A Momentary Lapse Of Reason

Comparing Davidson's account of rationality and reasoning to that of van Lambalgen & Stenning

MSc Thesis (*Afstudeerscriptie*)

written by

Olga Grigoriadou
(born June 1st, 1980 in Athens, Greece)

under the supervision of **Prof.dr Martin Stokhof** and **Prof.dr Michiel van Lambalgen**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
October 9th, 2009	Prof.dr Martin Stokhof
	Prof.dr Michiel van Lambalgen
	Dr. Paul J. A. Dekker
	Prof.dr Frank Veltman

Institute for Logic, Language and Computation

“One slip, and down the hole we fall
It seems to take no time at all
A momentary lapse of reason
That binds a life for life
A small regret, you will never forget,
There’ll be no sleep in here tonight”

Pink Floyd

Acknowledgements

To begin with, I want to thank Martin Stokhof without whom this thesis would have neither been started nor finished. For me he is a living example of a good teacher: knowing what the students' needs are at any moment—which is a kind of magic—and giving the clearest explanations even for the most complicated philosophical ideas. Coming from a mathematics background, I never imagined I would meet someone that would explain philosophical ideas with such an inviting way to make me want to investigate some of them deeper, drifting away from the certainty of mathematics. I don't know if I have managed to go deep enough into my investigations, and I certainly felt quite nostalgic about mathematics several times along the way, but I am grateful to Martin for all I have learned from him about philosophy.

I also want to thank Michiel van Lambalgen who, when hearing about my interest in the issue of rationality, suggested to look at some of his recent work for my thesis. I was thus introduced to the very interesting and entirely new to me fields of psychology of reasoning and cognitive science. While reading Michiel's work I was inspired to continue my research within mathematics education, and I discovered new possibilities for bringing together logic and research in education. His comments were always useful and this thesis would not have been completed without his valuable help.

I am thankful to the Dutch government for supporting working students. Studiefinanciering was the basic sponsor of my studies. Although it was at times too hard to keep up with my studies while working, I have certainly gained useful experience out there in the real world through waitressing, babysitting, playing music, and working as a student assistant in various projects at the University of Amsterdam. I want to thank all my employers for helping me stay alive.

I owe a big thanks to Ansten Mørch Klev, Edgar Andrade-Lotero, and Wolter Kaper who read earlier versions of this thesis and commented on it. Their comments and the discussions we had helped me gain a better insight into the ideas I am dealing with in this thesis. I am also deeply grateful to Tikitu for proofreading my thesis, for being my LaTeX-angel all these years and, most importantly, for being there for me.

I should not forget to thank Pink Floyd, first of all for being one of the greatest bands of all times, and secondly for inspiring me for coming up with a thesis title.

Finally, I would like to thank with all my heart my parents for always supporting me, in any way they can, whatever my decisions are in life. To them I want to dedicate this thesis.

Contents

Abstract

I examine and compare two accounts of rationality and logical reasoning: that of Donald Davidson on the one hand and that of Michiel van Lambalgen and Keith Stenning on the other. My first aim is to examine the differences between these two theories, and the criticisms that each account can bring against the other. My second aim is to explore how a combination of the two accounts might result in a more complete account of rationality and logical reasoning than what is offered by either one alone. I conclude that the two accounts, although they at first seem to differ entirely from each other, have much in common, and that there are important elements in each that might be useful for the other (e.g. Davidson's Principle of Charity, or the distinction between reasoning *to* and *from* an interpretation introduced by van Lambalgen and Stenning). A combination of the two accounts may also have interesting implications for research in the field of education.

Chapter 1

Introduction

What is rationality? What allows us to claim that one person is rational while some other person is not? Are there any global norms for deciding who is rational? And, what role does logic play in explaining and describing human rationality? These are some of the—as yet unanswered—philosophical questions that lead to the writing of this thesis. Various claims have been made from antiquity to our time on the nature of rationality and on how humans reason. Traditionally, classical logic has been considered as the normative system which determines the "right" way of reasoning whether within a formal system or within everyday communication.

Classical logic consists of a formal language, a semantics for the language, and certain rules of valid inference. The semantics of classical propositional logic is described by tables of the truth values '0' and '1', and its formal language can represent adequately only a fragment of natural language. It seems strange then, in terms of semantics, to claim that classical logic determines the correct way to perform everyday reasoning and the norms of rationality. In fact, as we will see in this thesis, presupposing that classical logic is the logic which explains human reasoning may lead to paradoxical conclusions, exactly because the semantics of classical logic can express only a part of natural language and often fails to explain everyday human reasoning.

Depriving classical logic of the descriptive and normative power that has been assigned to it for centuries is not an easy task since many will immediately object. However, the fact that during the second half of the 20th century many non-classical logics appeared which could adequately describe fragments of natural language that classical logic fails to describe, may give a hint to those who insist that classical logic possesses a special position among other logics. The possibility of a multiplicity of logics for describing and understanding human rationality and logical reasoning is an interesting one, yet the threat of relativism arises. Can we claim that different logics can describe human rationality at a local level better than classical logic does when considered as a universal normative system, without implying that anything goes in logical reasoning? These issues will be further examined in this thesis by looking at two specific accounts of rationality and logical reasoning in the area of philosophy of language and cognitive science; that of Donald Davidson on the one hand, and that of Michiel van Lambalgen and Keith Stenning on the other. A central idea that underlies both accounts is that all humans are rational in the sense that any instance of irrational behaviour can be attributed either to the fact that subjects some times apply different norms to reason about similar situations or to momentary lapses of reason (hence the title of this thesis), rather than to irrationality.

In the following section, we will briefly look at the until now established views of what rationality is and of how we can come to understand human reasoning. I will give a brief overview of this background as it has been until the end of the 20th century, and then we will look closely at the specific aims and outline of this thesis.

1.1 The background

1.1.1 Definitions of rationality

The standard picture of rationality is that 'to be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory, and so forth' [Ste96, p. 4]. According to this picture, there are certain rules which define rationality and any human being that deviates from these rules—the so called norms—is deemed to be irrational. One may then wonder, is there a definite and exhaustive list of such rules which define rationality? And what is the correct way, if one exists, to discover them?

Before looking at the above questions and the attempts that have been made during the last centuries towards answering them (we leave this for section 1.1.2), let us look at some ways that rationality has been defined in the past.

In [Noz93, p. 64] we read:

> It is natural to think of rationality as a goal-directed process. (This applies to both rationality of action and rationality of belief.) The stereotype of behaviour in traditional societies is that people act a certain way because things always have been done that way. In contrast, rational behaviour is aimed at achieving the goals, desires, and ends that people have. On this instrumental conception, rationality consists in the effective and efficient achievement of goals, ends, and desires. [. . .] If rational procedures are ones that reliably achieve specified goals, an action is rational when it is produced by such a procedure, and a person is rational when he appropriately uses rational procedures.

