in general.

If the class of problems in a class of search problems[16], the class of problems may be described using the following formalization proposed by Russell and Norvig (2010, p. 66):

- The initial state of the environment, including the initial state of the agent;
- A description of the actions which are available to the agent in a given state;
- A transition function which updates the state of the environment given the current state and an action[17];
- A goal test function, which determines whether the current state is a goal state;

[14] First, it is dubious if there exists an agent which exhibits maximum (epistemic) rationality for all possible problems. In addition, the number of problems is too large for any practical testing (simulation).

[15] Suppose that the pattern of inference of an agent returns a contradiction for some specific input. That agent may exhibit maximum (epistemic) rationality for the few and specific problems in which that specific input does not show up while exhibiting irrationality for all the other problems.

[16] A search problem is a problem which requires identifying a solution in a possibly infinite solution space. This contrasts with a decision problem, which merely asks whether a given answer is a solution.

[17] Together, the initial state, the available actions, the and transition function implicitly define the set of all states reachable from the initial state given any possible sequence of actions (the state space). The state space forms a graph in which the nodes represent states and the links represent actions.

- A path cost function, which accepts a list of actions as input and returns a number. The path cost assigns a numeric cost to each solution for the problem[18].

A class of problems is composed of the problems with the same available actions, transition function, goal test, and path cost function, but different initial state.

4.2.2 Agent

The hypotheses of agents which would exhibits maximum (epistemic) rationality for the chosen class of problems may be formed in different ways:

1. Using the epistemologist's intuitions about how to solve problems in that class.

2. Using empirical data about how humans solve problems in that class.

3. Using existing hypotheses about the rational way of solving problems in that class.

In (1), a study in CE starts exactly as a study in traditional epistemology starts: with epistemic intuitions. The using of epistemic intuitions in stating hypotheses is justified by the fact that a maximally (epistemically) rational agent may be a version of ourselves (it may act in the way we would act in ideal conditions) and our intuitions may reveal the way we would act in these conditions. The hypotheses are not guaranteed to be correct because the more complex are the problems that we are dealing with, the more confused our intuitions tend to be[19]. The difference in how CE and traditional epistemology use intuitions is that, in CE, intuitions play a merely heuristic role: they are used in stating hypotheses that are tested using other means.

In (2), the hypotheses are derived from empirical data about how humans or other animals deal with a class of problems. The reason why using empirical data may be interesting in constructing hypotheses in CE is that humans are the kind of agent known to be closer to exhibiting maximum (epistemic) rationality for most classes of problems[20].

[18] A solution for a problem is a sequence of actions which defines a path from the initial state to a goal state. The quality of a solution is measured by its cost, where the optimal solution has the lowest cost.

[19] There are known cases in which our intuitions simply get lost (e.g. epistemic paradoxes).

[20] There are classes of problems for which other animals perform better than humans: for example, chimpanzees surpass humans in certain memory tasks (Inoue and Matsuzawa, 2007).

In this context, the way humans or deal with some class of problems is a good (but not infallible) guide to how a maximally rational agent would deal with such class of problems. However, empirical data about humans do not always describe maximum (epistemic) rationality because humans are known to commit systematic mistakes (cognitive biases) in reasoning (see Pohl, 2004). Again, empirical data plays a heuristic role within CE.

In (3), the hypotheses are formed from existing theories in epistemology. In that case, I think that the investigation has a better start: generally, theories in epistemological literature have already been tested and shaped from years of discussion and refinements. This path is also interesting because comparing existing hypotheses for maximum (epistemic) rationality may help enriching long standing disputes in the field. This is the case because CE provides epistemology with consideration about efficiency (see section 4.1).

As I have said, an agent is a reasoner with two extra components: a set of goals and a function which has a set of explicit beliefs (`KB`) and a set of goals as inputs and which returns an action. The action that the function may return must be among the actions available in the chosen class of problems. The investigation that I am pursuing is focused only in epistemic reasoning and dealing with agents (instead of plain reasoners) may overcomplicate the study. In this context, we need to somehow 'neutralize' the extra components of agents. The first step is to use the same, reasonable, action function for all agents used in the investigation[21]. The second step is to homogenize the goals of all agents used in the investigation. I will assume that the goals of all agents used in the investigation are down to three: (i) maximizing function g; (ii) staying alive; and (iii) fulfilling the goal of the chosen class of problems. Goal (i) is obviously an epistemic goal and goal (ii) is, in some sense, a consequence of goal (i) (dead agents have zero (true) beliefs, then, maximizing function g implies staying alive). The non-epistemic goal (iii) is important in dealing with the interruptibility requirement: my claim is that if a pattern of reasoning enables an agent to act as to fulfill the (non-epistemic) goal of the chosen class of problems, then it is 'interruptible enough' for that class of problems.

[21] For example, in the literature rational practical reasoning is often modeled as an maximizing expected utility (see Weirich, 2010; Sobel, 1989).

4.2.3 Simulation

The computer simulation of how agent deals with a class of problems combines information from the description of the class of problems with information from the description of the agent. There exist different paradigms and programming languages in which the simulation may be designed and implemented[22], however, if the chosen class of problems and the agents are described using the classical tools of logic and mathematics, the simulation will probably be implementable in most common programming languages.

The simulation of how an agent deals with a problem works as a loop. The loop starts with the environment outputting information about its current state to the agent, which updates its beliefs, executes the necessary amount of reasoning, and outputs an action to the environment, which updates its current state and reinstates the loop. The loop stops when the goal test returns positive or when the agent dies. In order to produce data from the simulation, some tests (if statements) and counters (variables) may be included within the loop. For example, one may include tests and counters for (i) whether the agent solves the problem (i.e. reaches a goal state), (ii) the cost of the solutions generated by the agent, (iii) the time and memory requirements of the agent's patterns of inference.

Since maximum (epistemic) rationality cannot be tied up to specific problems, the data from the simulations of individual problems must be generalized to the class of problems. This may be done by simulating a large number of randomly generated problems in the class and averaging over the results. In addition, most of the requirements for maximum (epistemic) rationality deals with infinite quantities (infinitely many truths, logical consequences, etc), but infinite quantities cannot be directly implemented in a computer simulation. For example, the maximization of function g depends on the existence of infinitely many truths available to be believed, but implementable environments have only finitely many features. In order to deal with this issue, one can randomly generate problems according to the size of their environment and check how the averaged values change as the size of the environments grows. Then one can extrapolate the results using the limit of this sequence.

[22] There are imperative languages (e.g. Fortran, BASIC), functional (e.g. LISP, Haskell), logic-based (e.g. Prolog), and object oriented (e.g. C++, Java), etc (see Louden, 2002).

4.2.4 Analysis

In analyzing data from the simulation, one is able to compare the results for some parameters related to the requirements for maximum (epistemic) rationality, as, for example:

The value of function *g*: The most important goal for maximum (epistemic) rationality is the maximization of function g. Then the epistemologist must check to which value the agent's function g converges as the size of the environment grows.

Accuracy rate: In order to evaluate the agent towards the non-epistemic goal, one must check how often the agent solves problems as their difficulty (size) grows. This value needs not to be maximum, but only high enough for dealing with interruptibility.

Solution cost: The meaning of the cost of a solution varies from problem to problem, but the measurement is always of the same kind: one must include in the code a variable related with the cost of each action and check the final average value of the variable. In order to deal with interruptibility, the solution cost must grow not so fast (polynomially?) as the size of the problems grows[23].

Benchmarking: To 'benchmark' ordinarily means to write a program for an algorithm, run it on a computer, and measure speed in seconds and memory in consumption of bytes. Benchmarking can be unsatisfactory because it is too specific: it measures the performance of a particular program written in a particular language, running on a particular computer, with a particular compiler and particular input data. The technique may be interesting, however, when one is dealing with patterns of inference that are too complex to perform analytic analysis. Also, there are clever ways to benchmark. One may use random input, include in the code some variables related to the number of inferential steps (time) or beliefs (memory), then run the simulation in several trials and plot the average value variables as a function of the size of the input. In plotting the graph, one is able to evaluate the kinds of growth of that variable: polynomial, exponential, etc[24].

[23] The cost helps in dealing with interruptibility because, for an agent, it is always possible to gather more information before perform some actions. But, if there exist a cost for gathering information, the agent must deal with the trade-off between gather more information and acting upon a conclusion.

[24] Although complexity analysis is often made using worst-case analysis, there exists a large literature about average-case analysis (see Bogdanov and Trevisan, 2006). See appendix B.1.1 for a survey on computational complexity theory.

The measurement of the value of function g is relevant for the evaluation of epistemic rationality (requirement (iii)). The assessment of time and space complexity is relevant to deal with efficiency (requirement (iv)). The measurement of the accuracy rate for the non-epistemic goal and of the cost are important to dealing with interruptibility (requirement (v)). Note that I didn't list tests for believing all and only the logical consequences of an epistemic situation (requirement (i)) and for having a nontrivial set of beliefs (requirement (ii)). These requirements are 'easily' checked with analytical means (as in chapter 1). In other words, this requirement must be checked in choosing the hypotheses for agents.

In comparing agents regarding these features, it is possible to assess which of then is closer to maximum (epistemic) rationality for the chosen class of problems.

4.3 Discussion

Most of the discussion about the features of CE was made in the introduction of the dissertation. There, I have argued that:

1. CE is (part of) a normative research program;
2. CE employs well-stablished methods and provides a bridge between methods and results of traditional, naturalized, and formal epistemologies;
3. CE construes epistemology as an autonomous discipline;
4. CE provides epistemology with a new kind of consideration (about efficiency), which may help enriching long standing disputes in the field.

I also argued that these are desirable features for a research program in epistemology. There exists a diffuse field of CE in the literature (e.g. Pollock, 1995), but that field still lacks proper justification and presentation. In this chapter, I have intended to do both. The best argument for the relevance of CE is to show a case in which two theories prescribe otherwise equally reasonable patterns of inference which differ only in efficiency. Then show how the results of CE may help enriching the discussion. I make this case in chapter 5, where I present an example of investigation in CE about reasoning under uncertainty.

Chapter 5

Reason v the monster of uncertainty

> You just let the machines get on the adding up and we'll take care of the eternal verities. Under the law, the Quest for Ultimate Truth is quite clearly the inalienable prerogative of philosophers! Any bloody machine goes and actually finds it and we're straight out a job, aren't we? We demand rigidly defined areas of doubt and uncertainty!
> Douglas Adams, *The Hitchhiker's Guide to the Galaxy*.

In this chapter, I provide an example of investigation in computational epistemology. My aim is to show how considerations about computational complexity may help enriching long-standing disputes in epistemology. An example of a long-standing dispute in epistemology is that about the most adeqate model for maximum rationality for uncertain reasoning[1]. The formal investigations about uncertain reasoning may be organized in two broad categories (see Spohn, 2002):

1. Bayesian epistemology (Joyce, 2011; Williamson, 2010);

2. A divided lot of non-probabilistic frameworks such as:

    (a) the theory of defeasible reasoning (Pollock, 1995);

    (b) the AGM theory of belief revision (Gärdenfors, 1988);

    (c) the theory of rank functions (Spohn, 1988);

    (d) etc[2].

[1] 'Uncertain reasoning' meaning reasoning from uncertain premises or achieving uncertain conclusions.

[2] Other two examples of non-probabilistic frameworks are Dempster-Shafer theory (Shafer, 2010) and the certainty factors model (Heckerman and Shortliffe, 1992).

In this chapter, I compare how theories in these two classes model maximum rationality for uncertain reasoning. Within Bayesian epistemology, I focus on subjective Bayesianism (Joyce, 2011). Within the non-probabilistic frameworks, I focus on the theory of defeasible reasoning (Pollock, 1995). In order to compare how these theories model maximum rationality, I introduce an epistemic version of the Wumpus World (WW), a class of problems described in Russell and Norvig (2010) as a tool for investigating uncertain reasoning. I design agents based on these models, implement these agents in a computer simulation of the epistemic WW, and analyze the data from the simulation.

In section 5.1, I make a survey on Bayesian epistemology. I discuss the notions of degrees of belief, of maximally rational degrees of belief, and of a maximally rational update of degrees of belief. In section 5.2, I make a survey on theory of defeasible reasoning. I discuss the notions of (conclusive, defeasible) reasons, of (rebutting, undercutting) defeaters, and of a maximally rational computation of belief statuses (undefeated, defeated). In section 5.3, I describe the epistemic WW. Also, I describe the general features of agents based on the Bayesian and defeasible models. In section 5.4, I implement these agents in a computer simulation of the epistemic WW and analyze the data from the simulation. I analyze how much these agents succeed in solving problems in that class, the cost of their solutions, and the computational complexity of the patterns of inference of these agents.

5.1 Bayesian epistemology

Roughly, Bayesian epistemology is the use of developments in theory of probability for addressing problems in epistemology. Bayesian epistemologists often model maximum rationality as the generation and update of degrees of belief according to the principles of probability. The features of the model are (Joyce, 2011, p. 17-18):

1. beliefs are modeled as having degrees (degrees of belief);
2. the degrees of belief of a maximally rational reasoner express probabilities;
3. a maximally rational reasoner updates its degrees of belief given new information using some principle of probability (e.g. principle of conditionalization).

The traditional definition of what it means for a reasoner to have a given degree of belief is connected with the betting behavior of the reasoner as follows (Talbott, 2013):

Definition 5.1.1 (Degrees of belief ($pr(\phi)$)). A reasoner $\mathcal{R}$ is said to have a degree of belief x on a sentence ϕ ($pr(\phi) = x$) iff $\mathcal{R}$ would buy a \$1 wager on ϕ for a price equal or less (but not greater) than \$$x$ and would sell such a wager for a price equal or greater (but not less) than \$$(1 - x)$.

Bayesian epistemologists maintain that the degrees of belief of a maximally rational reasoner must express probabilities. A probability function for a language $\mathcal{L}$ is a function $pr : \mathcal{L} \to \mathbb{R}$ satisfying the following constraints for all $\phi, \psi \in \mathcal{L}$ (Demey et al., 2014)[3]:

(i) $pr(\phi) \geq 0$ (lower bound);

(ii) If $\overset{\text{c}}{\models} \phi$, then $pr(\phi) = 1$ (tautologies);

(iii) If $\overset{\text{c}}{\models} \neg(\phi \wedge \psi)$, then $pr(\phi \vee \psi) = pr(\phi) + pr(\psi)$ (finite additivity)[4].

If constraints (i)$-$(iii) hold, then the following are theorems given arbitrary $\phi, \psi \in \mathcal{L}$: $pr(\neg\phi) = 1 - pr(\phi)$ (negation)[5]; $pr(\phi) \leq 1$ (upper bound)[6]; if $\overset{\text{c}}{\models} \phi \leftrightarrow \psi$, then $pr(\phi) = pr(\psi)$ (equivalence)[7]; and $pr(\phi \vee \psi) = pr(\phi) + pr(\psi) - pr(\phi \wedge \psi)$[8].

[3] I think that the most adequate logic for expressing (i)$-$(iii) would be 'the logic of a priori knowledge' $\overset{\text{x}}{\models}$ (see chapter 2), but the exact properties of $\overset{\text{x}}{\models}$ are not easy to come by. Since $\overset{\text{x}}{\models}$ is most likely supraclassical, I will state (i)$-$(iii) in terms of classical logic, where $\overset{\text{c}}{\models}$ is the classical consequence relation.

[4] Probability functions are usually defined for a σ-algebra of subsets of Ω and required to satisfy countable additivity. In logical contexts, it is more natural to define probability functions for the logic's object language (Williamson, 2002). Then finite additivity suffices because $\overset{\text{c}}{\models}$ is finitary (Talbott, 2013).

[5] It is the case that $\overset{\text{c}}{\models} \neg(\phi \wedge \neg\phi)$, what entails that $pr(\phi \vee \neg\phi) = pr(\phi) + pr(\neg\phi)$ (finite additivity). But $\overset{\text{c}}{\models} (\phi \vee \neg\phi)$ and $pr(\phi \vee \neg\phi) = 1$ (tautologies). Then $pr(\phi) + pr(\neg\phi) = 1$ and $pr(\neg\phi) = 1 - pr(\phi)$.

[6] Suppose that $pr(\phi) > 1$. Since $pr(\phi) = 1 - pr(\neg\phi)$ (negation), the supposition entails that $1 - pr(\neg\phi) > 1$, what entails that $-pr(\neg\phi) > 0$ and that $pr(\neg\phi) < 0$. But this is a contradiction because $pr(\neg\phi) \geq 0$ (lower bound). Therefore, $pr(\phi) \leq 1$.

[7] Suppose that $\overset{\text{c}}{\models} \phi \leftrightarrow \psi$. Then $\overset{\text{c}}{\models} \neg(\phi \wedge \neg\psi)$ and $\overset{\text{c}}{\models} \neg(\neg\phi \wedge \psi)$. Also, $\overset{\text{c}}{\models} (\neg\phi \vee \psi)$ and $\overset{\text{c}}{\models} (\phi \vee \neg\psi)$ (De Morgan). Since $\overset{\text{c}}{\models} \neg(\phi \wedge \neg\psi)$, then $pr(\phi \vee \neg\psi) = pr(\phi) + pr(\neg\psi)$ (finite additivity). Then $pr(\phi \vee \neg\psi) = pr(\phi) + 1 - pr(\psi)$ (negation). But $\overset{\text{c}}{\models} (\phi \vee \neg\psi)$, then $pr(\phi \vee \neg\psi) = 1$ (tautologies). Then $pr(\phi) + 1 - pr(\psi) = 1$, what entails that $pr(\phi) = pr(\psi)$.

[8] It is the case that $\overset{\text{c}}{\models} \neg(\phi \wedge (\neg\phi \wedge \psi))$, what entails that $pr(\phi \vee (\neg\phi \wedge \psi)) = pr(\phi) + pr(\neg\phi \wedge \psi)$ (finite additivity). But it is also the case that $\overset{\text{c}}{\models} ((\phi \vee \psi) \leftrightarrow (\phi \vee (\neg\phi \wedge \psi)))$, what entails that (a) $pr(\phi \vee \psi) = pr(\phi) + pr(\neg\phi \wedge \psi)$ (equivalence). It is also the case that $\overset{\text{c}}{\models} \neg((\phi \wedge \psi) \wedge (\neg\phi \wedge \psi))$, what entails that $pr((\phi \wedge \psi) \vee (\neg\phi \wedge \psi)) = pr(\phi \wedge \psi) + pr(\neg\phi \wedge \psi)$ (finite additivity). But $\overset{\text{c}}{\models} (\psi \leftrightarrow ((\phi \wedge \psi) \vee (\neg\phi \wedge \psi)))$, then $pr(\psi) = pr(\phi \wedge \psi) + pr(\neg\phi \wedge \psi)$ (equivalence). Then $pr(\neg\phi \wedge \psi) = pr(\psi) - pr(\phi \wedge \psi)$. Therefore, $pr(\phi \vee \psi) = pr(\phi) + pr(\psi) - pr(\phi \wedge \psi)$ (substituting $pr(\neg\phi \wedge \psi)$ in (a)).

Bayesian epistemologists maintain probabilism: the claim that the degrees of belief of a maximally rational reasoner must express probabilities in the sense of fulfilling constraints (i)−(iii)[9]. This claim is often supported using Dutch Book arguments (DBAs) (e.g. Ramsey, 1926), but there also exist non-pragmatic defenses of probabilism (Joyce, 1998; Pettigrew, 2015a)[10].

Bayesian epistemologists also maintain that a maximally rational reasoner must update its degrees of belief upon new information conforming to some principle of probability. Most often, (subjective) Bayesianists defend that a maximally rational reasoner must update its degrees of belief conforming to the principle of conditionalization (e.g. Joyce, 2011, p. 34)[11]. The definition of the principle of conditionalization depends on the notion of the conditional probability of ϕ given ψ (i.e. $pr(\phi|\psi)$):

$$pr(\phi|\psi) = \frac{pr(\phi \wedge \psi)}{pr(\psi)}, \text{ where } pr(\psi) \neq 0. \tag{CP}$$

Then $pr(\phi \wedge \psi) = pr(\psi|\phi) \times pr(\phi)$ (conjunction)[12]. And $pr(\psi) = pr(\psi|\phi) \times pr(\phi) + pr(\psi|\neg\phi) \times pr(\neg\phi)$[13]. The principle of conditionalization is the following (Talbott, 2013):

Definition 5.1.2 (Principle of conditionalization). In acquiring a new piece of information ψ, a maximally rational reasoner must (i) update its degree of belief on ψ to 1 and (ii) update its other degrees of belief ψ conditionalizing on ϕ ($pr(\phi) = pr(\phi|\psi)$).

The claim that a maximally rational reasoner must update its degrees of belief upon new information conforming to the principle of conditionalization is also often supported using DBAs (Teller, 1976), but there also exist non-pragmatic defenses of this claim

[9] Objetive Bayesianism often maintain additional restrictions for the degrees of beliefs of a maximally rational reasoner. For example, some objective Bayesians maintain that the degrees of belief of a maximally rational reasoner must conform to the principle of maximum entropy (Williamson, 2010).

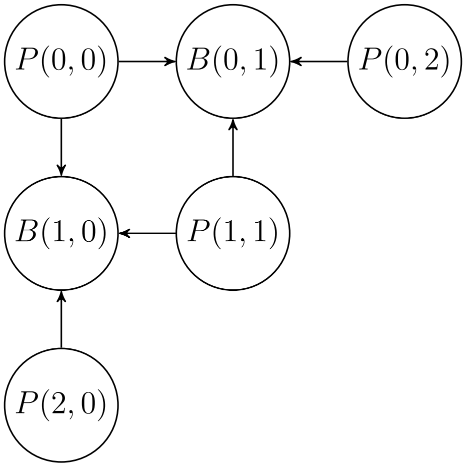
[10] A DBA shows that a reasoner whose degrees of belief violate some principle of probability is vulnerable to having a Dutch Book made against him, i.e. a combination of wager deals which the reasoner would accept but which entails a sure lost. In the traditional interpretation, a DBA states that having degrees of belief which do not express probabilities is pragmatically self-defeating (Talbott, 2013).

[11] Objective Bayesianists often claim that the degrees of belief of a reasoner must be generated and updated using other principles, as the principles of calibration and maximum entropy (Williamson, 2010).

[12] Since $pr(\psi|\phi) = (pr(\psi \wedge \phi))/pr(\phi)$ (CP) and $\overset{c}{\models} ((\phi \wedge \psi) \leftrightarrow (\psi \wedge \phi))$, then $pr(\phi \wedge \psi) = pr(\psi|\phi) \times pr(\phi)$.

[13] It is the case that $\overset{c}{\models} (\psi \leftrightarrow ((\psi \wedge \phi) \vee (\psi \wedge \neg\phi)))$ and $\overset{c}{\models} \neg((\psi \wedge \phi) \wedge (\psi \wedge \neg\phi))$. Then $pr((\psi \wedge \phi) \vee (\psi \wedge \neg\phi)) = pr(\psi \wedge \phi) + pr(\psi \wedge \neg\phi)$ (finite additivity) and $pr(\psi) = pr(\psi \wedge \phi) + pr(\psi \wedge \neg\phi)$ (equivalence). But then $pr(\psi) = pr(\psi|\phi) \times pr(\phi) + pr(\psi|\neg\phi) \times pr(\neg\phi)$ (conjunction).

(Easwaran, 2013; Pettigrew, 2015a). In the following, let 'Bayesian reasoner' denote a maximally rational reasoner according to Bayesian epistemology.



$pr(P(x,y)) = .2$	$pr(B(x,y)\vert \bigvee P(\pm x, \pm y)) = 1$
	$pr(B(x,y)\vert \bigwedge \neg P(\pm x, \pm y)) = .01$

Figure 5.1: A typical belief network. In the graph, nodes represent sentences and edges represent dependency between sentences. The table contain degrees of belief for all nodes without parents and conditional degrees of belief for all the other nodes given their parents. The sentence $\bigvee X(\pm x \pm y)$ means $X(x+1,y) \vee X(x-1,y) \vee X(x,y+1) \vee X(x,y-1)$ and $\bigwedge X(\pm x, \pm y)$ means $X(x+1,y) \wedge X(x-1,y) \wedge X(x,y+1) \wedge X(x,y-1)$.

The degrees of belief of a Bayesian reasoner may be represented using a belief network. A belief network is composed of a direct acyclic graph and of a table of degrees of beliefs[14]. The nodes of the graph represent sentences and edges represent relations of dependence between sentences. If there exists an edge from a node ϕ to a node ψ, then ϕ is a parent of ψ and ψ is a descendant of ϕ. The table contains unconditional degrees of belief for every node without parents and degrees of belief for every other nodes conditional on all possible assignments of their parents. In this context, a Bayesian reasoner updates its degrees of belief upon new information using the following procedure: (i) update the degree of belief on the new information to 1 and (ii) calculate new degrees of beliefs for every other sentence in the graph using joint probability tables. The new degrees of belief on a sentence ϕ is the sum of the probabilities of each assignment in the table that makes

[14] A direct acyclic graph is a collection of nodes and directed edges (edges with an associated direction) with no direct cycles. A directed cycles is a path that loops back to its initial node following the direction of the edges. It is usually accepted that the essential requirement for the soundness and completeness of a belief network is that the dependency relations in the graph reflect causality (Pearl, 1988, p. 14).

ϕ true divided by the sum of the probabilities of all assignments in the table.

For an example, consider a Bayesian reasoner $\mathcal{R}$ with the degrees of belief in figure 5.1. Suppose that $\mathcal{R}$ has acquired the information that $\neg P(0,0)$, $B(1,0)$, and $B(0,1)$. Then $\mathcal{R}$ (i) updates its degrees of belief to $pr(P(0,0)) = 0$ and $pr(B(1,0)) = pr(B(0,1)) = 1$; and (ii) updates the other degrees of belief on using the table 5.1. The new degree of belief on $P(2,0)$ is the sum of the probabilities of assignments 1-4 over the total. The new degree of belief on $P(0,2)$ is the sum of the probabilities of assignments 1, 3, 5, 7 over the total. Then $pr(P(2,0)) = pr(P(0,2)) = .3123$. The new degree of belief on $P(1,1)$ is the sum of the probabilities of the assignments 1, 2, 5, 6 over the total. Then $pr(P(1,1)) = .8525$.

$P(2,0)$	$P(1,1)$	$P(0,2)$	Probability
T	T	T	$(.2)^3 = .008$
T	T	F	$(.2)^2 \times .8 = .032$
T	F	T	$.2 \times .8 \times .2 = .032$
T	F	F	$.2 \times (.8)^2 \times .01 = .00128$
F	T	T	$.8 \times (.2)^2 = .032$
F	T	F	$.8 \times .2 \times .8 = .128$
F	F	T	$(.8)^2 \times .2 \times .01 = .00128$
F	F	F	$(.8)^2 \times (.01)^2 = 6.4 \times 10^{-5}$

Table 5.1: Joint probability table for updating the degrees of belief on $P(2,0)$, $P(1,1)$, and $P(0,2)$ given information that $pr(P(0,0)) = 0$, $pr(P(1,0)) = 1$, and $pr(P(0,1)) = 1$. This situation is represented in figures 5.1 and 5.5 ((a)). The total probability is .2346.

5.2 Defeasible reasoning

The theory of defeasible reasoning (Pollock, 1995) models uncertain reasoning as the adoption and retraction of beliefs given reasons and defeaters. Reasoning is uncertain in the sense that beliefs adopted on the basis of reasons may be retracted later on the basis of defeaters. Maximum rationality is the correct construction of reasons and defeaters and the correct assignment of beliefs statuses given the available reasons and defeaters, where 'correct' means 'as described by the theory'. These are the features of the model:

1. beliefs are modeled as having two possible statuses (undefeated, defeated);

2. a maximally rational reasoner constructs reasons and defeaters correctly;

3. a maximally rational reasoner assigns beliefs statuses correctly.

The notion of a reason is defined as follows (Pollock, 1995, p. 39):

Definition 5.2.1 (Reason). Γ is a reason for $\mathcal{R}$ to believe ϕ iff if $\mathcal{R}$ were to believe ϕ on the basis of Γ, then that belief would be justified,

where $\mathcal{R}$ is a reasoner and Γ is a set of mental states of $\mathcal{R}$ (e.g. beliefs and perceptions).

A reason may be conclusive or defeasible depending on the epistemic support that it provides. The epistemic support of conclusive reasons cannot be overridden by new information (defeaters). For example, believing $\phi \rightarrow \psi$ and believing ϕ is conclusive reason for believing ψ. The epistemic support of defeasible reasons can be overridden by the action of defeaters. For example, perceiving an object as red is defeasible reason for believing that the object is red. This reason is defeasible because its epistemic support may be overridden, for example, by the information that the object is under red light.

I will assume that conclusive reasons work as described by classical logic. For example, believing $\phi \rightarrow \psi$ and believing ϕ is conclusive reason for believing ψ because ψ follows from $\phi \rightarrow \psi$ and ϕ in classical logic. Finding the complete list of defeasible reasons which yield the most sensible results is a major burden for the theory of defeasible reasoning[15]. These are two examples of defeasible reasons (Pollock, 1995, p. 51)[16]:

> **Perceptual input:** x's appearing F is defeasible reason for believing $F(x)$.
>
> **Statistical syllogism:** If $r > .5$, then believing $pr(F|G) \geqslant r$ and $G(x)$ is defeasible reason for believing $F(x)$ (Pollock, 1995, p. 68)[17],

The epistemic support of defeasible reasons may be overridden by the action of defeaters. The definition of a defeater is as follows (Pollock, 1995, p. 85):

[15] I think that this investigation cannot be done on a priori basis because such a list is relative to a class of environments (see chapter 1, section 1.3). For example, the optimal value of r in the statistical syllogism is relative to class of environments. In this context, computational epistemology is an adequate framework for investigating the optimal list of defeasible reasons (see chapter 4, section).

[16] These are simplified glosses, omitting important qualifiers and details. For example, I am assuming that all defeasible reasons generate the same amount of epistemic support (see Pollock, 1995, p. 93, for a discussion about reasons with different degrees of epistemic support).

[17] These are indefinite probabilities, where $pr(F|G)$ means 'the proportion of physically possible Gs that would be Fs'. The optimal value of r depends on the environment (.5 is the lowest acceptable value).

Definition 5.2.2 (Defeater). Γ_2 is a defeater for Γ_1 as a reason for believing ϕ iff:

(i) Γ_1 is a defeasible reason for believing ϕ;

(ii) $\Gamma_1 \cup \Gamma_2$ is not a reason for believing ϕ.

A defeater may be a rebutting or an undercutting defeater. Rebutting defeaters attack the conclusion of a reason. For example, if perceiving at distance what appears to be a sheep in the field is defeasible reason for believing that there is a sheep in the field, then hearing from the shepherd that there are no sheep in the field is a rebutting defeater for that reason. The definition of a rebutting defeater is as follows (Pollock, 1995, p. 85):

Definition 5.2.3 (Rebutting defeater). Γ_2 is a rebutting defeater for Γ_1 as a reason for believing ϕ iff:

(i) Γ_2 is a defeater for Γ_1 as a reason for believing ϕ;

(ii) Γ_2 is reason for believing $\neg\phi$.

Undercutting defeaters attack the connection between a reason and its conclusion. For example, if perceiving an object as red is defeasible reason for believing that the object is red, then learning that the object is under red light is an undercutting defeater for that reason. Undercutting defeaters are reasons for believing that a reason does not support a conclusion. The definition of an undercutting defeater is as follows (Pollock, 1995, p.86):

Definition 5.2.4 (Undercutting defeater). Γ_2 is a undercutting defeater for Γ_1 as a reason for believing ϕ iff:

(i) Γ_2 is a defeater for Γ_1 as a reason for believing ϕ;

(ii) Γ_2 is a reason for believing (Γ_1 is not a reason for believing ϕ).

This is an example of an undercutting defeater (Pollock, 1995, p. 85):

> **Subproperty defeat:** Believing $H(x) \wedge pr(G|F \cap H) \neq pr(G|F)$ is an undercutting defeater for the statistical syllogism.

In the following, let 'defeasible reasoner' denote a maximally rational reasoner according to the theory of defeasible reasoning. The doxastic state of a defeasible reasoner may be represented as an inference graph. In the graph, nodes represent sets of mental states with content represented as sentences and edges represent dependency relations among mental states of three kinds: (a) conclusive support (arrows), (b) defeasible support (dashed arrows), and (c) defeat (red arrows). Compound arrows link the union of nodes to a node. Consider some examples in figure 5.2.

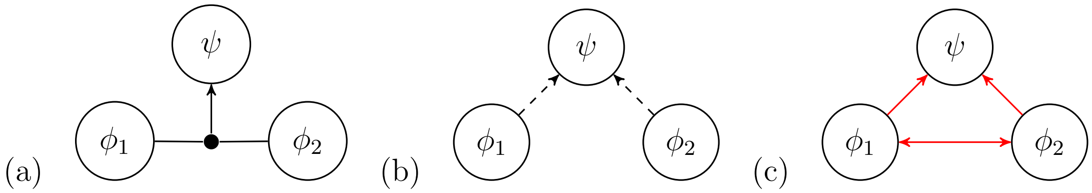


Figure 5.2: (a) $\{\phi_1, \phi_2\}$ is conclusive reason for $\{\psi\}$; (b) $\{\phi_1\}$ and $\{\phi_2\}$ are (independent) defeasible reasons for ψ; (c) $\{\phi_1\}$ and $\{\phi_2\}$ are undercut defeaters for $\{\psi\}$ and rebut each other. Rebutting defeaters are symmetrical and undercutting defeaters are not[18].

In the theory of defeasible reasoning, beliefs are not modeled as having degrees. The belief statuses of a reasoner towards a sentence ϕ may have two values: defeated, undefeated. Pollock have described several procedures for computing the statuses of beliefs given an inference graph. Here, I focus on the multiple-assignment semantics (Pollock, 1995). Let the support list of a node ϕ be composed of all nodes having support edges toward ϕ and the defeat list of ϕ be composed of all nodes having defeat edges towards ϕ. Then consider the definitions (Pollock, 1995, p. 308):

Definition 5.2.5 (Initial node). A node is an initial node iff both its support list and its defeat list are empty.

Definition 5.2.6 (Status assignments). An assignment σ of 'defeated' and 'undefeated' to a subset of the nodes of an inference graph is a status assignment iff:

1. σ assigns 'undefeated' to any initial node;
2. σ assigns 'undefeated' to a non-initial node ϕ iff σ assigns 'undefeated' to some of the members of the support list of ϕ and all defeaters of ϕ are assigned 'defeated';

[18] In the following, I will drop the set notation.

3. There does not exist an status assignment σ' such that $\sigma \subset \sigma'$.

The assignment of 'undefeated' and 'defeated' in a graph is done by supervaluation.

Definition 5.2.7 (Supervaluation). A node is undefeated iff every status assignment assigns 'undefeated' to it; otherwise it is defeated[19].