In the MIT Encyclopedia of Cognitive Science a similar definition is given for rational agency:

> [T]he agent must have a means-end competence to fit his actions or decisions, according to its beliefs or knowledge representations, to its desires or goal-structure.

Both the above definitions stress the relation of rationality to the way people act in order to achieve their goals. A person is considered to be rational when he acts driven by his desires and restricted by what he believes/knows to be the best means to satisfy his desires. For example, if you know that in front of you there is a fridge full of your favourite kind of beer and you feel the desire to drink one, it would be an irrational move to start walking backwards away from the fridge given that you have the desire to drink a beer, that you know in front of you there is a fridge full of your favourite beer, and that you believe that the best and shortest way to drink a beer is to open the fridge.

Another kind of definition for rationality, focusing more in the reasoning processes of humans rather than on their actions, is given in [Ste96, p. 2] where we read that there are three ways to be rational resulting from the apparent tension between the Aristotelian idea of a 'rational animal' and the Freudian 'irrational humans':

1. Rationality is reasoning ability; a human is then rational when he is able to consciously and explicitly reason, give arguments which support his beliefs etc.

2. Rationality is 'perfect' reasoning; since humans often make mistakes in reasoning, humans are irrational.

3. Humans are rational if the only mistakes they make in reasoning 'are attributable to interferences with human reasoning rather than to human reasoning ability itself' [Ste96, p. 3].

Under the first of the above categories, we have those who believe that human beings *are* rational and follow certain norms for reasoning which makes them deserve the characterization. On the other hand, we have those who believe that the only correct reasoning is 'perfect' reasoning which excludes any kind of divergence from what the norms of reasoning define as good reasoning. Thus they cannot allow any mistakes: human beings fall into mistakes, human beings are irrational. Lastly, there are those who believe that humans *are* rational regardless of the errors they might sometimes make. These errors, they claim, do not occur as a result of lack of reasoning ability but they are due to other reasons; they are *performance* errors and not *competence* errors.[1]

The last view of rationality I want to present is the one found in [EO96, p. 8] where Evans and Over distinguish between two ways of looking at rationality: a *personal* and an *impersonal*. They define them as follows:

> *Rationality*$_1$: Thinking, speaking, reasoning, making a decision, or acting in a way that is generally reliable and efficient for achieving one's goals.
>
> *Rationality*$_2$: Thinking, speaking, reasoning, making a decision, or acting when one has a reason for what one does sanctioned by a normative theory.

[1] Performance errors are mere mistakes of reasoning due to various interfering factors such as lack of sleep or concentration, alcohol or drug use etc. Competence errors on the other hand, are errors which are systematic violations of the norms.

The first of the above definitions is what they call a personal view of rationality. In this view, a person's acts are rational if they are useful in achieving the person's individual goals. The second view of rationality, the impersonal one, coincides with the standard view of rationality—the very first introduced in this section—which wants people to be called rational when they have the correct reasons according to the norms of logic (or some other normative system) for their actions and beliefs. If we adopt the above view about rationality, Evans and Over claim, we can explain why people sometimes make systematic errors in reasoning; they do so because of their limited ability to reason according to rationality$_2$.

Reasoning according to rationality$_2$ presupposes that there is a normative theory which defines the 'right' way of reasoning. Traditionally, as we saw, classical logic has been given the position of this theory. Evans and Over made apparent the need for a distinction between rationality$_2$ and a rationality which seems more appropriate for describing everyday reasoning (rationality$_1$). As we discussed earlier, classical logic fails to represent a fragment of natural language. Thus, if we assume that rationality and logical reasoning are guided by the norms of classical logic, an important part of natural language and everyday reasoning is left outside of this picture. It seems natural then to attempt a distinction between different kinds of rationality.

In this thesis we will examine a view (van Lambalgen and Stenning) in which we can still have a unique concept of rationality, by assuming that classical logic is not the only normative system of rationality. Evans and Over seem to be right to claim that people often fail to reason according to rationality$_2$, but we should wonder: is this failure due to the limited ability of human reasoning or maybe due to our failure to understand and adequately describe rationality?

1.1.2 Norms of rationality and how they are determined

Let us now look at the questions left aside earlier: Is there a definite and exhaustive list of rules which define rationality? And what is the correct way, if one exists, to discover them?

Normativity

The first of the above questions is directly related to the issue of normativity, which in general refers to the existence of an ideal defining how we ought to behave, what is right and what is wrong, what is good and what is bad. If we could somehow make evident that there exist certain rules—in other words, norms—defining rationality and make a list of them, then we would have solved the issue of the normativity of rationality and it would be an easy empirical task to check whether people act according to these rules and therefore whether or not humans are rational. However, discovering and listing the norms of rationality is not an easy task, and many (among them, as we will see later, Davidson) claim that such a list does not exist at all in a definite form. Moreover, there is the issue raised in the second of the above questions. Assuming that rationality is normative, what method should we follow to discover the norms of rationality?

Empirical vs. conceptual methods[2]

One suggestion towards finding an answer to the second of the questions posed at the beginning of this section was given by the *Cartesian project*: the only way to give an account of how we arrive at our beliefs (and thus of how we reason) is independent of experience since if we base our account on the shaky grounds of experience we might fall into mistakes. This view implies that the only way of arriving at a normative account of rationality is conceptual and that there is no empirical project which can determine how we ought to reason.

Others, like *W. V. O. Quine* with his essay 'Epistemology Naturalized', suggested that it is impossible to find some non-empirical standard on which we can base a normative account of rationality. What we should do instead is to base our account on what we have access to, namely on empirical evidence. We should thus replace *normative* epistemology with *descriptive* epistemology and, rather than looking at how we *ought to* reason, start looking at how we *actually* reason.

There is yet another suggestion coming from the field known as 'naturalized epistemology'[3] ('naturalized epistemology' should not be confused with Quine's paper 'Epistemology Naturalized'). What is suggested here falls between the two previous suggestions: it does not reject normativity and at the same time it considers empirical (scientific) investigation as an important element towards the determination of the rules that describe how we reason.

Experimental psychology of reasoning is an example of 'naturalized epistemology'. We are going to look at psychology of reasoning closely in the rest of this thesis. Psychology of reasoning has the role of 'asses[ing] empirically the nature and quality of human reasoning' [SvL08, p. 194], and is thought to be a descriptive enterprise: its focus is on investigating how people think and how they draw conclusions from premises. But can psychology of reasoning spell out the standards against which the reasoning abilities of humans can be judged? The 'traditional' epistemologist would say that there is no hope of arriving at such a list empirically, therefore psychology of reasoning has not much chance to estimate the real quality of human reasoning. However, experimental psychology of reasoning has as its starting point formal theories which provide us with some standards of good reasoning and proceeds in the following way: subjects are provided with reasoning tasks the 'right' answers of which are specified by the help of a formal theory, and then they are asked to solve the tasks. The quality of reasoning of the subjects is then judged by the answers they gave in comparison to the expected 'right' answers. We see that normativity is combined with empirical investigation, which can be characterized as 'naturalized epistemology' in the sense defined above.