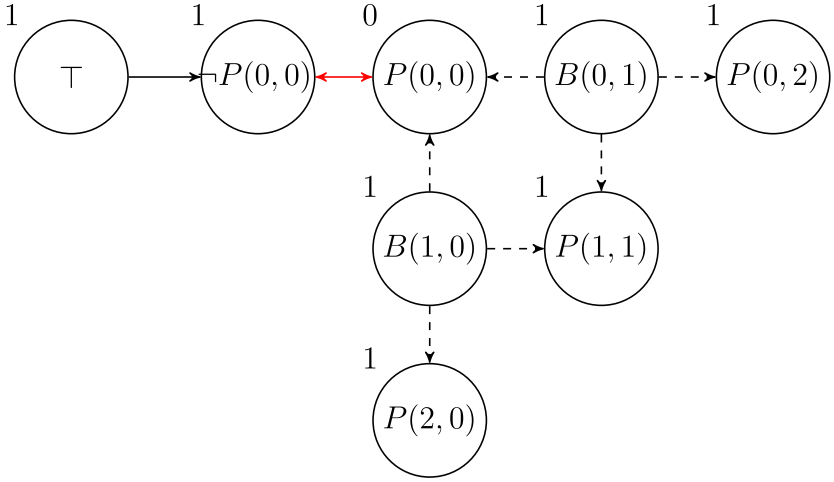


Figure 5.3: A typical inference graph. In the graph, nodes represent sets of mental states with content represented as sentences and edges represent dependency relations between mental states. Arrow represent conclusive support, dashed arrows represent defeasible support, and red arrows represent defeat. The labels '1' and '0' represent, respectively, undefeated and defeated.

For an example, consider a defeasible reasoner with an inference graph as shown in figure 5.3. Roughly, this is how the reasoner would compute its belief statuses. The nodes $\top$, $B(1,0)$, and $B(0,1)$ are initial nodes and, therefore, are undefeated. The node $\top$ is a conclusive reason for $\neg P(0,0)$, which makes $\neg P(0,0)$ also undefeated. The defeat list of $P(0,0)$ has a node which is undefeated ($\neg P(0,0)$), then $P(0,0)$ is defeated. All the nodes in the support list of $P(2,0)$, $P(1,1)$, and $P(0,2)$ are undefeated ($B(0,1)$ and $B(1,0)$) and these nodes do not have defeaters, then $P(2,0)$, $P(1,1)$, and $P(0,2)$ are undefeated.

[19] This definition differs from Pollock's: "A node is undefeated iff the graph contains an argument for the node which is such that every status assignment assigns an undefeated value to all nodes and links in the argument; otherwise it is defeated" (Pollock, 1995). I think that Pollock's definition is too strong and does not deal well with the case in which there exist defeasible reasons for ϕ_1 and for ϕ_2, ϕ_1 and ϕ_2 rebut each other, and both ϕ_1 and ϕ_2 are reasons for ψ (e.g. ϕ_1 = 'x is red', ϕ_2 = 'x is green', and ψ = 'x is colored'). In this case, I think that ψ is undefeated, but Pollock's definitions entail that it is defeated.

5.3 Wumpus World

The Wumpus World (WW) is a well-known class of problems for investigating uncertain reasoning (Russell and Norvig, 2003, p. 197). The WW is a cave consisting of rooms connected by passageways and surrounded by walls. Lurking somewhere in the cave is the Wumpus, a ferocious beast that kills anyone who enters its room. Also, some rooms contain a bottomless pit that kills anyone who enters the room (except for the Wumpus, which is too big to fall in). The only mitigating feature of WW is the existence of a heap of gold hidden somewhere in the cave. The goal of an agent in the WW is to explore the cave, find the gold, and escape the cave with the gold (without being killed).

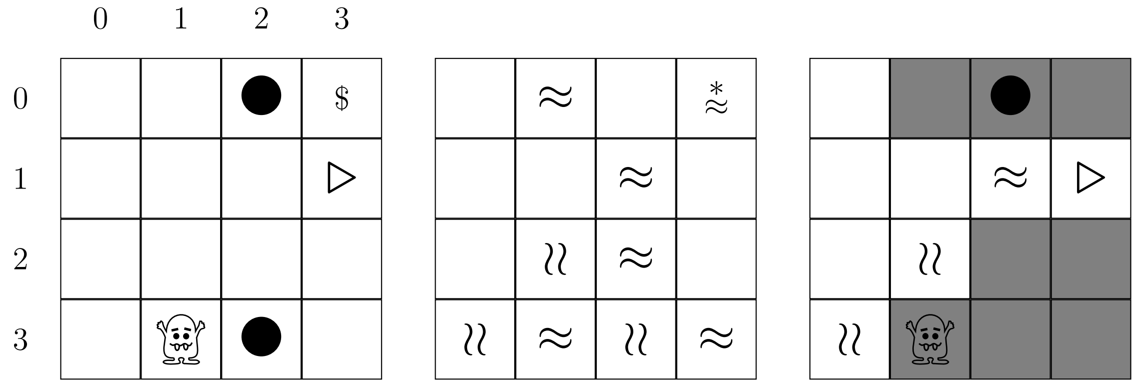


Figure 5.4: From left to right, the full map of an instance of WW, the percepts that are available in that map, and the known features of the map. The symbols mean: $\bullet$ = pit, 👻=Wumpus, \$ = gold, $\triangleright$ = agent, $\approx$ = breeze, $\wr\wr$ = stench, and $*$ = glitter. The reasoner has visited the squares in the order (0,0), (1,0), (2,0), (3,0), (2,0), (2,1), (1,1), (1,2), (1,3). Non-visited squares are in gray. At this point, the reasoner may have a (high degree of) belief that the Wumpus is located in square $(3, 1)$ and that there exists a pit at $(0, 2)$.

The WW may be represented as a grid of squares, where the square (x, y) is in the xth line and yth column of the grid (see figure 5.4). The reasoner starts at square $(0, 0)$ facing right. Every square other than $(0, 0)$ has probability .2 of containing a pit. The Wumpus and the gold are randomly positioned in a square other than $(0, 0)$ with uniform distribution. Moving around the grid, the agent encounters percepts containing information about the location of pits, Wumpus, and gold. In the directly (not diagonally) adjacent squares to a square containing a pit (Wumpus), there exists a breeze (stench). In a square containing a gold, there exists a glitter. The initial description of WW is the following:

- $pr(P(x,y)) = .2$ for $(x,y) \neq (0,0)$.
- $pr(W(x,y)) = pr(G(x,y)) = \dfrac{\text{\# squares -1}}{\text{\# squares}}$ for $(x,y) \neq (0,0)$.
- If $P(x,y)$, then $\bigwedge B(\pm x, \pm y)$ for existing $(\pm x, \pm y)$.
- If $W(x,y)$, then $\bigwedge S(\pm x, \pm y)$ for existing $(\pm x, \pm y)$.
- $G(x,y)$ iff $Gl(x,y)$.

The predicates $P(x,y)$, $W(x,y)$, $G(x,y)$, $B(x,y)$, $S(x,y)$, $Gl(x,y)$ mean, respectively, that there exists a pit, Wumpus, gold, breeze, stench, and glitter at square (x,y).

My goal is to compare two models of reasoning under uncertainty. For that reason, I will use an epistemic version of WW which has two modifications. The first modification is that, in addition to finding the gold, an agent in the epistemic WW has the goal of having the most comprehensive and accurate set of beliefs about the environment. In chapter 3, I have introduced a function $g(t,f) = \dfrac{t-f}{t+f+a}$ for evaluating the set of beliefs of a reasoner (t is the number of true beliefs, f is the number of false beliefs, and a is a small constant). In this context, it is a goal in the epistemic WW to maximize function g. The second modification is an additional layer of uncertainty: every square other than $(0,0)$ has a probability of .01 of containing a breeze (stench) which is independent of the actual position of pits (Wumpus)[20].

The final description of the epistemic WW is the following:

> **Performance measure:** $+1000$ for having the gold, $+1000$ for being alive, -1 for each action, and $+(\dfrac{t-f}{t+f+a}) \times 1000$, where t is the number of true beliefs of the agent, f is the number of false beliefs of the agent, and a is a constant.
>
> **Environment:** The agent starts at square $(0,0)$, facing right. Each square other than $(0,0)$ contains a pit with probability .2. Wumpus and gold are placed randomly

[20] This small probability may be interpreted as the reasoner's perception having a small probability of a persisting false positive (i.e. having a persisting perceptual illusion). This modification intends to block the use of deductive agents in the epistemic WW (e.g. Sardina and Vassos, 2005; Shapiro and Kandefer, 2005; Thielscher, 2005, use deductive agents).

> in a square other than $(0, 0)$ with uniform distribution. In the directly (not diagonally) adjacent squares to a square containing a pit (Wumpus), there exist a breeze (stench). The square containing the gold contains a glitter. Every square other than $(0, 0)$ has an independent probability of .01 of containing a breeze (stench).
>
> **Actuators:** The agent is able to move forward, turn left by 90°, turn right by 90°, grab something in the same square, and climb out the cave when at square $(0, 0)$. The agent dies if it enters a square containing a pit or the Wumpus. Moving forward has no effect if there is a wall in front of the agent.
>
> **Sensors:** In a square containing a breeze, the agent will perceive a breeze. In a square containing a stench, the agent will perceive a stench. In a square containing a glitter, the agent will perceive a glitter.

In the epistemic WW, the agent has initial information about the rules of the problem and about its initial conditions. The agent has information about the size of the grid, the probability of pits in squares other than $(0, 0)$ (.2), the number of Wumpusses and heaps of gold in the grid (1), that every square adjacent to a pit (Wumpus) contains a breeze (stench), that the square containing a glitter also contains a gold, and that there exists an independent probability (.01) of existing a breeze (stench) in squares other than $(0, 0)$. Also, the agent has the information that it begins at square $(0, 0)$, facing right, without the gold, and about the consequences of its actions (including the performance measure).

I use matrices of numbers for expressing sentences, where each position in a matrix represents the corresponding position in the cave. I use different matrices for expressing sentences about breeze, stench, glitter, pit, Wumpus, and gold. For example, the position (x, y) in the matrix about pits represent the sentence $P(x, y)$. A '1' in that position means that the reasoner believes $P(x, y)$, a '0' means that the reasoner believes $\neg P(x, y)$, a '.33' means that the reasoner has a .33 degree of belief on $P(x, y)$, and a 'null' means that the reasoner believes neither $P(x, y)$ nor $\neg P(x, y)$.

5.3.1 Bayesian reasoner

The design of the Bayesian reasoner follows the principles in section 5.1. The degrees beliefs of the Bayesian reasoner express probabilities. Then, for example, if the reasoner has a degree of belief of x in the sentence ϕ, then it has a degree of belief of $1 - x$ in the sentence $\neg\phi$. Also, the Bayesian reasoner updates its degrees of belief using joint probability tables as shown in section 5.1. The initial degrees of belief of the reasoner are:

(u$_1$) $pr(P(x,y)) = .2$ for all $(x,y) \neq (0,0)$.

(u$_2$) $pr(W(x,y)) = pr(G(x,y)) = \dfrac{\text{\#squares - 1}}{\text{\#squares}}$ for all $(x,y) \neq (0,0)$.

(u$_3$) $pr(P(0,0)) = pr(W(0,0)) = pr(G(0,0)) = 0$.

Sentence (u$_1$) states that, for every square (x,y) other than $(0,0)$, the reasoner has initial a degree of belief of .2 that (x,y) contains a pit. Sentence (u$_2$) states that, for every square (x,y) other than $(0,0)$, the reasoner has initial degree of belief of $\dfrac{\text{\#squares - 1}}{\text{\#squares}}$ that (x,y) contains the Wumpus (gold). Finally, sentence (u$_3$) states that the reasoner has initial degree of belief of 1 that square $(0,0)$ does not contain a pit, Wumpus, or gold. The conditional degrees of belief of the Bayesian reasoner are:

(c$_1$) $pr(B(x,y)|\bigvee P(\pm x, \pm y)) = 1$, $pr(B(x,y)|\bigwedge \neg P(\pm x, \pm y)) = .01$;

(c$_2$) $pr(S(x,y)|\bigvee W(\pm x, \pm y)) = 1$, $pr(S(x,y)|\bigwedge \neg W(\pm x, \pm y)) = .01$;

(c$_3$) $pr(Gl(x,y)|G(x,y)) = 1$, $pr(Gl(x,y)|\neg G(x,y)) = 0$.

Sentence (c$_1$) states that if a square is a neighbor of a square containing a pit, it contains a breeze with probability 1; else it contains a breeze with probability .01. Sentence (c$_2$) states that if a square is a neighbor of a square containing a Wumpus, it contains a stench with probability 1; else it contains a stench with probability .01. Sentence (c$_3$) states that a square contains a gold iff it contains a glitter.

(Part of) The belief network of the Bayesian reasoner is represented in figure 5.1. In encountering new information, the Bayesian reasoner updates its beliefs as following: (i) it updates the degree of belief on the new information to 1 and (ii) updates the other sentences in the belief network using joint probability tables[21].

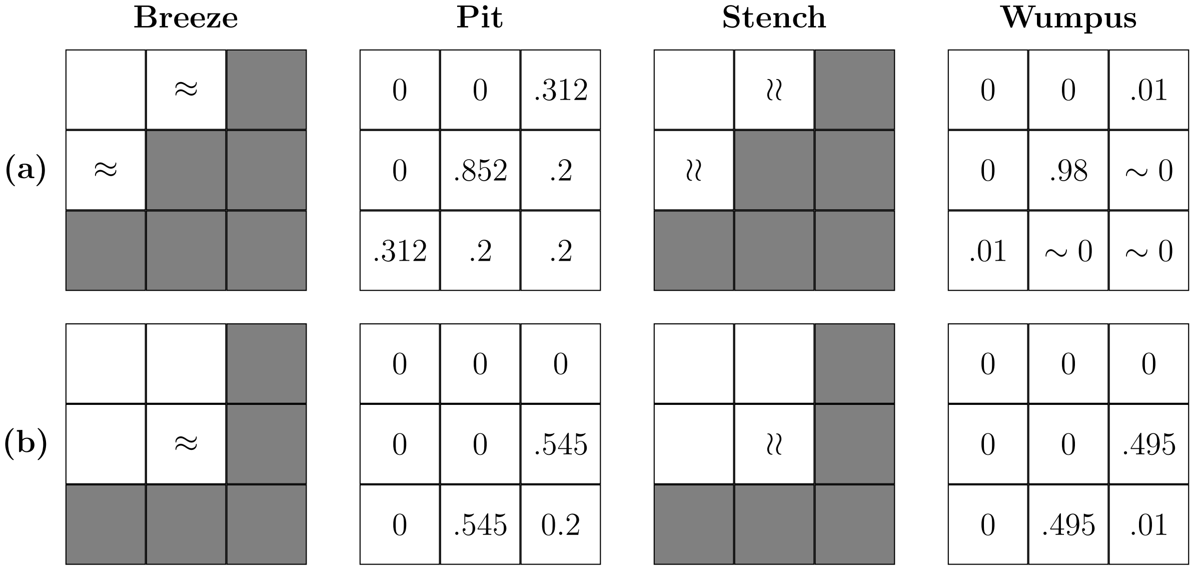


Figure 5.5: Two different configurations of the WW (lines). The percepts about breeze and stench and the degrees of beliefs about pits and the Wumpus (columns). The symbols mean: ≈= breeze, ≀≀ = stench, and ∼= approximately. Non-visited squares are in gray.

Consider how a Bayesian reasoner computes its degrees of belief about pits and Wumpus in the situations in figure 5.5. The joint probability table for dealing with pits in situation (a) is in table 5.1. The updated degrees of belief on $pr(P(2,0)) = pr(P(0,2)) = .312$ and $pr(P(1,1)) = .852$ are calculated using lines 1-4; 1, 3, 5, 7; and 1, 2, 5, 6 of the table. The joint probability table for dealing with Wumpus in situation (a) is in table 5.2. The updated degrees of belief of $pr(W(2,0)) = pr(W(0,2)) = .01$, and $pr(W(1,1)) = .98$ are calculated using lines 1, 2, and 3 of table (a). The joint probability table for dealing with pits and Wumpus in situation (b) is in table 5.2. The updated degree of belief $pr(P(2,1)) = .545$ and $pr(P(1,2)) = .545$ are calculated using lines 1, 2 and 1, 3 of table (b2). The updated degrees of belief $pr(W(2,1)) = .495$, $pr(W(1,2)) = .495$, and $pr(W(2,2)) = .01$ are calculated using lines 1, 2, and 3 of table (b1) respectively.

An advantage of the Bayesian reasoner is its high precision. In situation (a) of figure 5.5, for example, the Bayesian reasoner has a much higher degree of belief that $(1,1)$

[21] For efficiency, the reasoner may group in different tables different groups of independent sentences.

(a)

$W(2,0)$	$W(1,1)$	$W(0,2)$	Probability
T	F	F	$.167 \times (.833)^2 \times .01 = .0012$
F	T	F	$.833 \times .167 \times .833 = .1159$
F	F	T	$(.833)^2 \times .167 \times .01 = .0012$
F	F	F	$(.8)^3 \times (.01)^2 = 5.78 \times 10^{-5}$

(b1)

$W(2,1)$	$W(1,2)$	Probability
T	F	$.333 \times .666 = .222$
F	T	$.666 \times .333 = .222$
F	F	$(.666)^2 \times .01 = .004$

(b2)

$P(2,1)$	$P(1,2)$	Probability
T	T	$.2 \times .2 = .04$
T	F	$.2 \times .8 = .16$
F	T	$.8 \times .2 = .16$
F	F	$(.8)^2 \times .01 = .0064$

Table 5.2: Joint probability tables for situations in figure 5.5. Table (a) is about Wumpus in situation (a). The total probability is .1184, $pr(W(2,0)) = pr(W(0,2)) = .01$, and $pr(W(1,1)) = .98$. Table (b1) is about Wumpus in situation (b). The total probability is .448, $pr(W(2,1)) = pr(W(1,2)) = .495$. Table (b2) is for reasoning about pits in situation (b). The total probability is .3664, $pr(P(2,1)) = pr(P(1,2)) = .545$.

contains a pit in comparison to $(2,0)$ and $(0,2)$. Then the reasoner may choose to go to $(2,0)$ or $(0,2)$, but not to $(1,1)$. On the other hand, the pattern of inference of the Bayesian reasoner is very demanding. Joint probability tables may have 2^n lines, where n is the number of sentences in the table.

5.3.2 Defeasible reasoner

The design of the defeasible reasoner follows the principles in section 5.2. The belief state of the reasoner towards a sentence ϕ is modeled using three values: believing ϕ (1); believing $\neg\phi$ (0); and not believing either ϕ or $\neg\phi$ (null)[22]. The defeasible reasoner constructs reasons and defeaters and compute its belief statuses as described in section 5.2. In reasoning about pits, the reasoner uses the following rules:

(b_1) $B(x,y)$ is defeasible reason for $\bigwedge P(\pm x, \pm y)$.

(b_2) $\neg B(x,y)$ is conclusive reason for $\bigwedge \neg P(\pm x, \pm y)$.

[22] 'Believing ϕ' means 'having an undefeated belief that ϕ' and 'not believing ϕ' means either 'having a defeated belief that ϕ' or that ϕ is not even in the inference graph. The use of null values distinguishes defeasible and Bayesian agents, but this difference is not essential for our purposes (see note 28).

Rule (b_1) states that perceiving a breeze in a square is defeasible reason for believing that there exist pits in the adjacent squares. Rule (b_2) states that perceiving the absence of breeze in a square is conclusive reason for believing that there does not exist a pit in the adjacent squares. In reasoning about Wumpus, the reasoner uses the following rules[23]:

(s_1) $S(x_1, y_1) \wedge ... \wedge S(x_n, y_n)$ is defeasible reason for $\bigvee W(\pm x_1, \pm y_1) \vee ... \vee \bigvee W(\pm x_n, \pm y_n)$.

(s_2) $S(x, y)$ is defeasible reason for $\neg W(z, w)$ for $(z, w) \neq (\pm x, \pm y)$.

(s_3) $\neg W(x, y)$ for all (x, y) is an undercutting defeater for (s_2).

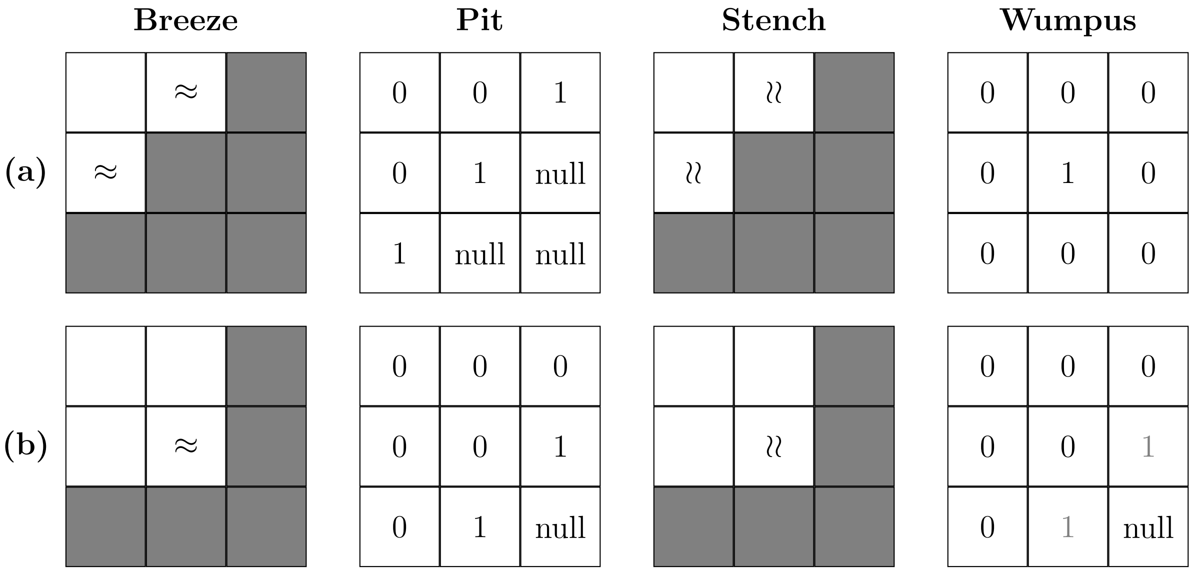
(s_4) $\neg S(x, y)$ is conclusive reason for $\bigwedge \neg W(\pm x, \pm y)$.

Rule (s_1) states that the reasoner has defeasible reason for believing that the Wumpus is in some square around the stenches it has perceived so far. Rule (s_2) states that each stench gives the reasoner defeasible reason for believing that the Wumpus is nowhere not around that stench. If the reasoner has encountered some stench not related with the position of the Wumpus, it is possible that no position of the Wumpus satisfies rule (s_2). In this case, rule (s_3) is an undercut defeater for rule (s_2). Rule (s_4) states that perceiving the absence of stench in a square is a conclusive reason for believing that the Wumpus is not in the adjacent squares. In reasoning about gold, the reasoner uses these two rules:

(g_{1-2}) $Gl(x, y)$ $(\neg Gl(x, y))$ is conclusive reason for $G(x, y)$ $(\neg G(x, y))$.

In situation (a) of figure 5.6, part of the defeasible reasoner's inference graph for dealing with pits is in figure 5.3. The belief that $\neg P(0, 0)$ is conclusive because of the definition of WW. Since $\neg B(0, 0)$, the beliefs that $\neg P(0, 1)$ and $\neg P(1, 0)$ are conclusive. Since $B(0, 1)$ and $B(1, 0)$, the beliefs that $P(0, 2)$, $P(1, 1)$, and $P(2, 0)$ are defeasible. The reasoning about Wumpus has the same structure for $\neg W(0, 0)$, $\neg W(0, 1)$, and $\neg W(1, 0)$. Since $S(0, 1)$ and $S(1, 0)$, the beliefs that $W(0, 2) \vee W(1, 1) \vee W(2, 0)$, $\neg W(0, 2)$, and $\neg W(2, 0)$ are defeasible. Then the reasoner has the defeasible conclusion that $W(1, 1)$

[23] The sentence $(x, y) \neq (\pm z, \pm w)$ means $(x, y) \neq (z + 1, w) \wedge (x, y) \neq (z - 1, w) \wedge (x, y) \neq (z, w + 1) \wedge (x, y) \neq (z, w - 1)$ (for existing $(\pm z, \pm w)$).



	Breeze			Pit			Stench			Wumpus		
(a)		≈	▨	0	0	1		≀≀	▨	0	0	0
	≈	▨	▨	0	1	null	≀≀	▨	▨	0	1	0
	▨	▨	▨	1	null	null	▨	▨	▨	0	0	0
(b)			▨	0	0	0			▨	0	0	0
		≈	▨	0	0	1		≀≀	▨	0	0	1
	▨	▨	▨	0	1	null	▨	▨	▨	0	1	null

Figure 5.6: The epistemic state of the defeasible reasoner in two configurations of WW. In the lines, two different configurations of WW. In the columns, the percepts about breeze and stench and the beliefs about pits and the Wumpus. The symbols mean: $\approx$ = breeze, $\wr\wr$ = stench. Non-visited squares are in gray. Gray letters mean disjunctions. In (b), the reasoner believe that the Wumpus either in square $(1, 2)$ or $(2, 1)$.

and $\neg W(x, y)$ for all $(x, y) \neq (1, 1)$. In situation (b) of figure 5.6, the reasoning about pits has the same structure of the former for $\neg P(0, 0)$, $\neg P(0, 1)$, and $\neg P(1, 0)$. Since $\neg B(0, 1)$ and $\neg B(1, 0)$, the beliefs that $\neg P(0, 2)$, $\neg P(1, 1)$, and $\neg P(2, 0)$ are conclusive. Since $B(1, 1)$, the beliefs that $P(1, 2)$ and $P(2, 1)$ are defeasible. The reasoning on Wumpus has the same structure for $\neg W(0, 0)$, $\neg W(0, 1)$, and $\neg W(1, 0)$. Since $\neg S(0, 1)$ and $\neg S(1, 0)$, the beliefs that $\neg W(0, 2)$, $\neg W(1, 1)$, and $\neg W(2, 0)$ are conclusive. Since $S(1, 1)$, the belief that $W(1, 2) \vee W(0, 1)$ is defeasible.

The defeasible reasoner is less precise than the Bayesian reasoner. In situation (a), for example, the Bayesian reasoner has a much higher degree of belief that $(1, 1)$ contains a pit in comparison to $(0, 2)$ and $(2, 0)$, where the defeasible reasoner believes indistinctly that there exist pits in squares $(0, 2)$, $(1, 1)$, and $(2, 0)$. On the other hand, the defeasible reasoner seems to have a much less demanding pattern of inference than the Bayesian reasoner. In a situation such as (a), for example, the Bayesian reasoner must construct a joint probability table with 2^n possible assignments, where the defeasible reasoner needs to execute only about n^c rules, where c is a constant.

5.4 Results

In order to compare the Bayesian and defeasible models of maximum rationality for uncertain reasoning, I have designed agents based on these models and implemented these agents in a computer simulation of the epistemic WW. The agents are composed of four modules: perception, memory, practical cognition, and epistemic cognition. The agents have the same perception, memory, practical cognition, and part of epistemic cognition, but differ in the relevant aspects of epistemic cognition. The modules of perception and memory are very straightforward. Perception simply receives the available percepts from the environment, encodes the percepts in the format of beliefs (matrices of values as specified in section 5.3), and stores the corresponding perceptual beliefs in memory. Memory simply stores beliefs and makes them available for practical and epistemic cognition.

The role of practical cognition is to select actions. In the implementation, the work of practical cognition is to construct and execute plans (sequences of actions). First, practical cognition checks whether there is a plan being executed. If there is, it returns the next action in the plan. If there is not, it formulates a plan and returns the first action in the plan. In formulating a plan, practical cognition chooses the sub-goal with higher utility among the following: 'grab the gold', 'get out of the cave', and 'move to a fringe square (x, y)', where a fringe square is a non-visited square neighbor of a visited square. The utility of moving to a square (x, y) is calculated differently depending on the agent[24]; the utility of grabbing the gold (if the square contains the gold) is 1000; and the utility of getting out of the cave is 0. The plan for moving to a square is constructed using a greedy search algorithm with cost computed as number of actions; the plan for grabbing the gold is composed by the action 'grab the gold'; and the plan for getting out of the cave is a plan for moving to square $(0, 0)$ plus the action 'get out'.

The role of epistemic cognition is to update beliefs on the face of new information. Epistemic cognition queries memory for the new perceptual beliefs and updates the other

[24] The Bayesian agent calculates the utility of (x, y) as $1000 \times pr(G(x, y)) - 2000 \times pr(P(x, y) \vee W(x, y))$. The defeasible agent assigns utility -2000 if $b(P(x, y) \vee W(x, y)) = 1$, where $b(\phi)$ is the value of the belief on ϕ; else: it assigns utility 0 if $b(G(x, y)) = 0$; it assigns utility 500 if $b(G(x, y)) = \text{null}$; and it assigns utility 1000 if $b(G(x, y)) = 1$. Epistemic cognition execute these calculations.

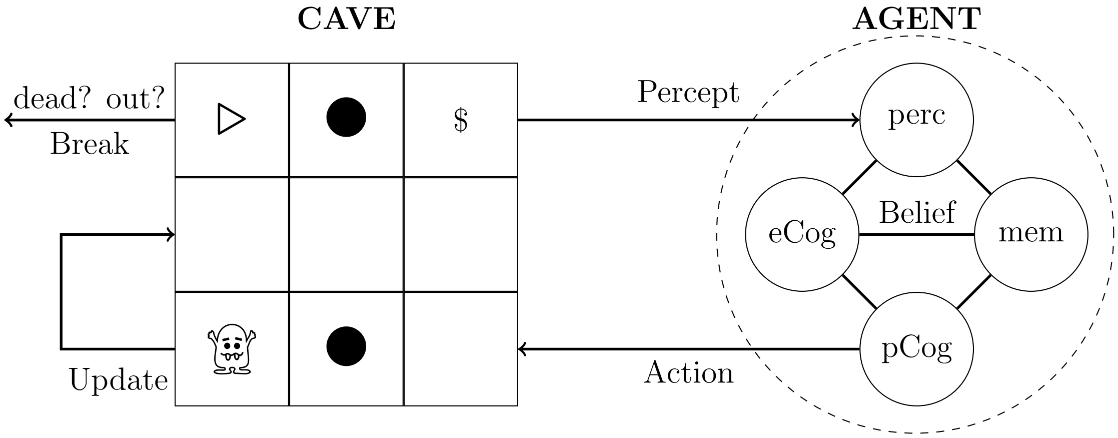


Figure 5.7: The simulation works as a loop. In each iteration, the cave outputs the available percepts for the agent. Then the agent updates its beliefs and returns an action. Finally, the cave updates its current state given the action. The loop stops when the agent dies or gets out of the cave. The labels mean the following: 'perc'= Perception, 'eCog'= Epistemic Cognition, 'mem'= Memory, and 'pCog'= Practical Cognition.

beliefs given the new information. For simplicity, both agents (Bayesian and defeasible) encode and update beliefs about breezes, stenches, and glitter in the same 'all or nothing' way (0 or 1): if the agent perceives glitter in a square (breeze, stench), it concludes that there exist glitter in that square (breeze, stench); else it concludes that there does not exist glitter in that square (breeze, stench)[25]. Then the Bayesian and defeasible agents share part of epistemic cognition. These agents differ fundamentally in how they encode and update beliefs about gold, pits, and Wumpus. The Bayesian agent encodes these beliefs using floating point numbers between 0 and 1 and updates them using joint probability tables as described in section 5.3.1. The defeasible agent encodes these beliefs as having only three values (0, 1, null) and updates them using the rules described in section 5.3.2.

I have implemented the Bayesian and defeasible agents in a computer simulation of the epistemic WW. I have run the simulation using caves with size from 2 to 10 and 1000 trials for each size[26]. In each trial, the main control of the simulation calls the environment constructor for generating a cave with the features described in section 5.3 and the specific size. Then main control calls the agent constructor for generating a Bayesian or defeasible

[25] More naturally, the Bayesian agent would deal with breezes, stenches, and glitters using degrees of belief, but this would only introduce unnecessary calculations. To keep track of the probabilities for glitter, breeze, or stench does not help in solving instances of WW.

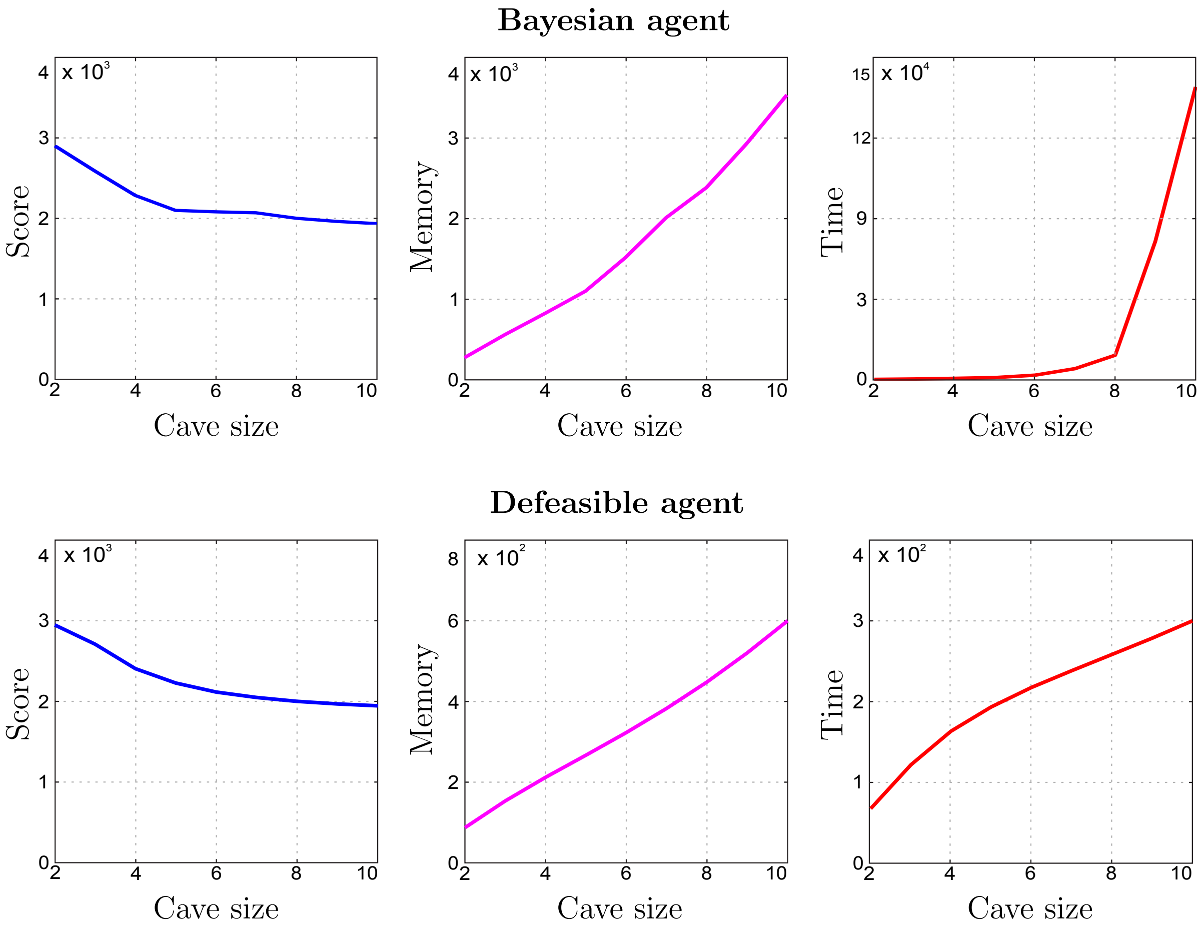
[26] I have used caves with square dimensions. Then a cave with size s has s lines and s columns.

agent with the features described in sections 5.3.1 and 5.3.2 (respectively). After that, the simulation works as a loop (see figure 5.7). In each iteration, main control feeds the agent with the current percepts, queries the agent for its action, and updates the state of the cave given the action. The loop stops when the agent dies or gets out of the cave. When the loop stops, main control returns the final score and the measures for time and space requirements for that trial. I have averaged the results of the 1000 trials. The results are depicted in figure 5.8. Roughly, the Bayesian agent has a score which stabilizes around 2000, a polynomial growth in space requirements, and an exponential growth in time requirements; the defeasible agent has similar score, but a polynomial growth in both space and time requirements.