[2] For the historical review in this section I have used [Ste96].

[3] This term is used by Edward Stein in [Ste96, p. 16], and describes the project which combines elements of traditional epistemology and descriptive epistemology.

1.1.3 Experimental psychology of reasoning

A main focus of research within experimental psychology of reasoning, and a main focus in the work of van Lambalgen and Stenning (especially in [SvL08] which we will discuss later), has been deductive reasoning. Deductive reasoning has to do with deciding the validity of arguments. An argument is said to be valid if, whenever its premises are true, its consequences are also true. An example of deductive reasoning would be the following,

> All men are mortal.
>
> Socrates is a man.
>
> Therefore, Socrates is mortal.

The above is a classical example taken from Aristotle, the 'father' of logic and deductive reasoning, who gave an account of deductive inference already more than two thousand years ago. One can easily see that in case the premises (the two first lines) of this argument are true, then the conclusion (the third line) of the argument is also—unavoidably—true. We can say thus that the above argument is valid, although this tells us nothing in general about the truth of the premises of the argument.

Despite the fact that psychologists had been long studying human reasoning, it was not until Piaget and his 'logicism' [Pia53], about half a century ago, that anyone tried to give a thorough account of the ability of people to make inferences. In Piaget's opinion, human beings are equipped with a mental logic which is fully developed when they are young adults.[4] If one takes this as a fact, which most of the psychologists of the time did, it only remains to discover the rules defining the principles of this logic. For a long time, it was thought that classical logic is the one people are equipped with and thus, this logic was thought to be the one that describes the inference mechanism of human minds. This idea, however plausible it seemed to be then, was later challenged by the facts: in many—not to say most—of the cases in everyday reasoning, people fail to derive the 'right' conclusions (that is, the ones classical logic suggests) from the given premises, an observation made by Peter Wason in 1968.

Peter Wason was a cognitive psychologist who with his work tried to explain why human beings quite often do not reason in accordance with the rules that rationality should, under 'normal' circumstances, have imposed on them. The widely discussed 'selection task' suggested by Wason (first introduced in [Was66] and later reported more extensively in [Was68]) is examined in [SvL08] and

[4] In fact, the idea that all adults reach the formal operations stage was later abandoned by Piaget himself. In [Pia08, p. 44] Piaget suggests that whether adults reach the formal operations stage or not depends to an extent on their environment and the education they receive: 'the formation of operations always requires a favorable environment for cooperation', that is to say, operations carried out in common (e.g., the role of discussion, mutual criticism or support, problems raised as the result of exchanges of information, heightened curiosity due to the cultural influence of a social group, etc.). [...I]n principle all normal individuals are capable of reaching the level of formal structures on the condition that the social environment and acquired experience provide the subject with the cognitive nourishment and intellectual stimulation necessary for such a construction.'

concerns the following experiment: Four cards are laid in front of the subject (usually an educated adult) which on their visible sides have either a number or a letter ordered as follows:

$\boxed{A}$ $\boxed{K}$ $\boxed{4}$ $\boxed{7}$

The subject is told that on the other side of each card there is either a letter or a number. Also, the following rule is provided to the subject: *If there is a vowel on one side of the card, then there is an even number on the other side.* Given the fact that this rule applies only to the four cards, the subject is then asked to decide which, if any, of these four cards he or she *must* turn in order to decide if the above rule is correct.

If we trust the idea of Piaget that all humans are equipped with a mental logic, we should expect any adult to be able to give the correct answer to the above question, since every human adult is supposed to be equipped with a mental logic which would be enough to yield the desired answer, that is, to turn the A and 7 cards.[5] However, the results of the selection task are not at all what would be expected. Most of the (always adult) subjects of the experiment fail to give the expected answer. Of course, this observation had a huge impact in the 'mental logic' view. So, what is going wrong? Why aren't people following the rules of the normative theory of reasoning (as classical logic was then believed to be)? Wason went on with conducting thematic variations of the same experiment and noticed that, when the content of the experiment resembles an everyday situation more than the original version of the task does, people manage to a greater degree to give a successful response, that is, one equivalent to the A, 7 response. What these results suggested, Wason thought, is that theories of reasoning should probably no longer be based on logic. The paradoxical nature of the results of experiments like Wason's led many—Wason included—to the conclusion that humans are irrational. But is this conclusion correct?

1.2 Aim of the thesis

In this thesis, I will look at two accounts on reasoning, rationality and interpretation: that of Donald Davidson, a philosopher of language, and that of the combined work of Michiel van Lambalgen and Keith Stenning, a logician and a cognitive scientist. I will investigate and then compare the two accounts with a main focus on the role logic plays in the understanding of reasoning, rationality and interpretation. In fact, both accounts would reject Wason's conclusion that humans are irrational.

[5] This answer is yielded if one assumes that the 'If ...then' clause in the task rule is interpreted as the material conditional. The material conditional is false when the antecedent is true and the consequent is false; in all other cases, it is true. Although Wason expected that most subjects would interpret the task rule in this way, this was not the case. In fact, the rule may be interpreted in many ways other than the classical logical interpretation of the task which Wason assumed. In [SvL08, pp. 52–91], van Lambalgen and Stenning provide multiple possible interpretations of the rule. We will come back to this point later in this thesis.

During the process of investigating the two accounts, I will try to give an answer to the following questions:

(a) What are the common elements of, and differences between, the two accounts?

(b) What is the innovation that van Lambalgen and Stenning's account offers?

(c) How can van Lambalgen and Stenning's work escape the threat of relativism (of which it is accused)? How can their theory be improved in that respect?

(d) What kind of criticism can be brought against a Davidsonian account from the van Lambalgen and Stenning perspective and vice versa?

(e) Would combining the two approaches result in a more 'complete' explanation of human reasoning and rationality?

(f) How can the two approaches be used in interpreting specific instances of discourse?

1.3 Outline of the thesis

In this chapter I gave an overview of the background against which the two accounts compared in this thesis will be discussed. The aim of the thesis was explained and the main questions to which we seek an answer were presented. Chapter 2 examines the Davidsonian account of rationality, reasoning and linguistic communication. Davidson's ideas have changed significantly over the course of his career, with 1984 (the year he published the essay 'A Nice Derangement of Epitaphs' [Dav86]) serving as a notional turning point. We will see Davidson's early ideas on rationality (section 2.1) and will compare them to his post-1984 ideas (section 2.2). Both Davidson's early and later ideas are relevant to, and possibly useful for, part of the van Lambalgen and Stenning account described in chapter 3. The innovations of the van Lambalgen and Stenning account of logical reasoning and rationality are presented (section 3.1), and some evidence taken from empirical investigations that support this account are briefly discussed (section 3.2). In the same chapter I discuss the criticism about relativism that the van Lambalgen and Stenning account has received and I make an attempt to defend the account against this criticism (section 3.3). In chapter 4, a comparison of the two accounts is made based on three main points: the overall aims of the two accounts (section 4.1), the definition of rationality and the role of logic in this definition (section 4.2), and the connection of the property of graceful degradation to the process of logical reasoning (section 4.3). The conclusion with the answers to the questions of section 1.2 together with some further plans for research is given in chapter 5. Finally, appendix A offers a short analysis of some empirical data based on the combination of the two accounts.