Bayesian and defeasible agents achieve very similar scores in WW (figure 5.8, blue). As the cave grows, the score of both agents stabilizes around 2000 (2/3 of the maximum). This means that, as the cave grows, these agents tend to find the gold in 2/3 of the trials and to have five times more true beliefs than false beliefs ($g(5f, f) = 2/3$). This is a good result because 21% of the instances of the WW are impossible to solve in the sense of having the gold surrounded by pits or in the same square as the Wumpus (Russell and Norvig, 2010, p. 198). Then the Bayesian and defeasible agents are equally accurate (and accurate 'enough') in dealing with instances of WW. This means that these two models are equally close to maximum rationality for the epistemic WW regarding the generation of beliefs and the accuracy of those beliefs (value of function g).

In this context, the results about time and space (memory) requirements are specially relevant for comparing these models[27]. In the next section, I present these results in details. In the following, 'main beliefs' refers to the beliefs about breezes, stenches, glitter, pits, Wumpus, gold, visited squares, and about the current state of an agent (its position, orientation, and whether it has the gold); 'update procedures' refers to procedures used in updating the main beliefs.

[27] In chapter 4, I have argued that maximum rational patterns of inference must have polynomial space and time requirements if this is possible (see chapter 4, section 4.1 for the argument and the appendix B.1 for a survey on computational complexity).



Bayesian agent

Defeasible agent

Figure 5.8: Results of the Bayesian and defeasible agents for caves with size from 2 to 10. In blue, the score of both agents stabilizes around 2000 (2/3 of the maximum). In magenta, both agents have a polynomial growth in space requirements. In red, the defeasible reasoner has a polynomial growth in time whereas the Bayesian reasoner has an exponential growth. The result were averaged over 1000 trials for each size.

5.4.1 Space (memory)

The space requirements were measured in terms of number of bytes stored in memory instead of number of main beliefs. I have made this choice for two reasons. First, I want to stress a difference between the Bayesian and defeasible agents: while the defeasible agent encodes beliefs about gold, pits, and Wumpus using Boolean values (0, 1, null), the Bayesian agent encodes these beliefs using double values (floating point numbers between 0.0 and 1.0). Since storing a double value requires more bits than a Boolean value, the Bayesian agent uses, other things being equal, more memory than the defeasible reasoner. This difference, however, is not relevant to our analysis because we are interested in classes of computational complexity (exponential, polynomial, etc) and not in absolute values. The second reason is that the measurement in terms of bytes is sensible not only to the amount of memory used for storing the main beliefs, but also to the overall amount of memory used in the update procedures (e.g. the amount of memory necessary for storing joint probability tables and support/defeat relations). An issue is that memory used in the update procedures may be allocated only temporarily (e.g. a joint probability table may be deleted after used). In this context, the measurement of space considers only the maximum amount of bytes allocated in memory at the same time through the whole trial.

The amount of bytes necessary for storing the main beliefs grows polynomially on the size of the cave for both Bayesian and defeasible agents. The beliefs about breezes, stenches, glitter, pits, Wumpus, gold, and visited squares are expressed as values in a matrix with the same dimensions of the cave. Then the maximum number of each kind of these beliefs is s^2, where s is the size of the cave[28]. The number of beliefs about the current state of an agent is constant because the reasoner always have exactly one belief about its current position, orientation, and whether it has the gold. Since the number of bytes used for storing the main beliefs is a product of the number of beliefs, the amount of memory necessary for storing the main beliefs grows polynomially on the size of the cave. Then whether an agents is polynomial in space depends on their update procedures.

The Bayesian agent has polynomial growth in space requirements (figure 5.8, magenta). It is sometimes stated that, in order to update its degrees of belief adequately, a Bayesian agents would need to store degrees of belief for all sentences it believes conditional on all combinations of the sentences it believes (Pollock, 2008). This problem may be avoided in WW using beliefs networks. For example, instead of storing degrees of belief about each square containing a breeze conditional on each other squares containing pits, the agent may store only a belief network composed of $2s$ nodes and three probabilities. In addition, the Bayesian agent does not need to store more than one line of the joint probability table at the same time because each line may be constructed independently. For those reasons, the Bayesian agent has a polynomial growth in space requirements.

[28] This number may be smaller if the agent uses null values, but cannot be larger. For this reason, the use null values is not directly for the difference between pollynomial and exponential space requirements.

The defeasible agent also is polynomial in space requirements (figure 5.8, magenta). The update procedure of the defeasible agent needs to store reasons, defeaters, and support/defeat relations. In WW, reasons and defeaters are among the main beliefs themselves. For example, believing that a square contains a breeze is defeasible reason for believing that it neighbors contain a pit. Also, in WW, the information for support/defeat relations may be stored as a small number of general rules as described in section 5.3.2. For this reason, the defeasible agent also has polynomial growth in its space requirements.

5.4.2 Time

The time requirements were measured in terms of the number of inferential steps instead of number of belief updates. An inferential steps is any change in the value of a variable. For example, changing or assigning n values in a matrix consists in n inferential steps. I have made this choice because I want to consider not only changes in the main beliefs, but the overall time necessary for updating the main beliefs (the update procedures). More specifically, I consider the necessary time for constructing joint probability tables and for computing beliefs statuses.

The Bayesian agent has an exponential growth in time requirements (figure 5.8, red). This is the case due to the use of joint probability tables for updating degrees of belief about pits and Wumpus. It is known that, in the worst-case situation, a joint probability table has 2^n lines, where n is the number the sentences in the table[29]. In WW, the number of lines in a joint probability table for updating the degrees of belief about pits, for example, is exponential on the number of fringe squares which may contain pits. In figure 5.9, situation (a), for example, the probability of a pit in any fringe square depends on the probability of a pit in any other fringe square. In this situation, a table for updating degrees of belief about pits has $2^9 = 512$ lines. The same is true for Wumpus[30].

[29] The worst-case situation is the situation in which all lines in the joint-probability table express possible assignments. The worst-case situation is always the case in the epistemic WW when dealing with pits (Wumpus) due the possibility of breezes (stenches) not related with pits (Wumpus).

[30] In WW, the use of belief networks may be optimized in several ways, but such optimization is not enough for achieving polynomial time requirements. Instead of constructing joint probability tables for all fringe squares, the Bayesian agent may include in the same table only the dependent squares. For example, in figure 5.9, situation (b), the agent constructs two joint probability tables with 32 lines each (instead of a table with 1024 lines). However this strategy is not available for all situations (for example,

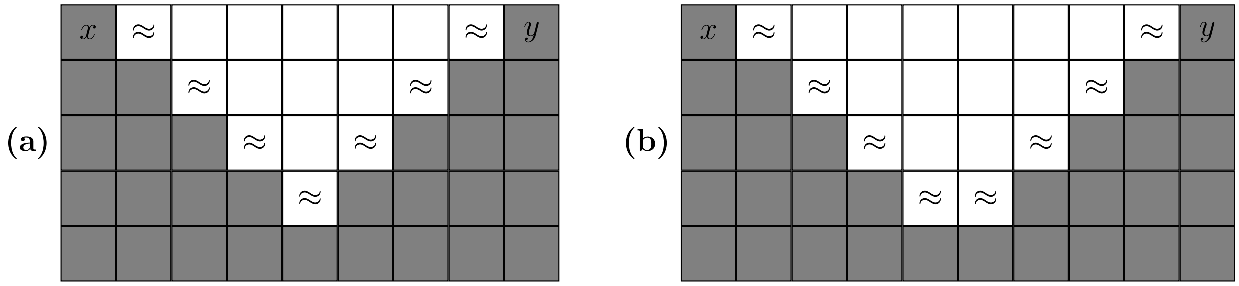


Figure 5.9: Two possible situations for caves with size larger than 9 squares. In situation (a), the joint probability table for updating the degrees of beliefs about pits has 512 lines. This is the case because the probability of pit in each fringe square depends on every fringe square. In situation (b), it is possible to update the degrees of beliefs about pits using two joint probability tables with 32 lines each. This is the case because it is possible to divide the cave into two independent 5×5 partitions. In other words, x and y are conditionally dependent in (a), but are independent in (b).

There are algorithms for performing inference from belief networks in polynomial time (for example, the belief propagation algorithm: Pearl, 1986). These algorithms only work for singly-connected networks when WW requires using multiply-connected networks[31]. In many situations, the belief network about pits (Wumpus) must be multiply-connected because the same breeze (stench) may be caused by pits (Wumpus) in different positions and the same pit (Wumpus) may case breezes (stenches) in different positions. For example, the belief network in figure 5.1 presents more than one undirected paths between nodes $P(1,1)$ and $P(0,0)$: $P(1,1) \rightarrow B(1,0) \rightarrow P(0,0)$ and $P(1,1) \rightarrow B(0,1) \rightarrow P(0,0)$. There are other algorithms for performing inference from multiply-connected belief networks, but those algorithms are all exponential in time. Probably, a polynomial algorithm for inference from multiply-connected networks is impossible because this problem is NP-hard (Cooper, 1990). Then a Bayesian agent for WW has exponential time requirements.

The defeasible reasoner has polynomial growth in time requirements (figure 5.8, red). The assignment of belief statuses about pits is very simple. For every square without breeze, the agent assigns 0 for pits for all surrounding squares. This belief is conclusive. For every square with a breeze, the reasoner assigns 1 for pits for all non-0 surrounding

situation (a) of figure 5.9).

[31] A graph is singly-connected of there exists at most one unidirected path between any two nodes in the graph and multiply-connected otherwise.

squares. The assignment about Wumpus is a little more complicated, but it exploits the fact that reason and defeaters have a fixed structure in WW. If the agent perceives a new stench, then, for each stench it has perceived so far, the reasoner adds one degree of support about Wumpus to all non-0 squares surrounding the stench and subtracts one degree to all other non-0 squares (this is done simply by adding or subtracting an unit in the value of the belief). If, at the end of this process, there exist a square with more support than all others, the agent assigns 1 to that square and 0 to all others. If not, the reasoner assigns null to the squares with more support (it cannot choose between them) and 0 to the others. This procedure results in the pattern of inference described in section 5.3.2 and involves at most s^c belief changes, where s is the size of the cave and c is a constant. For this reason, the defeasible agent has a polynomial time requirements.

5.5 Discussion

In section 5.4, I have shown that Bayesian and defeasible agents are equally accurate (and accurate 'enough') in the environments of WW. Then both models qualify as candidates for modeling maximum rationality for uncertain reasoning. The only relevant difference between these models is that the Bayesian agent is exponential in time and the defeasible agent is polynomial in time (see figure 5.8, red). In chapter 4, I have argued that the maximally rational patterns of inference for a class of environments must be polynomial in space and time if possible. In this case, the results in section 5.4 suggest that maximum rationality for uncertain reasoning should be understood in terms of defeasible reasoning and not of Bayesian epistemology. These results, however, must be interpreted carefully.

The results in section 5.4 were generated using the epistemic WW and the generalization of these results to problems about uncertain reasoning in general depends on how much the features of this class of problem may be generalized. It is true that the epistemic WW has some features which are present in problems of uncertain reasoning in general:

(i) (partially) unreliable informational input;

(ii) no direct access to relevant information;

(iii) revision of (degrees of) belief.

The main sources of uncertainty in reasoning are (i) and (ii). In situations of (i), reasoning is uncertain because premises are uncertain. In WW, (i) happens when the agent has perceptual input about breezes (stenches) which is not related to pits (Wumpus)[32]. In situations of (ii), reasoning is uncertain because some of the information relevant for supporting the conclusion is missing. In WW, (ii) is the case because the agent does not have direct access to information about pits and Wumpus. When (i) or (ii) hold, a reasoner often needs to revise its (degrees of) beliefs on the face of new information (e.g. information which contradicts prior perceptual input). In WW, this happens when, for example, the agent discovers that a breeze (stench) is not related to pits (Wumpus).

On the other hand, there exist some features which are specific of the epistemic WW and which seems to favor the defeasible over the Bayesian model. These are the features:

(iv) diagnostic inference, but not both causal and diagnostic inference;

(v) conditional dependence may be ignored in most cases.

If the difference in efficiency between Bayesian and defeasible agents is a direct consequence of these features which are specific of the epistemic WW, then in other (possibly, more relevant) situations of uncertain reasoning, a Bayesian agent may be equally or more efficient than a defeasible agent. If this is the case, then the results in section 5.4 should not be generalized for uncertain reasoning in general. I will argue that this is the case.

In the epistemic WW, (iv) an agent is required to perform diagnostic inference but not causal inference. Diagnostic inference is inference from effects to causes. For example, concluding that there exist pits in the neighborhood from the fact that there exists a breeze in a square is a diagnostic inference. Causal inference is inference from causes to effects. For example, concluding that there exist breezes in the neighborhood from the fact that there exists a pit in a square is a causal inference. Let e denote a sentence about effects (e.g. $B(x, y)$) and c denote a sentence about causes (e.g. $P(x, y)$). For the Bayesian agent,

[32] This feature of WW models a (persistent) perceptual illusion.

diagnostic inference is more difficult than causal inference because the probability $pr(c|e)$ is not fixed whereas the probability $pr(e|c)$ is fixed (Pearl, 1988, p. 35). In this case, in order to infer c from e (diagnostic inference), the Bayesian agent needs to calculate $pr(c|e)$ before calculating $pr(c)$ because $pr(c|e)$ is not fixed[33]. The inference of e from c may be done directly from $pr(e|c)$ because $pr(e|c)$ is fixed[34]. For a defeasible agent, diagnostic inference and causal inference are equally difficulty. The defeasible agent infers ϕ from ψ if ψ is a reason for ϕ no matter whether ϕ causes ψ, ψ causes ϕ, etc.

This feature is essential for the results in section 5.4 because the results would be different if the problem requires only causal inference or both causal and diagnostic inference. In the first case, the Bayesian agent is as efficient as the defeasible agent because causal inference is equally difficulty for both agents ($pr(e|c)$ is fixed). In the second case, the Bayesian agent may even be more efficient because the defeasible agent has difficulty in dealing with both causal and diagnostic inference. In order to be able to perform both inferences at the same time, the defeasible agent needs to have two reasons forming a cycle (from effect to cause and from cause to effect). In those cycles, a slight evidence may be amplified, generating at each loop an even stronger but unwarranted confirmation[35]. The Bayesian reasoner may perform both causal and diagnostic inference from the same belief network. Then the fact that an agent is required to perform diagnostic inference but not causal inference in the epistemic WW favors defeasible over Bayesian agents.

The second specific feature of the epistemic WW which favors the defeasible over the Bayesian agent is that (v) most relation of conditional dependence may be ignored in solving that class of problems. Two sentences ϕ_1 and ϕ_2 are conditionally dependent given ψ iff $pr(\phi_1 \wedge \phi_2|\psi) \neq pr(\phi_1|\psi) \times pr(\phi_2|\psi)$. In WW, there exist relations of conditional dependence among the positions of pits and Wumpus. For example, a breeze $(1,1)$ may be caused by pits at $(2,1)$ or $(1,2)$. Then discovering that there exists a pit at $(2,1)$

[33] For example, if the agent has only the information that $B(0,1)$, then $pr(P(0,2)|B(0,1)) = .545$; but if the agent acquires the additional piece of information that $B(1,0)$, then $pr(P(0,2)|B(0,1)) = .312$.

[34] For example, the probability $pr(B(0,1)|P(0,2)) = 1$ is fixed.

[35] There are clever ways to deal with causal-diagnostic cycles. Pollock was able to deal with this kind of cycles in his last attempt of providing a semantics for the theory of defeasible reasoning, the critical link semantics (Pollock, 2009). It is worth noticing that the critical link semantics is exponential in time.

decreases the probability of a pit at $(1, 2)$[36]. This pattern of inference is often known as 'explaining away' (Pearl, 1988, p. 7). The Bayesian agent exhibits explaining away 'naturally' in its reasoning procedure from a belief network, although dealing with this kind of relation is the main cause for this agent being exponential in time (see figure 5.9). A defeasible agent may only mimic this pattern of inference using undercutting defeaters, but each relation of undercutting defeater must be added as an independent edges in the inference graph. In this case, if the number of undercutting defeaters grew exponentially on the size of the cave, the defeasible agent would be exponential in both space and time.

This problem may be avoided because the defeasible reasoner is able to ignore most relations of conditional dependence without damaging its performance. In dealing with pits, the defeasible agent ignores relations of conditional dependence altogether. Then any slight reason for a square containing a pit is 'reason enough' for the agent adopting that belief unless there exist conclusive reason for believing otherwise. No undercut defeater is needed here. In reasoning about Wumpus, this strategy is not available because there is only one Wumpus. In this case, undercut defeaters are needed an the pattern on inference is much more complex. However, even in this case, the defeasible agent may still ignore most aspects of the conditional dependence. It is reasonable to suppose that a conclusion based on equivalent reasons from independent sources is, all things being equal, more well supported than a conclusion based on reasons from correlated sources. The defeasible agent deals with beliefs about Wumpus by simply adding reasons and subtracting defeats disregarding relations of conditional dependence. This was crucial for this agent achieving polynomial requirements in space and time. This procedure is only warranted because, in WW, reasons an defeaters are equally probable and are equally dependent of each other given their causes. Otherwise, the agent could be in the position of counting 'twice' the same reason or defeater. Then the fact that most relations of conditional dependence may be ignored in the epistemic WW favors the defeasible over the Bayesian agent.

In the last paragraphs, I have argued that the advantage for the defeasible agent shown in the results in section 5.4 is a direct consequence of features of the epistemic

[36] Then $P(2, 1)$ and $P(1, 2)$ are conditionally dependent given $B(1, 1)$.

WW which are not general to problems of uncertain reasoning. In different conditions, the results could be different: the Bayesian agent could be polynomial in time or the defeasible agent could be exponential in time or memory. Then the results in section 5.4 should not be interpreted as indicating that defeasible reasoning provides a better model of maximum rationality for uncertain reasoning in general than Bayesian epistemology. However, for problems of uncertain reasoning with features (iv) and (v), those results do indicate that defeasible reasoning provides a better model of maximum rationality than Bayesian epistemology. This result is corroborated by several results for similar situations (see Gelfond et al., 2006, for an example).

The difference between the defeasible and the Bayesian models is a case of the difference between extensional and intensional systems in general (Pearl, 1988, p. 3). In an extensional system, the value of a belief is assigned by applying a set of rules to other beliefs. Nonmonotonic formalisms (including the theory of defeasible reasoning) often yield extensional systems and so does the certainty factors model (Heckerman and Shortliffe, 1992). In an intensional system, beliefs pose constraints over the set of acceptable models and the value of a belief is calculated using operations over models[37]. Bayesian epistemology and Dempster-Shafer theory (Shafer, 2010) often yield intensional systems. The real gap between extensional and intensional systems is the difficulty to define extensionally an efficient procedure to do the work of conditional probabilities. The main function of a conditional probability $pr(\phi|\psi)$ is to define the context ψ under which a sentence ϕ may be believed (Pearl, 1988, p. 24). Interpreted as a rule, $pr(\phi|\psi) = x$ does not give us license to do anything without checking the whole context. As soon as new information γ is gathered, the license to assert $pr(\phi) = x$ from $pr(\psi) = 1$ is revoked and we need to look up $pr(\phi|\psi, \gamma)$ instead. A contextual device such as conditional probabilities is necessary for dealing with causal-diagnostic cycles, explaining away, etc. On the other hand, calculations involving conditional probabilities are not efficient in most situations.

[37] For example, the conditional probability $pr(\phi|\psi)$ expresses the weight of the models in which ϕ is true among the models in which ψ is true. This conditional probability cannot be determined from the individual probabilities $pr(\phi)$ and $pr(\psi)$ alone.

Appendix A

Accessible beliefs as modalities

A reasoner is a triple $\mathcal{R} = \langle \mathcal{L}, \texttt{KB}, \pi \rangle$, where $\mathcal{L}$ is a formal language, $\texttt{KB}$ is a set of sentences in $\mathcal{L}$ (explicit beliefs), and $\pi : 2^{\mathcal{L}} \times \mathbb{Z}^+ \rightarrow 2^{\mathcal{L}}$ is an update function (pattern of inference) (def. 1.1.1). A reasoner may be represented as a directed graph in which the nodes are sets of sentences in $\mathcal{L}$ and the edges represent applications of function π to a node an and integer i (an inference). The graph is constructed by adding $\texttt{KB}$ as the initial node ($\texttt{KB}_0$) and then by adding recursively the result of applying function π to the nodes already in the graph and an integer i[1]. Then there exists an arrow labeled i from a node $\texttt{KB}_w$ to a node $\texttt{KB}_{w'}$ iff $\texttt{KB}_{w'}$ is the output of applying function π to $\texttt{KB}_w$ and i (i.e. $\texttt{KB}_{w'} = \pi(\texttt{KB}_w, i)$).

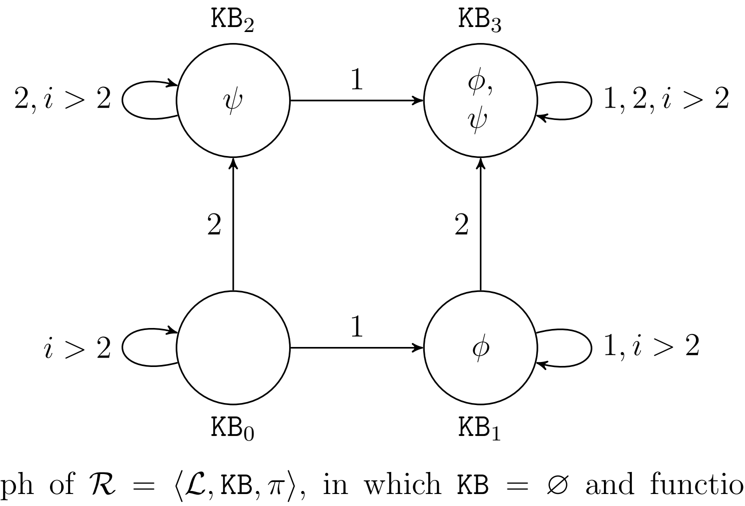


Figure A.1: Graph of $\mathcal{R} = \langle \mathcal{L}, \texttt{KB}, \pi \rangle$, in which $\texttt{KB} = \varnothing$ and function π is such that $\pi(\texttt{KB}_w, 1)$ adds ϕ to $\texttt{KB}_w$, $\pi(\texttt{KB}_w, 2)$ adds ψ to $\texttt{KB}_w$, and $\pi(\texttt{KB}_w, i)$ does nothing for $i > 2$.

[1] There exist sensible orderings for constructing this graph because it may be infinite.

The graph of a reasoner $\mathcal{R}$ represents both the explicit and the accessible beliefs of $\mathcal{R}$. The explicit beliefs of $\mathcal{R}$ (beliefs$_{ex}$) in an epistemic situation $\mathtt{KB}_w$ are the sentences in $\mathtt{KB}_w$ (def. 1.2.1). In the example in figure A.1, $\mathcal{R}$ believes$_{ex}$ ϕ and nothing else in the epistemic situation $\mathtt{KB}_1$. The accessible beliefs of $\mathcal{R}$ (beliefs$_{ac}$) in an epistemic situation $\mathtt{KB}_w$ are the sentences in a $\mathtt{KB}_{w'}$ such that there exists an arrow from $\mathtt{KB}_w$ to $\mathtt{KB}_{w'}$ (def. 1.2.5). In other words, the set of beliefs$_{ac}$ of $\mathcal{R}$ in the situation $\mathtt{KB}_w$ is the set $\pi(\mathtt{KB}_w) = \bigcup \pi(\mathtt{KB}_w, i)$. In the example, $\mathcal{R}$ believes$_{ac}$ ϕ and ψ and nothing else in the epistemic situation $\mathtt{KB}_1$.

In the graph of $\mathcal{R}$, the epistemic situations $\mathtt{KB}_w$ may be interpreted as (im)possible worlds and the edges may be interpreted as an accessibility relation between (im)possible worlds[2]. The pair $\langle \mathtt{KB}_w, \mathtt{KB}_{w'} \rangle$ is in the accessibility relation π (i.e. $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$) iff there exists an i such that $\mathtt{KB}_{w'} = \pi(\mathtt{KB}_w, i)$. In this context, we may give a modal interpretation for the notion of beliefs$_{ac}$. In the modal interpretation, $\Diamond\phi$ is true of $\mathtt{KB}_w$ iff $\mathcal{R}$ believes$_{ac}$ ϕ at $\mathtt{KB}_w$ (i.e. $\phi \in \pi(\mathtt{KB}_w)$)[3]. The evaluation of $\Diamond\phi$ is, as it were, *external.* Since the $\mathtt{KB}_w$ are (im)possible worlds, the fact that $\Diamond\phi$ is true *of* $\mathtt{KB}_w$ does not entail that $\Diamond\phi \in \mathtt{KB}_w$. For example, $\mathcal{L}$ may not contain modal operators.

Let a function π be strictly ideal$_{ac}$ for a logic $\overset{\mathrm{x}}{\models}$ iff the following requirements hold:

(i) $\phi \in \pi(\mathtt{KB}_w)$ iff $\mathtt{KB}_w \overset{\mathrm{x}}{\models} \phi$;

(ii) $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$ for all $\mathtt{KB}_{w'}$ such that $\mathtt{KB}_{w'} \subseteq \pi(\mathtt{KB}_w)$.

In this context, there exists a relation between properties of $\overset{\mathrm{x}}{\models}$, properties of a function π which is strictly ideal$_{ac}$ for $\overset{\mathrm{x}}{\models}$, and the modal formulas π validates (see table A.1).

Logic $\overset{\mathrm{x}}{\models}$	Function π	Modal formulas
Reflexivity	Reflexivity	$\phi \to \Diamond\phi$ (T)
Cut	Transitivity	$\Diamond\Diamond\phi \to \Diamond\phi$ (4)
Cautious monotony	Euclidian relation	$\Diamond\phi \to \Box\Diamond\phi$ (5)

Table A.1: Relation between properties of a logic, properties of a function π which is strictly ideal$_{ac}$ for that logic, and the modal formulas that π validates.

[2] Impossible worlds are worlds in which the laws of classical logic fail (see Priest, 1997). For example, the law of excluded middle may fail in an epistemic situation $\mathtt{KB}$ because it it is the case that $(\phi \vee \neg\phi) \notin \mathtt{KB}$.

[3] In other words, $\Diamond\phi$ is true of $\mathtt{KB}_w$ iff $\exists i(\phi \in \pi(\mathtt{KB}_w, i))$ or $\exists w'(\pi(\mathtt{KB}_w, \mathtt{KB}_{w'}) \wedge \phi \in \mathtt{KB}_{w'})$.

The relation between properties of an accessibility relation π and the modal formulas which these π validates is well-known in the literature (Blackburn et al., 2001, p. 128-129). The relation between properties of logic $\overset{\mathtt{x}}{\models}$ and properties of a function π which is strictly ideal$_{ac}$ for $\overset{\mathtt{x}}{\models}$ is the following. Reflexivity for a logic $\overset{\mathtt{x}}{\models}$ is the property that $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_w$ for all $\mathtt{KB}_w$[4]. Reflexivity for function π is the property that $\pi(\mathtt{KB}_w, \mathtt{KB}_w)$ for all $\mathtt{KB}_w$. Suppose that π is reflexive. Then $\pi(\mathtt{KB}_w, \mathtt{KB}_w)$ for all $\mathtt{KB}_w$. Since $\pi(\mathtt{KB}_w, \mathtt{KB}_w)$, then $\mathtt{KB}_w \subseteq \pi(\mathtt{KB}_w)$ and $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_w$ (def. π). Therefore $\overset{\mathtt{x}}{\models}$ is reflexive. Suppose that $\overset{\mathtt{x}}{\models}$ is reflexive. Then $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_w$ for all $\mathtt{KB}_w$. Since $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_w$, then $\mathtt{KB}_w \subseteq \pi(\mathtt{KB}_w)$ and $\pi(\mathtt{KB}_w, \mathtt{KB}_w)$ (def. π). Therefore $\overset{\mathtt{x}}{\models}$ is reflexive. Therefore $\overset{\mathtt{x}}{\models}$ is reflexive iff π is reflexive.

Cut for a logic $\overset{\mathtt{x}}{\models}$ is the property that if $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w'}$ and $\mathtt{KB}_{w'} \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$, then $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$. Transitivity for function π is the property that if $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$ and $\pi(\mathtt{KB}_{w'}, \mathtt{KB}_{w''})$, then $\pi(\mathtt{KB}_w, \mathtt{KB}_{w''})$. Suppose that π is transitive. Suppose that $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w'}$ and $\mathtt{KB}_{w'} \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$. Since $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w'}$, then $\mathtt{KB}_{w'} \subseteq \pi(\mathtt{KB}_w)$ and $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$ (def. π). Since $\mathtt{KB}_{w'} \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$, then $\mathtt{KB}_{w''} \subseteq \pi(\mathtt{KB}_{w'})$ and $\pi(\mathtt{KB}_{w'}, \mathtt{KB}_{w''})$ (def. π). Then $\pi(\mathtt{KB}_w, \mathtt{KB}_{w''})$ (π is transitive). Since $\pi(\mathtt{KB}_w, \mathtt{KB}_{w''})$, then $\mathtt{KB}_{w''} \subseteq \pi(\mathtt{KB}_w)$ and $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$ (def. π). Therefore $\overset{\mathtt{x}}{\models}$ exhibits cut. Suppose that $\overset{\mathtt{x}}{\models}$ exhibits cut. Suppose that $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$ and $\pi(\mathtt{KB}_{w'}, \mathtt{KB}_{w''})$. Since $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$, then $\mathtt{KB}_{w'} \subseteq \pi(\mathtt{KB}_w)$ and $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w'}$ (def. π). Since $\pi(\mathtt{KB}_{w'}, \mathtt{KB}_{w''})$, then $\mathtt{KB}_{w''} \subseteq \pi(\mathtt{KB}_{w'})$ and $\mathtt{KB}_{w'} \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$ (def. π). Then $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$ ($\overset{\mathtt{x}}{\models}$ exhibits cut). Since $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$, then $\mathtt{KB}_{w''} \subseteq \pi(\mathtt{KB}_w)$ and $\pi(\mathtt{KB}_w, \mathtt{KB}_{w''})$ (def. π). Therefore π is transitive. Therefore $\overset{\mathtt{x}}{\models}$ exhibits cut iff π is transitive.

Cautious monotony (CM) for a logic $\overset{\mathtt{x}}{\models}$ is the property that if $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w'}$ and $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$, then $\mathtt{KB}_{w'} \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$. Function π is Euclidian when if $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$ and $\pi(\mathtt{KB}_w, \mathtt{KB}_{w''})$, then $\pi(\mathtt{KB}_{w'}, \mathtt{KB}_{w''})$. Suppose that π is Euclidian. Suppose that $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w'}$ and $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$. Since $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w'}$, then $\mathtt{KB}_{w'} \subseteq \pi(\mathtt{KB}_w)$ and $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$ (def. π). Since $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$, then $\mathtt{KB}_{w''} \subseteq \pi(\mathtt{KB}_w)$ and $\pi(\mathtt{KB}_w, \mathtt{KB}_{w''})$ (def. π). Then $\pi(\mathtt{KB}_{w'}, \mathtt{KB}_{w''})$ (π is Euclidian). Since $\pi(\mathtt{KB}_{w'}, \mathtt{KB}_{w''})$, then $\mathtt{KB}_{w''} \subseteq \pi(\mathtt{KB}_{w'})$ and $\mathtt{KB}_{w'} \overset{\mathtt{x}}{\models} \mathtt{KB}_{w''}$ (def. π). Therefore $\overset{\mathtt{x}}{\models}$ is CM. Suppose that $\overset{\mathtt{x}}{\models}$ is CM. Suppose that $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$ and $\pi(\mathtt{KB}_w, \mathtt{KB}_{w''})$. Since $\pi(\mathtt{KB}_w, \mathtt{KB}_{w'})$, then $\mathtt{KB}_{w'} \subseteq \pi(\mathtt{KB}_w)$ and $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w'}$ (def. π). Since $\pi(\mathtt{KB}_w, \mathtt{KB}_{w''})$, then $\mathtt{KB}_{w''} \subseteq \pi(\mathtt{KB}_w)$

[4] Where $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \mathtt{KB}_{w'}$ means that if $\phi \in \mathtt{KB}_{w'}$, then $\mathtt{KB}_w \overset{\mathtt{x}}{\models} \phi$.

and $\mathtt{KB}_w \overset{\mathrm{x}}{\models} \mathtt{KB}_{w''}$ (def. π). Then $\mathtt{KB}_{w'} \overset{\mathrm{x}}{\models} \mathtt{KB}_{w''}$ ($\overset{\mathrm{x}}{\models}$ is CM). Since $\mathtt{KB}_{w'} \overset{\mathrm{x}}{\models} \mathtt{KB}_{w''}$, then $\mathtt{KB}_{w''} \subseteq \pi(\mathtt{KB}_{w'})$ and $\pi(\mathtt{KB}_{w'}, \mathtt{KB}_{w''})$ (def. π). Therefore $\overset{\mathrm{x}}{\models}$ is CM iff π is Euclidian.

The relation between properties of an accessibility relation π and the modal formulas which these π validates is well-known in the literature. Axiom T ($\phi \rightarrow \Diamond\phi$) holds of a graph $\mathcal{R}$ iff the accessibility relation π is reflexive. Then axiom T holds iff π is reflexive iff $\overset{\mathrm{x}}{\models}$ is reflexive. In our interpretation, axiom T states that if $\mathcal{R}$ believes$_{ex}$ ϕ, then $\mathcal{R}$ believes$_{ac}$ ϕ (i.e. $\mathcal{R}$ accesses all of its beliefs$_{ex}$). Axiom 4 ($\Diamond\Diamond\phi \rightarrow \Diamond\phi$) holds of a graph $\mathcal{R}$ iff the accessibility relation π is transitive. Then axiom 4 holds iff π is transitive iff $\overset{\mathrm{x}}{\models}$ exhibits cut. In our interpretation, axiom 4 states that the set of beliefs$_{ac}$ of $\mathcal{R}$ does not increase over any amount of reasoning (i.e. if $\mathcal{R}$ does not believe$_{ac}\phi$, then $\mathcal{R}$ will not believe$_{ac}\phi$ after performing any number of inferences). Axiom 5 ($\Diamond\phi \rightarrow \Box\Diamond\phi$) holds of a graph $\mathcal{R}$ iff the accessibility relation π is Euclidian. Then Axiom 5 holds iff π is Euclidian iff $\overset{\mathrm{x}}{\models}$ is CM. In our interpretation, axiom 5 states that the set of beliefs$_{ac}$ of $\mathcal{R}$ does not decrease over any amount of reasoning (i.e. if $\mathcal{R}$ believes$_{ac}\phi$, then $\mathcal{R}$ will still believe$_{ac}\phi$ after performing any number of inferences)[5].