Chapter 2

Donald Davidson

In this chapter I will present Davidson's views on rationality, interpretation and on what is needed for building a theory of meaning—if we can say that such a theory exists. There are essential differences between Davidson's early work, mainly that on Radical Interpretation, and later Davidson (post 1984), in particular the turn which is obvious in his essay 'A Nice Derangement of Epitaphs' [Dav86]. The global view of rationality in early Davidson is seemingly abandoned in Davidson's later writings and is replaced by a more local view. The account of rationality that van Lambalgen and Stenning give also has important local elements, so it will be useful to examine what caused Davidson to move from a global to a local view and to use the outcome of this examination for the later comparison of the two accounts.

In what follows, I will present the views in early and later Davidson and locate their main differences, on which the comparison with the work of van Lambalgen and Stenning will be later based. The first and essential step in this direction will be to roughly present the general views that Davidson holds about rationality, about mental events and about whether there can be a science (in the ordinary sense) of rationality. This step is essential for obtaining a clear and coherent overview of Davidson's work since such an overview can only be obtained by looking at the tightly interrelated ideas spread across many pieces of Davidson's work.

2.1 Early Davidson

I will begin by presenting Davidson's ideas on rationality, interpretation and linguistic communication, as they appear before the turn which is obvious in his essay 'A Nice Derangement of Epitaphs.'

2.1.1 Rationality is global: Davidson and Decision Theory

Language and rationality are two main, distinctive human characteristics and, for Davidson, understanding linguistic communication is a basic step towards understanding human rationality. In fact, one of Davidson's main goals through his philosophical inquiries is to understand how language works and what actually the phenomenon of language is. In Davidson's work, rationality is considered to be a global characteristic of human beings and he is using this characteristic as a basis on which he can explain linguistic communication in a way that will become obvious in the progression of this thesis.

Davidson's ideas on what rationality is seem to have been much affected by decision theory and especially by the work of Frank P. Ramsey and Richard C. Jeffrey.[1] Much of Davidson's early work concerned decision theory and one can see references to decision theory in many of his later essays. For instance, in the essay 'What Thought Requires' [Dav01c, p. 146], in an attempt to examine what is required for a theory to be called scientific, Davidson writes:

> What one can say is that 'given the right conditions', *ceteris paribus*, the laws of decision theory do describe how people make real choices. [...] These are laws of human behaviour we depend upon, rough and fallible as they are.
>
> In the same way the laws of logic are laws of thought—always, of course, given the right conditions, and so forth. Tarski-type truth definitions, modified to fit natural languages, describe the basic semantic structure that informs the human language ability. [...] These three structures, of logic, decision theory, and formal semantics, have the characteristics of serious theories in science: they can be precisely, that is, axiomatically, stated, and, given empirical interpretation and input, they entail endless testable results. Furthermore, logic, semantics, and decision theory can be combined into a single unified theory of thought, decision, and language [...].

Later in the same essay we see Davidson's worries about what the actual standards of rationality are, when we are outside decision theory or logic, as well as his attempt to defend the extensional character of the Unified Theory (which we will briefly discuss later) even though the theory cannot be specified

[1] In fact, rationality as Davidson understands it, entails (at the least) logical consistency and speakers can be interpreted only if their beliefs are considered to be true and consistent with each other. This is supported by what Davidson maintains in [Dav85], for instance in [Dav85, p. 192], namely that the only clear case of irrationality is when there is inconsistency within a set of beliefs combined with principles. (I suppose Davidson, when referring to 'principles', he implies the principles of logic since, as we will see later in this thesis, Davidson claims that the semantic contents of beliefs—and actually all other propositional attitudes—are related to the world in consistent ways determined by certain norms and principles among which are those of logic.) For instance, if a person believes p and $\neg p$ at the same time then this person is irrational.

In addition, in [Dav80, pp. 156–157] Davidson claims that 'we take it as a constraint on possible interpretations of sentences held true that they are logically consistent with one another' or, in other words, that we 'assume the speaker's beliefs are logically consistent (up to a point at least).'

in physical terms.[2] The Unified Theory cannot be specified in physical terms since it is about psychological concepts; however, the theory depends on the attitude of preferring certain sentences true rather than others, and this implies that the theory can be stated in extensional terms.

Davidson thinks that a theory similar in certain respects to decision theory would be most suitable for describing human behaviour since it would allow the theory to be stated in a scientific way. As we will see in what follows, Davidson proposes the Unified Theory in order to explain verbal communication and to give a foundational account of language, and this theory has as a crucial part a version of decision theory. But before describing this theory, it is worth looking at Davidson's ideas on thought and mental events since it will help us formulate a clearer view on his ideas about rationality and, more generally, about human agency.

2.1.2 The mental

One of Davidson's main claims about the connection of the mental to the physical is what he calls 'The Anomalism of the Mental' [Dav70, p. 208]:

> [T]here are no strict deterministic laws on the basis of which mental events can be predicted and explained.

The fact that Davidson considers rationality as following certain normative principles which entail overall consistency and coherence makes him think of physical events as substantially different from mental events in that respect. In other words, Davidson thinks that what can be (and is) explained by physical descriptions are events very different in kind from mental events in the sense that they are not subject to the rules of rationality.

[2] The following quote [Dav01c, pp. 148–149] supports this claim:

> What makes the empirical application of decision theory or formal semantics possible is that the norms of rationality apply to the subject matter. In deciding what a subject wants or thinks or means, we need to see their mental workings as more or less coherent if we are to assign contents to them. As in any science, we must be able to describe the evidence in terms the relevant theory accepts. The trouble with the study of thoughts is that the standards of rationality, outside of decision theory or logic at least, are not agreed upon. We cannot compare our standards with those of others without employing the very standards in question. This is a problem that does not arise when the subject matter is not psychological.
>
> In one respect, the unified theory of thought and meaning which I described is a little better off than one might think. The important primitive term in that theory is the one expressing the attitude of preferring one sentence true rather than another. This is certainly a psychological concept, and a pretty complicated one. So there is no chance that the theory can be specified in physical terms. On the other hand, the theory is entirely stated in extensional terms. The relation of preferring true is a relation between an agent and two sentences, and it holds no matter how these entities are described.