Classical logic $\overset{\mathrm{c}}{\models}$ is reflexive, exhibits cut, and is CM, then a function π which is strictly ideal$_{ac}$ for $\overset{\mathrm{c}}{\models}$ is reflexive, transitive, and Euclidean, and, therefore, an equivalence relation. This means that the set of beliefs$_{ac}$ ($\pi(\mathtt{KB})$) of a reasoner $\mathcal{R}$ with a function π which is strictly ideal$_{ac}$ for $\overset{\mathrm{c}}{\models}$ does not change with any amount of reasoning ($\pi(\mathtt{KB}) = \pi(\pi(\mathtt{KB}))$). This seems to be a good result. However, a well-behaved nonmonotonic logic $\overset{\mathrm{n}}{\models}$ is also reflexive, exhibits cut, and is CM (see Antonelli, 2005, p. 4-7). This means that the set of beliefs$_{ac}$ ($\pi(\mathtt{KB})$) of a reasoner $\mathcal{R}$ with a function π which is strictly ideal$_{ac}$ for $\overset{\mathrm{n}}{\models}$ also does not change with any amount of reasoning ($\pi(\mathtt{KB}) = \pi(\pi(\mathtt{KB}))$). This is not a good result because the main feature of nonmonotonic inference is that conclusions may be withdrawn after some amount of reasoning[6]. For this reason, the notion of beliefs$_{ac}$ is not adequate for expressing maximum rationality for nonmonotonic reasoning (see section 1.3).

[5] In this context, we may interpret the formula $\Box\Diamond\phi$ as stating that $\mathcal{R}$ is certain about ϕ. Axiom B ($\phi \rightarrow \Box\Diamond\phi$), which follow from T and 5, may be read as 'if $\mathcal{R}$ believes$_{ex}\phi$, then $\mathcal{R}$ is certain about ϕ'.

[6] Nonmonotonic frameworks are usually used to model a posteriori reasoning, whereas the model in this appendix is of a priori reasoning. But this feature of the notions of beliefs$_{ac}$ has consequences for both a priori and a posteriori reasoning (see argument in section 1.3).

Appendix B

Concepts

B.1 Computational complexity

For more information about theory of computational complexity, see Sipser (2012, p. 273).

B.1.1 Asymptotic analysis

The analysis of the complexity of algorithms is a part of computational complexity theory. The complexity of an algorithm is often understood as the rate in which the running time (time complexity) or the memory requirements (space complexity) of the algorithm grows in relation to the size of the input. The actual complexity of an algorithm usually requires assumptions about implementation. In order to abstract from these assumptions, an asymptotic analysis is often employed.

In an asymptotic analysis, where the complexity of the algorithm is determined for arbitrarily large inputs. The asymptotic analysis of an algorithm is often done under some extra abstractions. The first step is to abstract over the input of the algorithm, using some parameter to characterize the size of the input. The second step is to use some parameters to characterize the running time or memory requirements of the algorithm. In addition, the complexity of an algorithm is usually measured in relation to its worst-case scenario (the most demanding case for the algorithm).

Nevertheless, the exact complexity of an algorithm is often a complex expression. In most cases, the big-O notation is used for simplification (e.g. $f(n) = O(g(n))$, where n is the size of input). In big-O notation, only the highest order term of the expression of the complexity of the algorithm is considered and both the coefficient of that term and any lower order terms are disregarded. The idea is that, when the input grows arbitrarily, the highest order term dominates the other terms. For example, the function $f(n) = 7n^3+5n^2+10n+34$ has four terms and the highest order term is $7n^3$. Disregarding the coefficient 7, we say that $f(n) = O(n^3)$.

Definition B.1.1 (Big-O notation ($O(g(n))$)). Let f and g be functions $f, g : \mathcal{N} \to \mathcal{R}^+$. Then $f(n) = O(g(n))$ iff there exist positive integers c and n_0 such that for every integer $n \geq n_0$, $f(n) \leq cg(n)$ (Sipser, 2012, p. 227).

In this framework, algorithms are often classified under two categories: polynomial and exponential complexity. A polynomial algorithm is an algorithm whose complexity is expressed using a function of the kind $O(n^c)$, where c is a constant. An exponential algorithm is an algorithm whose complexity is expressed using a function of the kind $O(2^{n^c})$, where c is a constant greater than 0. While the actual complexity of an algorithm depends on low-level encoding details, where an algorithm falls on the polynomial/exponential dichotomy is independent of almost all such choices (for reasonable models of computation).

B.1.2 Complexity classes

In computational complexity theory, a complexity class is a class of computational problems of related resource-based computational complexity. The most important complexity classes are P, NP, and NP-complete. The class P (polynomial time) is the class of decision problems[1] that are solvable by a deterministic Turing machine in polynomial time. Meanwhile, NP (nondeterministic polynomial-time) is the class of decision problems for which a solution may be checked by a deterministic Turing machine in polynomial time, even if the solution cannot be found in polynomial time[2].

Since all problems that are solvable in polynomial time may have its solution checked in polynomial time, P $\subseteq$ NP. Whether P = NP is an open question[3]. Of special interest for the solution of this question is the class NP-complete. NP-complete is the class of

[1] A decision problem is a problem with a yes-or-no answer.

[2] To solve a decision problem is to return the correct answers for all instances of the problem; to check the solution of a decision problem is to solve for all 'yes' instances of the problem.

[3] Most theorists proceed on the assumption that P $\neq$ NP even if the claim is open to question.

decision problems such that all decision problems in NP are reducible to these problems in polynomial time[4]. In this context, if a problem in NP-complete is shown to be solvable in polynomial time, then all problems in NP are solvable in polynomial time and P=NP. An example of NP-complete problem is the satisfiability problem: given a sentence of propositional logic, is there an assignment of truth values to the propositional constants of the sentence that make it true?

[4] To reduce a problem to another is to create an algorithm which maps the solution for the first problem to the solution for the second.

Appendix C

Proofs

Proof C.1 (If $\mathcal{R}$ is maximally rational, then $\mathcal{R}$ is ideal). Suppose that $\mathcal{R}$ is maximally rational. Suppose that ϕ is a logical consequence of $\mathcal{R}$'s epistemic situation. Then all ideal reasoners in $\mathcal{R}$'s situation believe ϕ (def. 1.1.3, ideal). Then $\mathcal{R}$ ought to believe ϕ (def. 1.1.4, ought). Then $\mathcal{R}$ believes ϕ (def. 1.1.6, maximally). Therefore $\mathcal{R}$ believes all logical consequences of its epistemic situation. Suppose that $\mathcal{R}$ has a trivial set of beliefs. Then $\mathcal{R}$ may have a trivial set of beliefs (def. maximally). Then there exists an ideal reasoner in $\mathcal{R}$'s situation which has a trivial set of beliefs (def. 1.1.5, may). This is a contradiction (def. ideal). Therefore $\mathcal{R}$ has a nontrivial set of beliefs. $\mathcal{R}$ believes all logical consequences of its epistemic situation and $\mathcal{R}$ has a nontrivial set of beliefs. Therefore $\mathcal{R}$ is ideal (def. ideal).

Proof C.2 ($\mathcal{R}$ is maximally rational iff $\mathcal{R}$ is strictly ideal). Proof in two steps:

Suppose that $\mathcal{R}$ is maximally rational. Therefore $\mathcal{R}$ is an ideal reasoner (proof C.1). Suppose that $\mathcal{R}$ believes ϕ. Then $\mathcal{R}$ may and ought to believe ϕ (def. 1.1.6, maximally). Then there exists an ideal reasoner in $\mathcal{R}$'s epistemic situation and all ideal reasoners in $\mathcal{R}$'s situation believe ϕ (def. 1.1.5 and 1.1.4, may and ought). Then there exists an ideal reasoner in $\mathcal{R}$'s situation which believes ϕ only if ϕ is a consequence of $\mathcal{R}$'s situation. Then ϕ is a consequence of $\mathcal{R}$'s situation. Therefore $\mathcal{R}$ believes only the logical consequences of its epistemic situation. $\mathcal{R}$ is an ideal reasoner and $\mathcal{R}$ believes only the logical consequences of its epistemic situation. Therefore $\mathcal{R}$ is strictly ideal.

Suppose that $\mathcal{R}$ is strictly ideal. Suppose that $\mathcal{R}$ ought to believe ϕ. Then all ideal reasoners in $\mathcal{R}$'s situation believe ϕ (def. ought). $\mathcal{R}$ is an ideal reasoner in $\mathcal{R}$'s situation (def. 1.1.9, strictly). Then $\mathcal{R}$ believes ϕ. Therefore if $\mathcal{R}$ ought to believe ϕ, then $\mathcal{R}$ believes ϕ. Suppose that $\mathcal{R}$ believes ϕ. Then ϕ is a consequence of $\mathcal{R}$'s situation. Then all ideal reasoners in $\mathcal{R}$'s situation believe ϕ (def. 1.1.3, ideal). Then $\mathcal{R}$ ought to believe ϕ (def. ought). Therefore $\mathcal{R}$ ought to believe ϕ iff $\mathcal{R}$ believes ϕ. Also, if $\mathcal{R}$ believes ϕ, there is an ideal reasoner in $\mathcal{R}$' situation which believes ϕ ($\mathcal{R}$ itself). Therefore, if $\mathcal{R}$ believes ϕ, then $\mathcal{R}$ may believe ϕ. Therefore $\mathcal{R}$ believes ϕ iff $\mathcal{R}$ ought to believe ϕ and only if $\mathcal{R}$ may believe ϕ. Therefore, $\mathcal{R}$ is maximally rational.

Proof C.3 (if $t' > t$, then $\frac{t'-f}{t'+f+a} > \frac{t-f}{t+f+a}$). Suppose that $t' > t$. Multiplying both sides by $(f + a + f)$, it follows that $t'(f + a + f) > t(f + a + f)$. Distributing, it follows that $t'f + t'a + t'f > tf + ta + tf$. Adding $(-t'f - tf)$ to both sides, it follows that $t'f + t'a - tf > tf + ta - t'f$. Since $tf = ft$ and $t'f = ft'$ (commutativity), this is equivalent to $t'f + t'a - ft > tf + ta - ft'$. Adding $(tt' - ff - fa)$ to both sides, it follows that $tt' + t'f + t'a - ft - ff - fa > tt' + ft + ta - ft' - ff - fa$. Since $t't = tt'$ and $ft = tf$, this is equivalent to $t't + t'f + t'a - ft - ff - fa > tt' + tf + ta - ft' - ff - fa$ (commutativity). From distribution, it follows that $(t' - f)(t + f + a) > (t - f)(t' + f + a)$. Divinding both sides by $(t + f + a)(t' + f + a)$, it follows that $\frac{t'-f}{t'+f+a} > \frac{t-f}{t+f+a}$.

Proof C.4 (if $f' > f$, then $\frac{t-f'}{t+f'+a} < \frac{t-f}{t+f+a}$). Suppose that $f' > f$. Then $-f' < -f$. Multiplying both sides by $(t + t + a)$, it follows that $-f'(t + t + a) < -f(t + t + a)$. Distributing, it follows that $-f't - f't - f'a < -ft - ft - fa$. Since $-f't = -tf'$ and $-ft = -tf$, this is equivalent to $-tf' - f't - f'a < -tf - ft - fa$ (commutativity). Adding $(tf' + tf)$ to both sides, it follows that $tf - f't - f'a < tf' - ft - fa$. Adding $(tt + ta - f'f)$ to both sides, it follows that $tt + tf + ta - f't - f'f - f'a < tt + tf' + ta - ft - f'f - fa$. Since $-f'f = -ff'$, this is equivalent to $tt + tf + ta - f't - f'f - f'a < tt + tf' + ta - ft - ff' - fa$ (commutativity). From distribution, it follows that $(t - f')(t + f + a) < (t - f)(t + f' + a)$. Divinding both sides by $(t + f' + a)(t + f + a)$, it follows that $\frac{t-f'}{t+f'+a} < \frac{t-f}{t+f+a}$.

Proof C.5 ($\lim\limits_{x\to\infty} \frac{a^x}{x^b} = \infty$, where $a > 1$ and b are constants). Assume that b is an integer, since if the proof holds for $\lceil b \rceil$, it holds for b. We use induction on b. The assertion is true for $b \leq 0$, since as $x \to \infty$, $a^x \to \infty$ while x^b is constant ($b = 0$) or $x^b \to 0$ ($b < 0$). Suppose that the assertion fails for some b and choose b minimal such that it fails. Given the last remarks, $b \geq 1$ and $\lim_{x\to\infty} x^b = \infty$. Since the theorem holds for $b-1$, $\lim_{x\to\infty} \frac{a^x}{x^{b-1}} = \infty$. Let $f(x) = a^x$ and $g(x) = x^b$. L'Hopital's rule tells us that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)} = \lim_{x\to\infty} \frac{ln(a)a^x}{bx^{b-1}} = \left(\frac{ln(a)}{b}\right) \lim_{x\to\infty} \frac{a^x}{x^{b-1}} = \left(\frac{ln(a)}{b}\right)\infty = \infty$.

Proof C.6 ($\lim\limits_{x\to\infty} \frac{x^a}{(log_c(x))^b} = \infty$, where a, b, c are constants). Since the ratio of $(log_c(x))^b$ to $(ln(x))^b$ is a nonzero constant ($(log_c(e))^b$), we may restrict to the natural algorithm ($c = e$). By the same reasoning in proof C.5, we assume that b is an integer. We use induction on b. As in proof C.5, this theorem is obviously true if $b \leq 0$. If the theorem fails for some b, choose b minimal such that it fails. By the remark above, $b \geq 1$, and $\lim_{x\to\infty}(ln(x))^b = \infty$. Since the theorem holds for $b-1$, $\lim_{x\to\infty} \frac{x^a}{ln(x)^{b-1}} = \infty$. Let $f(x) = x^a$ and $g(x) = (ln(x))^b$. L'Hopital's rule tells us that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)} = \lim_{x\to\infty} \frac{ax^{a-1}}{blb(x)^{b-1}(1/x)} = \left(\frac{a}{b}\right) \lim_{x\to\infty} \frac{x^a}{ln(x)^{b-1}} = \left(\frac{a}{b}\right)\infty = \infty$.

Bibliography

Aaronson, S. (2011). Why philosophers should care about computational complexity. *CoRR*, abs/1108.1791.

Almeder, R. (1998). *Harmless Naturalism: The Limits of Science and the Nature of Philosophy*. Peru, Illinois: Open Court.

Alston, W. P. (1985). Concepts of epistemic justification. *The Monist*, 68(1):57–89.

Antonelli, A. (2005). *Grounded Consequence for Defeasible Logic*. Cambridge University Press.

Antonelli, A. (2015). The completeness of classical propositional and predicate logic (c2p2l). URL: <http://aldo-antonelli.org/Papers/C2P2L.pdf>.

Baltag, A. and Smets, S. (2005). Qs7 complete axiomatizations for quantum actions. *International Journal of Theoretical Physics*, 44(12):2267–2282.

Bealer, G. (1998). Intuition and the autonomy of philosophy. In *Rethinking Intuition*, pages 201–239.

Bekenstein, J. D. (1981). Universal upper bound on the entropy-to-energy ratio for bounded systems. *Physical Review*, 23(2):287–298.

Bernstein, M. (2007). Experimental philosophy meets experimental design: 23 questions. In *MidSouth Philosophy Conference*.

Binkley, R. (1968). The surprise examination in modal logic. *Journal of Philosophy*, 65(5):127–136.

Bishop, M. and Trout, J. D. (2005). *Epistemology and the Psychology of Human Judgment*. New York: Oxford University Press.

Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press, New York, NY, USA.

Böerger, E., Grädel, E., and Gurevich, Y. (1997). *The Classical Decision Problem*. Springer-Verlag, Berlin.

Bogdanov, A. and Trevisan, L. (2006). Average-case complexity. *CoRR*, abs/cs/0606037.

Boghossian, P. A. (2003). The normativity of content. *Philosophical Issues*, 13(1):31–45.

Bohm, D. (1952). A suggested interpretation of the quantum theory in terms of "hidden" variables. *Phys. Rev.*, 85:166–179.

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Harvard University Press.

BonJour, L. (1994). Against naturalized epistemology. *Midwest Studies in Philosophy*, XIX:283–300.

Boolos, G., Burgess, J., P., R., and Jeffrey, C. (2007). *Computability and Logic*. Cambridge University Press.

Bovens, L. and Hartmann, S. (2004). *Bayesian Epistemology*. Oxford: Oxford University Press.

Buckwalter, W. and Stich, S. (2011). Gender and the philosophy club. *The Philosophers' Magazine*, 52:60–65.

Bykvist, K. and Hattiangadi, A. (2007). Does thought imply ought? *Analysis*, 67(296):277–285.

Bykvist, K. and Hattiangadi, A. (2013). Belief, truth, and blindspots. In Chan, T., editor, *The Aim of Belief*. Oxford University Press.

Carnap, R. (1946). Modalities and quantification. *The Journal of Symbolic Logic*, 11(2):33–64.

Chalmers, D. (2002). *Conceivability and Possibility*, chapter Does Conceivability Entail Possibility?, pages 145–200. Oxford University Press.

Chalmers, D. (2012). *Constructing the World*. Oxford University Press.

Chalmers, D. J. (2010). *The Character of Consciousness*. Oxford University Press.

Chisholm, R. (1977). *Theory of Knowledge, 2nd ed.* Englewood Cliffs: Prentice Hall.

Chisholm, R. M. (1982). A version of knowledge. In *The Foundations of Knowing*. Brighton, U.K.: The Harvester Press.

Church, A. (1936a). A note on the entscheidungsproblem. *Journal of Symbolic Logic*, 1:40–41.

Church, A. (1936b). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58:345–363.

Cohnitz, D. (2012). *New Waves in Philosophical Logic*, chapter The Logic(s) of Modal Knowledge. Palgrave Macmillan.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42:393–405.

Cummins, R. C. (1998). Reflection on reflective equilibrium. In DePaul, M. and Ramsey, W., editors, *Rethinking Intuition*, pages 113–128. Rowman & Littlefield.

David, M. (2001). Truth as the epistemic goal. In *Knowledge, Truth, and Duty*, pages 151–169. New York: Oxford University Press.

Davidson, D. (1965). Theories of meaning and learnable languages. In Bar-Hillel, Y., editor, *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science*, pages 3–17. North-Holland.

Demey, L., Kooi, B., and Sack, J. (2014). Logic and probability. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2014 edition.