Therefore, even though there is token identity[3] between mental and physical events, Davidson claims that the mental cannot be reduced to the physical by any explicit and strict laws and descriptions. The observation of this problem, that is, the apparent impossibility of constructing an adequate and complete theory of the mental in the form of a unified physical science, leads Davidson to attempt to find the best theory that would fit the mental and would be as close as possible to a unified theory of physical science by pairing mental events to the physical events which follow them. In the essay 'Problems in the Explanation of Action' [Dav87, p. 112], Davidson claims that any serious theory of the mental should pair mental to physical events, and the laws which would do this pairing cannot be reduced to the laws of a science like that of physics.[4]

In the same essay [Dav87, pp. 114–115], Davidson explains the main reason why the mental cannot be reduced to the physical, namely 'the normative character of mental concepts':

> The reason mental concepts cannot be reduced to physical concepts is the normative character of mental concepts. Beliefs, desires, intentions, and intentional actions must, as we have seen, be identified by their semantic contents in reason-explanations. The semantic contents of attitudes and beliefs determine their relations to one another and to the world in ways that meet at least rough standards of consistency and correctness. Unless such standards are met to an adequate degree, nothing can count as being a belief, a pro-attitude, or an intention. But these standards are norms, our norms, there being no others. [. . . I]n explaining action, we are identifying the phenomena to be explained, and the phenomena that do the explaining, as directly answering to our own norms; reason-explanations make others intelligible to us only to the extent that we can recognize something like our own reasoning powers at work. [. . .] We have noticed the obvious fact that a belief and a desire explain an action only if the contents of the belief and desire entail that there is something desirable about the action, given the description under which the action is being explained. This entailment marks a normative element, a primitive aspect of rationality.

Here we see that Davidson regards rationality as a normative characteristic of all human beings and their mental concepts. This fundamental characteristic

[3] "The token identity theory (defended by Kim (1966) and Davidson (1980) among others) maintains that every token mental event is some token physical event or other, but it denies that a type match-up must be expected. So for example, even if pain in humans turns out to be c-fiber stimulation, there may be other life forms that lack c-fibers but have pains too. And even if consciousness in humans turns out to be a brain waves that occur 40 times per second, perhaps androids have consciousness even if they lack such brain waves." [Weted]

[4] "Explanations of mental events must include reference to physical causes (as in perception etc.), and as we have seen, actions are typically characterized in terms of their physically described consequences. So any 'theory' of the mental must cover interactions between the mental events (i.e., events described in mentalistic ways) and physical events (i.e., events characterized in physical ways). The basic difference that I think exists between reason-explanations and the explanations of an ultimate physics can therefore be put this way: laws relating the mental and the physical are not like the laws of physics, and cannot be reduced to them. Since action explanations require such laws, action explanations are not like explanations in physics, and cannot be reduced to them" [Dav87, p. 112].

of mental concepts, normativity, is what prevents us from adequately describing them in physical terms. So, is there any kind of science that could describe the mental behaviour of humans in a similar way as a physical science? Davidson is trying to build such a theory, and we are now going to look at his attempts.

2.1.3 The Unified Theory

A science of rationality

In the essay 'Could there be a science of rationality?' [Dav95] Davidson examines whether there can be a science, like that of physics, about the mental and verbal behaviour of human beings. He discusses the reasons why the mental is irreducible to the physical and he actually stresses two of them: *normativity* which is introduced by the necessity of the application of the Principle of Charity,[5] and *the irreducibly causal character of mental events*.

The obvious problem here is that when one is trying to explain human behaviour, that is, human actions, usually it is propositional attitudes (such as beliefs, desires, hopes etc.) that are considered as giving the appropriate explanations. What several scientists and philosophers (e.g. Fodor) have been trying to do is to find some purely internal (i.e., independent of any relations to the external world) aspect of propositional attitudes, since this would allow one to make lawlike connections solely between the propositional attitudes of an agent and his or her (inner) physical properties. For instance, Fodor claims that there is no hope in accounting for a serious science for linguistic phenomena if there is no obvious and lawlike ways of pairing the meanings of expressions to internal representations [Dav95, p. 112]. This is a form of internalism.

Davidson's stance with respect to internalism is quite different from that of Fodor. It is clearly suggested by Davidson's writings that he is an externalist which means that he believes in the doctrine that mental states (of an agent) supervene[6] not only on the physical properties of the agent but in addition on the physical properties of the world outside the agent [Dav95, p. 122]. This implies that the content of propositional attitudes is partly determined by the external objects they are about. Davidson moreover claims that normativity, holism and externalism are interrelated in a way that they cannot be separated and they are all together significant features of mental events; they either stand or fall together. Therefore, if any of the above elements in psychological concepts is eliminated, there can be no serious science of the mental, unless we radically change the subject of our inquiry.[7]

Given all the above, Davidson sets a twofold task; to defend a 'theory' which gives an explanation of thought, language and action, and to explain why, if at all, it should be considered as scientific. The theory Davidson proposes, the *Unified Theory*, relates the concepts of belief, desire, and linguistic meaning.

[5] We will discuss the Principle of Charity later in the section on Radical Interpretation.

[6] We say that a set of properties A supervenes on a set of properties B if and only if any two objects x, y that share all properties in B must also share all properties in A.

[7] See [Dav95, pp. 122–123] and [Dav70] for more details on this subject.

The theory comprises by two main parts. In the first part, by using tools from decision theory (measuring belief by subjective probabilities, and desires on an interval scale) predictions can be made about intentional actions. The second part of the theory treats linguistic meaning, since it includes a Tarski-style theory of truth which can be used for giving the truth conditions of all utterances of sentences in a given language. Finally, decision theory and truth theory are joined by a formal device, the details of which are offered by Davidson in [Dav80, pp. 160–165]. For now, it suffices to say that the theory Davidson has in mind, uses Jeffrey's version of decision theory which assigns to agents preferences among propositions, with the following alteration: Davidson applies it to sentences the meaning of which is not yet known; subjective desirabilities and probabilities of all sentences are calculated and only after that the meaning of them is derived (see also footnote 9 below).

The empirical interpretation of the Unified Theory

It is interesting to take a look at the way Davidson describes how the Unified Theory can be interpreted empirically. As we will see, it seems that the proposed theory, unlike any other scientific theory, does not have any practical applications, which should make us suspicious about its plausibility. The description Davidson gives is the following.

First, in the decision theoretic part of the theory, the utterer's propositional attitudes are identified by making the assumption that the norms of rationality are followed. In this step, Jeffrey's version of decision theory is used in order to assign subjective probabilities to agents, where subjective probability is the degree of belief of an agent about some proposition. As we read in Jeffrey's [Jef83, p. 62] 'subjective probability is partial belief; beliefs are manifest in action and in the agent's attitudes toward alternative courses of action'. The assumption of rationality places certain constraints which help the interpreter estimate the likelihood of the utterer holding one sentence,[8] rather than another, to be true, and express this by subjective probabilities. These probabilities are later used to interpret the sentences that are uttered by the agent. The basic step of this procedure is the determination by empirical means of the preference of an agent that one sentence rather than another is true.[9] This method of

[8] We should note here that Davidson uses Jeffrey's theory slightly modified. Jeffrey is showing how to assign subjective probabilities to agent's preferences about propositions rather than sentences. But propositions (as Davidson stresses in [Dav80, p. 160]) are sentences with meaning, and the theory Davidson is trying to construct does not presuppose meaning. Therefore the theory that Davidson seeks should make use of Jeffrey's results on uninterpreted sentences rather than propositions. The details of how this change can be expressed in decision theoretic terms are given in [Dav80, pp. 161–164].

[9] Davidson gives the following rough description of how Jeffrey's version of decision theory can be applied to sentences in order to help the interpreter figure out the meaning of simple sentences and sentential connectives, as well as more theoretical terms [Dav95, p. 127]:

> Jeffrey's version of decision theory, applied to sentences, tells us that a rational agent cannot prefer both a sentence and its negation to a tautology, nor a tautology to both a sentence and its negation. This fact makes it possible for an interpreter to identify, with no knowledge of the meanings of the agent's sen-

the interpretation of the Unified Theory is what Davidson has called Radical Interpretation.

2.1.4 Radical Interpretation

In this section we will discuss in more detail Radical Interpretation (first introduced in [Dav73]), a project mainly concerned with the two following questions: *What would be sufficient for an interpreter to know in order to understand the utterances of a speaker when the language of the speaker is completely alien to the interpreter*, and, *how would he come to know it*? With these two central questions as a starting point, a theory of how-to-build-a-theory-of-meaning is developed, a theory which seeks to give an answer to the more general questions of what meaning is and how linguistic communication is achieved.

Non-linguistic evidence as a guide to interpretation

To give a rough outline of one of the central points of the theory, which would answer the first of the above questions, we should say that in Radical Interpretation the only useful evidence which is available to us if we are asked to interpret a foreign speaker's utterance is the empirical, non-linguistic evidence relevant to that particular utterance, such us the time and place of the utterance and all the states of affairs that come along with it. This empirical evidence, which essentially determines the truth conditions of the utterance, is what fixes the meaning of the utterance. It is important to note that, here, it is truth that determines meaning and not the converse.

According to Davidson, truth sits at the heart of any theory of meaning for natural language. In the essay 'Truth and Meaning' [Dav67, p. 24], first published in 1967, that is, several years before the first publication of 'Radical Interpretation', we read:

> There is no need to suppress, of course, the obvious connection between a definition of truth of the kind Tarski has shown how to construct, and the concept of meaning. It is this: the definition works by giving necessary and sufficient conditions for the truth of every sentence, and to give truth conditions is a way of giving the meaning of a sentence. To know

> tences, all of the pure sentential connectives, such as negation, conjunction, and the biconditional. This minimal knowledge suffices to determine the subjective probabilities of all of the agent's sentences—how likely the agent thinks those sentences to be true—and then, in turn, to fix the relative values of truth of these sentences (from the agent's point of view of course). The subjective probabilities can then be used to interpret the sentences. For what Quine calls observation sentences, the changes in probabilities provide the obvious clues to first order interpretation when geared to events and objects easily perceived simultaneously by interpreter and the person being interpreted. Conditional probabilities and entailments between sentences, by registering what the speaker takes to be evidence for his beliefs, provides the interpreter with what is needed to interpret more theoretical terms and sentences.

For more details and examples which illustrate how the theory works the interested reader should look at [Dav80].

> the semantic concept of truth for a language is to know what it is for a sentence—any sentence—to be true, and this amounts, in one good sense we can give to the phrase, to understanding the language.

As indicated above, when the radical interpreter sets off to interpret the utterances of an alien speaker, he is assigning truth conditions to the utterances; these truth conditions determine the meaning of utterances and at the same time the interpreter is also able to assign beliefs to the alien speaker.[10]

But what exactly is the evidence that allows the radical interpreter to do that? First, knowledge of the alien speaker's propositional attitudes towards the truth of the utterances in the alien language and, second, knowledge of the alien speaker's interactions with his environment as well as knowledge of the physical world [LL05, p. 157]. We should keep in mind however, that the radical interpreter need not have any information of the meanings of any of the alien utterances or of the alien speaker's general propositional attitudes.

Davidson claims that there is no hope for interpreting linguistic activity if we have no knowledge of what a speaker believes, and that we cannot found a theory of what the speaker means based on prior knowledge of what the speaker believes and intends. This means that in Radical Interpretation, a theory of meaning and a theory of belief are delivered at the same time. But what makes this possible? Here, a central concept of Radical Interpretation, the *Principle of Charity*, comes to the rescue.

The Principle of Charity

It is this principle that allows us to presuppose, given the fact that all we humans share the same basic physical properties, that we all also share the same basic beliefs (relative to a specific context) which helps the interpreter to simultaneously assign content to someone's beliefs and meanings to his utterances.

[10] We can find these kind of indications in much of Davidson's work. For instance in 'Thought and Talk' [Dav75, pp. 161–162]:

> The interlocking of the theory of action with interpretation will emerge in another way if we ask how a method of interpretation is tested. In the end, the answer must be that it helps bring order into our understanding of behaviour. But at an intermediate stage, we can see that the attitude of *holding true* or *accepting as true*, as directed towards sentences, must play a central role in giving form to a theory. On the one hand, most uses of language tell us directly, or shed light on the question, whether a speaker holds a sentence to be true. [. . .] On the other hand, knowledge of the circumstances under which someone holds sentences true is central to interpretation.

And in 'Belief and the Basis of Meaning' [Dav74, p. 144]:

> I hope it will be granted that it is plausible to say we can tell when a speaker holds a sentence to be true without knowing what he means by the sentence, or what beliefs he holds about its unknown subject matter, or what detailed intentions do or might prompt him to utter it. It is often argued that we must assume that most of a speaker's utterances are of sentences he holds true: if this is right, the independent availability of the evidential base is assured.

Thus, the interpreter takes for granted that the utterer is a creature with beliefs and intentions and that he both intends to utter something true and believes that his utterance is true and, in addition, that the beliefs the utterer holds are of the same nature as his (the interpreter's) own.

The interpreter begins with establishing laws of the form 'for speaker S, at time t, *ceteris paribus*, S holds true some sentence s at t, if and only if p', where s is an alien sentence uttered on a certain occasion and which the interpreter wants to interpret, and p is a sentence which specifies the truth conditions under which the alien speaker utters s. The truth conditions are relative to time and any other extra-linguistic evidence which forms the context of the utterance, and they are the same truth conditions under which the interpreter would utter a sentence s', equivalent to s, in his own language. After that, the interpreter will apply the Principle of Charity and will be able to arrive at sentences of the form 'For all speakers S, at time t, s is true for S at t if and only if p.' From this kind of sentences the radical interpreter will in the end be able to predict the axioms of the truth theory. This procedure sums up in a way the strategy of Radical Interpretation.