Dennett, D. (1984). Cognitive wheels: The frame problem of ai. In Hookway, C., editor, *Minds, Machines and Evolution*, pages 129–151. Cambridge University Press, Cambridge.

Duc, H. N. (1995). Logical omniscience vs. logical ignorance on a dilemma of epistemic logic. *Lecture Notes in Computer Science*, 990:237–248.

Easwaran, K. (2013). Expected Accuracy Supports Conditionalization—and Conglomerability and Reflection. *Philosophy of Science*, 80(1):119–142.

Easwaran, K. (2015). Dr. truthlove or: How i learned to stop worrying and love bayesian probabilities*. *Noûs*, pages 1–38.

Evnine, S. (2008). Modal epistemology: Our knowledge of necessity and possibility. *Philosophy Compass*, 3/4:664–684.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. (1995). *Reasoning about Knowledge.* Cambridge: MIT Press.

Field, H. (1998). Epistemological nonfactualism and the a prioricity of logic. *Philosophical Studies*, 92(1/2):1–24.

Fitelson, B. and Easwaran, K. (2015). Accuracy, Coherence and Evidence. In Gendler, T. S. and Hawthorne, J., editors, *Oxford Studies in Epistemology, Volume 5*. Oxford: Oxford University Press.

Foley, R. (1993). *Working Without a Net: A Study of Egocentric Epistemology.* Oxford University Press.

Foley, R. (1994). Quine and naturalized epistemology. *Midwest Studies in Philosophy*, XIX:243–260.

Ford, K., Glymour, C., and Hayes, P. (2006). *Thinking about Android Epistemology.* AAAI Press & The MIT Press.

Fumerton, R. (1994). Skepticism and naturalistic epistemology. *Midwest Studies in Philosophy*, XIX:321–340.

Gärdenfors, P. (1988). *Knowledge in Flux. Modelling the Dymanics of Epistemic States.* Cambridge, MA: MIT Press.

Garfield, J. L. (1990). The dog: Relevance and rationality. In *Truth or Consequences*, pages 97–109. Kluwer.

Gelfond, M., Rushton, J. N., and Zhu, W. (2006). Combining logical and probabilistic reasoning. In *AAAISS*.

Gendler, T. S. and Hawthorne, J. (2002). *Conceivability and Possibility*. Oxford University Press.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis*, 23:121–123.

Giunchiglia, E. and Giunchiglia, F. (2001). Ideal and real belief about belief. *J. Logic Computat*, 11:157–192.

Gödel, K. (1930). Die vollständigkeit der axiome des logischen funktionenkalküls. *Monatshefte für Mathematik und Physik*, 37(1):349–360.

Gold, E. M. (1965). Limiting recursion. *The Journal of Symbolic Logic*, 30:28–48.

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge: Harvard University Press.

Goldman, A. (1992). *Liaisons: Philosophy Meets the Cognitive and Social Sciences*. Cambridge: MIT Press.

Gottlob, G. (1994). From carnap's modal logic to autoepistemic logic. In MacNish, C., Pearce, D., and Pereira, L. M., editors, *Logics in Artificial Intelligence(JELIA)*, pages 1–18. Springer, Berlin, Heidelberg.

Greaves, H. and Wallace, D. (2006). Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind*, 115(459):607–632.

Grim, P. (1988). Logic and limits of knowledge and truth. *Noûs*, 22(3):341–367.

Haack, S. (1993). *Evidence and Inquiry: Towards Reconstruction in Epistemology*. Oxford: Blackwell.

Halpern, J. Y. and Moses, Y. (1985). Towards a theory of knowledge and ignorance: Preliminary report. In *Logics and Models of Concurrent Systems*, pages 459–476. Springer-Verlag.

Halpern, J. Y., Samet, D., and Segev, E. (2009). Defining knowledge in terms of belief: The modal logic perspective. *Review of Symbolic Logic*, 2(3):469–487.

Haugeland, J. (1982). Weak supervenience. *American Philosophical Quarterly*, 19(1):93–103.

Heckerman, D. E. and Shortliffe, E. H. (1992). From certainty factors to belief networks. *Artif. Intell. Med.*, 4(1):35–52.

Hendricks, V. F. and Symons, J. (2006). Where's the bridge? epistemology and epistemic logic. *Philosophical Studies*, 128(1):137–167.

Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell University Press.

Inoue, S. and Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. *Current Biology*, 17(23):R1004–R1005.

Joyce, J. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65:575–603.

Joyce, J. M. (2011). The development of subjective bayesianism. In Gabbay, D. M., Hartmann, S., and Woods, J., editors, *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 415–475. Elsevier.

Kelly, K. (1988). Artificial intelligence and effective epistemology. In Fetzer, J., editor, *Aspects of Artificial Intelligence*, volume 1 of *Studies in Cognitive Systems*, pages 309–322. Springer Netherlands.

Kelly, K. T. (1990). Learning theory and descriptive set theory. *Department of Philosophy. Paper 464. http://repository.cmu.edu/philosophy/464/.*

Kim, J. (1988). *Philosophical Perspectives*, chapter What is Naturalized Epistemology?, pages 2:381–406. Asascadero, CA: Ridgeview Publishing Co.

Kirk, R. (2012). Zombies. URL: <http://plato.stanford.edu/entries/zombies/>.

Kitcher, P. (1993). *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. Oxford University Press.

Kornblith, H. (1994). *Naturalizing Epistemology, 2nd Edition*. Cambridge: MIT Press.

Kraus, S., Lehman, D., and Magidor, M. (1990). Nonmonotonic reasoning, preferential models and comulative logics. *Artificial Intelligence*, 44(1-2):167–207.

Kripke, S. (1959). A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24(1):1–14.

Lehrer, K. (2000). *Theory of Knowledge*. Westview Press.

Leitgeb, H. and Pettigrew, R. (2010a). An Objective Justification of Bayesianism I: Measuring Inaccuracy. *Philosophy of Science*, 77(2):201–235.

Leitgeb, H. and Pettigrew, R. (2010b). An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science*, 77(2):236–272.

Lokhorst, G. (2000). Why i am not a super-turing machine. In *Hypercomputation Workshop, University College, London.*

Louden, K. C. (2002). *Programming Languages: Principles and Practice.* Course Technology.

Maher, P. (2002). Joyce's Argument for Probabilism. *Philosophy of Science*, 69(1):73–81.

Makinson, D. (1994). General patterns in nonmonotonic reasoning. In Gabbay, D., Hogger, C., and Robinson, J., editors, *Handbook of Logic in Artificial Intelligence and Logic Programming, volume 3*, chapter General patterns in nonmonotonic reasoning, pages 35–110. Oxford and New York: Oxford University Press.

McCarthy, J. (1980). Circumscription - a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39.

McCauley, R. N. (1988). Epistemology in an age of cognitive science. *Philosophical Psychology*, 1(2):143–152.

Menzies, P. (1998). *European Review of Philosophy, Volume 3: Response-Dependence*, chapter Possibility and Conceivability: A Response-Dependent Account of Their Connections, pages 255–277. Stanford.

Moore, R. C. (1985). Semantic considerations on nonmonotonic logic. *Artificial Intelligence*, 25:75–94.

Moser (1985). *Empirical Justification.* Dordrecht: Reidel.

Nagel, J. (2012). Intuitions and experiments: A defense of the case method in epistemology. *Philosophy and Phenomenological Research.*

Neta, R. (2013). Does the epistemic 'ought' imply the cognitive 'can'? In Fairweather, A. and Flanagan, O., editors, *Naturalizing Epistemic Virtue*, chapter Does the Epistemic 'Ought' Imply the Cognitive 'Can'?'. Cambridge University Press.

Nozick, R. (1981). *Philosophical Explanations.* Harvard University Press.

Oddie, G. (1997). Conditionalization, Cogency, and Cognitive Value. *British Journal for the Philosophy of Science*, 48(4):533–541.

Ojeda, A. E. (2012). *A Computational Introduction to Linguistics: Describing Language in Plain Prolog.*

Parent, T. (2014). An objection to the laplacean chalmers. URL: <http://www.unc.edu/ tparent/Chalmers.pdf>. Unpublished draft.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–288.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, California.

Pettigrew, R. (2015a). *Accuracy, Chance, and the Laws of Credence.* Oxford University Press.

Pettigrew, R. (2015b). Jamesian epistemology formalised: an explication of 'the will to believe'. *Episteme.*

Plantinga, A. (1993). *Warrant: The Current Debate.* Oxford University Press.

Pohl, R. (2004). *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory.* Psychology Press.

Pollock, J. (1987). Defeasible reasoning. *Cognitive Science*, 11(4):481–518.

Pollock, J. (1998). Procedural epistemology. In Bynum, T. and Moor, J., editors, *The Digital Phoenix: How Computers are Changing Philosophy*, pages 17–36. Oxford: Blackwell.

Pollock, J. and Cruz, J. (1999). *Contemporary Theories of Knowledge 2nd Edition.* Lanham, MD: Rowman & Littlefield.

Pollock, J. L. (1990). *Nomic Probability and the Fundations of Induction.* Oxford: Oxford University Press.

Pollock, J. L. (1995). *Cognitive Carpentry: a blueprint for how to build a person.* The MIT Press.

Pollock, J. L. (2008). *Reasoning: Studies of Human Inference and its Foundations*, chapter Defeasible reasoning. Cambridge University Press.

Pollock, J. L. (2009). *Argumentation and Artificial Intelligence*, chapter A recursive semantics for defeasible reasoning. Springer.

Poole, D., Mackworth, A. K., and Goebel, R. (1998). *Computational intelligence: A logical approach.* Oxford University Press, Oxford, UK.

Priest, G. (1997). Impossible worlds - editor's introduction. *Notre Dame Journal of Formal Logic*, 38(4):481–487.

Putnam, H. (1965). Trial and error predicates and the solution to a problem of mostowski. *The Journal of Symbolic Logic*, 30(1):49–57.

Quine, W. V. (1969). *Ontological Relativity and Other Essays*, chapter Epistemology Naturalized, pages 69–90. New York: Columbia UP.

Quine, W. V. (1992). *Pursuit of Truth*, volume 103. Harvard University Press.

Quine, W. V. (1998). *The Philosophy of W. V. Quine*, chapter Reply to Morton White, pages 663–665. Chicago: Open Court.

Raleigh, T. (2013). Belief norms & blindspots. *Southern Journal of Philosophy*, 51(2):243–269.

Ramsey, F. P. (1926). Truth and probability. In Braithwaite, R. B., editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. McMaster University Archive for the History of Economic Thought.

Ray, C. (1991). *Time, Space, and Philosophy.* Routledge.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, 13:81–132.

Russell, B. (2013). A priori justification and knowledge. URL: <http://plato.stanford.edu/entries/apriori/>.

Russell, S. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach 2nd Edition.* Prentice Hall.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach 3rd Edition.* Upper Saddle River, EUA: Prentice-Hall.

Salmon, N. (1989). The logic of what might have been. *Philosophical Review*, 98(1):3–34.

Sardina, S. and Vassos, S. (2005). The wumpus world in indigolog: A preliminary report. In *In Proc. NRAC-05*, pages 90–95.

Schurz, G. (2000). *Rudolf Carnap's Modal Logic.* University of Erfurt.

Sellars, W. (1956). *Empiricism and the Philosophy of Mind.*

Shafer, G. (2010). A betting interpretation for probabilities and dempster-shafer degrees of belief. *International Journal of Approximate Reasoning.*

Shapiro, S. C. and Kandefer, M. (2005). A sneps approach to the wumpus world agent or cassie meets the wumpus. In *Proceedings of The Sixth Workshop on Nonmonotonic Reasoning, Action, and Change.*

Siegel, H. (1996). Naturalism, instrumental rationality, and the normativity of epistemology. *Protosociology*, 8/9:97–110.

Sipser, M. (2012). *Introduction to the Theory of Computation, Third Edition.* Cengage Learning.

Sobel, J. (1989). Utility theory and the bayesian paradigm. *Theory and Decision*, 26(3):263–293.

Sorensen, R. A. (1988). *Blindspots.* Oxford University Press.

Sosa, E. (1985). Knowledge and intellectual virtue. *The Monist*, 68(2):226–245.

Sosa, E. (1999). How must knowledge be modally related to what is known? *Philosophical Topics*, 26(1/2):373–384.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, 132:99–107.

Spohn, W. (1988). *Causation in Decision, Belief Change, and Statistics, vol.II*, chapter Ordinal Conditional Functions: a dynamic theory of epistemic states, pages 105–134. Dordrecht: Kluwer.

Spohn, W. (2002). A brief comparison of pollock's defeasible reasoning and ranking functions. *Synthese*, 131:39–56.

Stalnaker, R. (1991). The problem of logic omnicience, i. *Synthese*, 89(3):425–440.

Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies*, 128(1):169–199.

Stanalker, R. (1994). Nonmonotonic consequence relations. *Fundamenta Informaticæ*, 21:7–21.

Sun, R. (2008). *The Cambridge Handbook of Computational Psychology.* Cambridge University Press, New York.

Swain, S., Alexander, J., and Weinberg, J. (2008). The instability of philosophical intuitions: running hot and cold on truetemp. *Philosophy and Phenomenoloical Research*, 76(1):138–155.

Talbott, W. (2013). Bayesian epistemology. URL: <http://plato.stanford.edu/entries/epistemology-bayesian/>.

Teller, P. (1976). Conditionalization, observation, and change of preference. In Harper, W. and Hooker, C., editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, volume 6a of *The University of Western Ontario Series in Philosophy of Science*, pages 205–259. Springer Netherlands.

Tesfatsion, L. (2006). *Handbook of Computational Economics*, chapter Agent-Based Computational Economics: A Constructive Approach to Economic Theory, pages 831–880.

Thielscher, M. (2005). A flux agent for the wumpus world.

Tononi, G. and Koch, C. (2008). The neural correlates of consciousness. *Annals of the New York Academy of Sciences*, 1124(1):239–261.

Treanor, N. (2013). The measure of knowledge. *Noûs*, 47(3):577–601.

van Ditmarsch, H., van der Hoek, W., and Kooi, P. (2006). *Concurrent Dynamic Epistemic Logic.* Synthese Library Series. Dordrecht: Springer.

von Wright, G. (1986). Is and ought. In Doeser, M. and Kraay, J., editors, *Facts and Values*, volume 19 of *Martinus Nijhoff Philosophy Library*, pages 31–48. Springer Netherlands.

von Wright, G. H. (1951). Deontic logic. *Mind*, 60(237):1–15.

Wallace, R. J. (2014). Practical reason. URL: <http://plato.stanford.edu/entries/practical-reason/>.

Weinberg, J. S., Nichols, S., and Sitich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29(1):429–460.

Weirich, P. (2010). Does collective rationality entail efficiency? *Logic Journal of IGPL*, 18(2):308–322.

Whiting, D. (2010). Should i believe the truth? *Dialectica*, 64(2):213–224.

Williamson, J. (2002). Probability logic. In D. Gabbay, R. Johnson, H. J. O. and (, J. W., editors, *Handbook of the Logic of Argument and Inference: the Turn Toward the Practical*, pages 397–424. Amsterdam: Elsevier.

Williamson, J. (2010). *In Defense of Objective Bayesianism*. Oxford: Oxford University Press.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.

Wrenn, C. B. (2006). Epistemology as engeneering? *Theoria*, 72(1):60–79.

Danilo Fraga Dantas
December 2016
Graduate Program in Philosophy

ALMOST IDEAL:
Computational Epistemology and the Limits of Rationality for Finite Reasoners

Abstract

The notion of an ideal reasoner has several uses in epistemology. Often, ideal reasoners are used as a parameter of (maximum) rationality for finite reasoners (e.g. humans). However, the notion of an ideal reasoner is normally construed in such a high degree of idealization (e.g. infinite/unbounded memory) that this use is unadvised. In this dissertation, I investigate the conditions under which an ideal reasoner may be used as a parameter of rationality for finite reasoners. In addition, I present and justify the research program of computational epistemology, which investigates the parameter of maximum rationality for finite reasoners using computer simulations.

In chapter 1, I investigate the use ideal reasoners for stating the maximum (and minimum) bounds of rationality for finite reasoners. I propose the notion of a strictly ideal reasoner which coincides with the notion of maximum rationality. The notion of a strictly ideal reasoner is relative to a logic and a notion of beliefs (explicit, implicit, etc). I argue that, for some relevant logics, a finite reasoner may only approach maximum rationality at the limit of a reasoning sequence (stable beliefs)[1]. In chapter 2, I investigate the use of ideal reasoners in the zombie argument against physicalism (Chalmers, 2010). This notion is used in the principle that ideal negative conceivability entails possibility. The conclusion is that the zombie argument is neither an a priori nor a conclusive argument against physicalism. In chapter 3, I investigate the notion of maximum (and minimum) *epistemic* rationality for finite reasoners. Epistemic rationality is often related

[1] Informally, the reasoning sequence of a reasoner is how the reasoner would reason from the available information if it had enough cognitive resources (e.g. time for reasoning). The notion of a stable belief is related to the beliefs that the reasoner would hold at the limit of a reasoning sequence.

to maximizing true beliefs and minimizing false beliefs. I argue that most of the existing models of maximum epistemic rationality have problems in dealing with blindspots and propose a model in terms of the maximization of a function g, which evaluates sets of beliefs regarding truth/falsehood. However, function g may only be maximized at the limit of a reasoning sequence. In chapter 4, I argue that if maximum (epistemic) rationality for finite reasoners must be understood in terms of the limit of a reasoning sequence, then issues about the computational complexity of reasoning are relevant to epistemology. Then I propose the research program of computational epistemology, which uses computer simulations for investigating maximum (epistemic) rationality for finite reasoners and considers the computational complexity of reasoning. In chapter 5, I provide an example of an investigation in computational epistemology. More specifically, I compare two models of maximum rationality for situations of uncertain reasoning: theory of defeasible reasoning (Pollock, 1995) and Bayesian epistemology (Joyce, 2011).