Davidson describes the Principle of Charity in the essay 'Paradoxes of Irrationality' [Dav82, p. 138] as follows:

> If we imagine ourselves starting out from scratch to construct a theory that would unify and explain what we observe—a theory of the man's thoughts and emotions and language—we should be overwhelmed by the difficulty. There are too many unknowns for the number of equations. We necessarily cope with this problem by a strategy that is simple to state, though vastly complex in application: the strategy is to assume that the person to be understood is much like ourselves. That is perforce the opening strategy, from which we deviate as the evidence piles up. We start out assuming that others have, in the basic and largest matters, beliefs and values similar to ours. We are bound to suppose someone we want to understand inhabits our world of macroscopic, more or less enduring, physical objects with familiar causal dispositions; that his world, like ours, contains people with minds and motives; and that he shares with us the desire to find warmth, love, security, and success, and the desire to avoid pain and distress. [. . . U]nless we can interpret others as sharing the vast amount of what makes up our common sense we will not be able to identify any of their beliefs and desires and intentions, any of their propositional attitudes. The reason is the holistic character of the mental.

Thus, Charity is about agreement (in general), between interpreter and speaker, about their shared environment in the sense that what the interpreter holds to be true in a certain context will be thought of as held true also by the speaker. And this principle is thought of by Davidson not as giving advice to actual interpreters on an interpretation strategy but rather as a condition of the possibility of interpretation.[11] Moreover, the whole method of Radical Interpretation, as Davidson clearly states, does not suggest an actual strategy

[11] For that see also [Ram89, p. 74].

for interpretation but rather it indicates which preconditions exist for interpretation. We should thus think of it not as a suggested theory of meaning, but rather as a model of a process[12] towards a theory of meaning.

The process of Radical Interpretation

I will now give a sketch of how Radical Interpretation works in 6 steps.[13]

1. With the aid of the Principle of Charity, the radical interpreter has to describe in his own language the utterer's beliefs and desires and take notes on how they relate to the non-linguistic evidence (time and place of the utterance etc.) which comes along with them and which constitutes the truth conditions of the utterer's utterances.

2. Assuming that the utterer holds his utterances true at the time and place of the utterance (that is, excluding the cases where lie or deceit on the utterer's part is obvious) one can conclude that the utterer believes his utterances. The radical interpreter is ready now to write down in the alien language the sentences of the alien language which correspond to the utterer's beliefs as already expressed in the radical interpreter's language.

3. By use of the Triangulation Principle[14] the radical interpreter should now try to write down which sentences, both in the interpreter's language and in the utterer's language, satisfy the truth conditions that were observed and written down in the previous steps. Thus, the interpreter will now have a way to match the utterer's sentences to sentences of his own language. In this step the interpreter has to moreover make use of the Principle of Charity so that he can simultaneously revise the utterer's beliefs as stated in the interpreter's language and the truth conditions of the utterances.[15]

4. By continuing to observe the utterer's behaviour, the interpreter should now revise first the utterer's desires as stated in the interpreter's language and then, by use of the Triangulation Principle again, also the utterer's desires as stated in the alien language.

5. The interpreter has to go back to step 2 and check whether the correspondence between the sentences of our own language and of the alien language which express the same beliefs has changed after the revision of step 4. If something has changed the interpreter has to start over at step

[12] "Process" here is used in the sense of "procedure."

[13] For the steps that I will present here, I have followed partly Lewis's essay on Radical Interpretation [Lew74] where, apart from proposing his own version, Lewis presents Davidson's version of Radical Interpretation.

[14] See the discussion on the next page about what the principle of triangulation is.

[15] It is important to mention that this step is the basic problem of Radical Interpretation as introduced by Davidson, that is, the simultaneous filling in of the utterer's beliefs in the interpreter's language and of the meaning (the truth conditions) of the utterer's utterances as expressed in the utterer's language.

2, now taking as a starting point the revised set of beliefs as stated in the interpreter's language. If nothing has changed then the interpreter has reached a stable solution.

6. Finally, the interpreter has to check his findings against the other speakers of the utterer's community and make the necessary revisions.

From the above one realizes that Radical Interpretation seems to be a process with a non-clear end point rather than a stable theory of meaning which models meaning in a fixed way once and for all. The radical interpreter might need a huge number of iterations before he comes up with a stable solution to the problem of interpretation, which means that to achieve interpretation probably he needs many more years than his life will last. However, Davidson in his early work about Radical Interpretation claims that such a solution to the problem of communication must exist, that is, he thinks that a theory of meaning can be reached after the process of Radical Interpretation. Actually, compositionality here comes to the rescue and, as Davidson clearly states, it should be one of the main characteristics of any theory of meaning and truth. Compositionality would make sure that we can infer the meaning of any sentence of a language by knowing only a finite number of fixed meanings and rules of the language.

The Principle of Triangulation

At this point, we should take a closer look at the Principle of Triangulation which was mentioned in step 3 of the process of Radical Interpretation. Triangulation is essential as an argument for the necessity of language for thought. Davidson has claimed that communication can be achieved in a context where 'each of two individuals (at least) find themselves responding to the same object and to each other, thereby forming a triangle, where, as Davidson says, the base line is the communication between them' [LL05, p. 404]. Thus, it does not make sense, according to Davidson, to talk about having thoughts at all if there are not at least two people to communicate with each other and interact with the external world. Here we can see again how the externalist stance of Davidson is obvious in his work. In slightly more detail, what is taking place during triangulation is the following. There is an object in the real world to which interpreter and speaker react with the same response and, granting the presupposition that all human beings respond in similar ways to the same objects, we conclude that interpreter and speaker share similar thoughts.

In the essay 'What Thought Requires' (2001) [Dav01c], where an attempt is made to discover the right criteria for thinking, the relation between language, thought, and world, is re-examined by Davidson who stresses the importance of triangulation. For triangulation to work though, the 'creatures involved must be very much alike' [Dav87, p. 143]. What is meant here is that, because all human beings share the same discriminatory abilities,[16] each one of us has the

[16] Discriminatory abilities are the abilities of humans to distinguish the same objects from their surroundings. Having the same discriminatory abilities does not imply that we all see

same concepts which define the same classes of objects and therefore we give the same (or at least similar) responses to the same stimuli. It is much more helpful, Davidson concludes, to have a third-personal approach to language and thought if we want to understand thought. Studying how it is possible for one person to understand the speech and thoughts of another (what Radical Interpretation is actually about) instead of what happens inside someone's mind in isolation is what we should do.

Radical interpretation is not a theory of meaning

To conclude the section on Radical Interpretation, we should stress once more that Radical Interpretation is not meant to be a theory of truth and meaning. It is a process during which the Radical Interpreter has to, step by step, eliminate possibilities of certain correspondences between truth conditions and utterances aiming at an ultimate theory of truth that would describe all linguistic communication. This procedure looks more like an ongoing process of exchanging one truth theory for another rather than a step-by-step construction of one theory, as Ramberg—rightly I think—concludes in [Ram89, p. 80]. The following quotation [Ram89, Ch. 6, p. 78] describes nicely how one should think of Radical Interpretation.

> [. . . T]he radical-interpretation model must be understood as a model of a process, not as a model of a static state of semantic competence. More precisely, we might say that semantic competence as it is modeled by radical interpretation is a process, and so cannot be modeled by any one theory of truth. Talking as if any particular, more or less complete, theory of truth might represent a level of semantic competence might lead us to seriously misconstrue the nature of this competence, by ignoring the essentially dynamic nature of semantic understanding.

Finally, it is worth mentioning what Davidson himself says about Radical Interpretation and in general about whether any truth theory is adequate for natural language. In [Dav94, pp. 126–127] we read:

> It is a big question whether such theories [theories of truth constructed more or less along Tarski's lines] can be made adequate to natural languages; what is clear is that they are adequate to powerful parts of natural languages, parts with great expressive power. Of course I did not say speakers or interpreters actually formulate such theories. It does seem to me though, that if we can describe how they could we will gain an important insight into the nature of the intentional (including, of course, meaning), in particular into how the intentional supervenes on the observable and non-intentional.
>
> [. . .]

exactly the same thing. Rather, it means that when humans look at say a chair, all of them can discriminate the object called 'chair' from its surroundings. If we did not have the same discriminatory abilities, then communication would probably be impossible.

> I believe that what people mean by what they say derives from the occasions of successful communication, so I restrict the evidence to what would be plainly available to an observer unaided by instruments. The other important restriction is to assume no prior detailed knowledge of any of the propositional attitudes: beliefs, desires and intentions. As a result, I consider a theory of meaning to be an undetachable part of a more general theory of human behaviour.

2.2 Later Davidson

As we saw, Davidson's earlier writings support that if there exists a theory of meaning then it will have to be a Tarski-style theory, a real scientific theory which can be axiomatized and which explains linguistic communication. Actually, Davidson maintains that such a theory must exist and sets the goal to discover the way to this theory. This is what Radical Interpretation is supposed to do: show the way to an adequate theory of meaning. Once we reach it, the theory will be sufficient for explaining how natural language works. In the process of Radical Interpretation, for the interpretation of discourse, there is no prior knowledge needed of the beliefs, desires, and intentions of the speakers, which suggests that the theory of meaning that Davidson has in mind is an undetachable part of a general theory of behaviour.

We have already discussed how, by taking as first evidence certain observable aspects of verbal behaviour, Davidson is establishing a theory of truth which he later combines with a version of decision theory (a theory of belief, meaning and desire) in order to arrive at a theory of meaning. What Davidson thus seems to be claiming in his early writings is that *there is a single systematic theory of meaning* (along the Tarskian lines) which perfectly describes natural language and therefore linguistic communication. However, we will see that Davidson seems to change his mind in his later writings, and especially in the essay 'A Nice Derangement of Epitaphs' [Dav86]. In this later work, Davidson seems to abandon the idea of a single theory of meaning and suggests a multiplicity of meanings that different people may be equipped with at different times, depending on the context of a conversation. Davidson's later view will be relevant for the comparison with the multiple-logics view of rationality introduced by van Lambalgen and Stenning.

2.2.1 The turn

Does a *single* theory of meaning exist?

The later (post 1984) Davidson is no longer hoping for a *single* systematic theory which explains all linguistic communication. Malapropisms[17] and other

[17] A malapropism is the act of using a wrong word in the place of the right word because the two words sound similar. In fact, the title of Davidson's essay 'A Nice Derangement of Epitaphs' is based on a famous example of such a speech mis-use borrowed from R. B. Sheridan's play *The Rivals*, from where actually the term *malapropism* originates. A character in this

irregularities that we come across in everyday-life discourses, led Davidson to conclude that it does not make sense to claim that all humans are armed with a theory of meaning which gives them in advance the 'tools' to interpret every utterance they will stumble upon. The fact that we can deal with malapropisms and other irregularities, seems to suggest that there are no conventional rules that we first learn and then apply to instances of communication. Language users must be—and actually are—creative and able to figure out the (contextual) meaning of an utterance right on the spot. With Davidson's own words [Dav86, p. 446]:

> We must give up the idea of a clearly defined shared structure which language-users acquire and then apply to cases. And we should try again to say how convention in any important sense is involved in language; or, as I think, we should give up the attempt to illuminate how we communicate by appeal to conventions.

Does *any* theory of meaning exist?

This conviction does not make Davidson conclude that there is *no* kind of systematic theory at all that people are equipped with, as we might have expected. He simply abandons the idea that there is a *unique* systematic theory with which each one of us is equipped and which is complete and available to us in advance. In Davidson's judgment, each person is equipped with a systematic theory of how-to-assign-meaning-in-context *but* this theory is not fixed; it changes and is revised according to the circumstances.

Before explaining the new picture of interpretation suggested by Davidson, we should take a look at the path he follows for arriving at it and elucidate the goal he aims at achieving. Instead of starting at the beginning, I will start from Davidson's own concluding words [Dav86, p. 446]:

> I conclude that there is no such thing as a language, not if a language is anything like what many philosophers and linguists have supposed.

These words suggest that Davidson's main aim in [Dav86] is to attack the, until that point, standard descriptions of linguistic competence (and therefore the, until then, established idea of what a language is). As he mentions, these descriptions include also some of his own earlier work. What is now needed, according to Davidson, is a deeper notion of the contextual meaning of words which will permit us to distinguish between speaker meaning and literal meaning. The only way this can be done, Davidson claims, is by modifying the accepted view of what natural language is and of what it is 'to know a language.'

We therefore need to see how this accepted view—that Davidson is attacking—is formed. At this point we have to look at a term introduced by Davidson [Dav86, p. 434] which will be central in our further discussion.

play, Mrs. Malaprop misuses the words *derangement* and *epitaphs* in the place of the words *arrangement* and *epithets*: "There, sir, an attack upon my language! what do you think of that?—an aspersion upon my parts of speech! was ever such a brute! Sure, if I reprehend any thing in this world it is the use of my oracular tongue, and a nice derangement of epitaphs!" Assuming that the audience will recognize the malapropism, Mrs. Malaprop sounds funny.

First meanings

'First meaning' is the term Davidson uses for the literal meaning of a word and the concept of first meaning, as he says, 'applies to words and sentences as uttered by a particular speaker on a particular occasion'. Actually the word 'first' implies that this meaning comes first in the order of interpretation, so the 'first meaning' is the meaning that the speaker intends the words he is using to have and sometimes this coincides with the literal meaning of those words (that is, the conventional one which can be found in dictionaries) while at other times it is just the first-in-order meaning according to the intentions of the speaker. Therefore, the literal meaning of a word should not be *identified* with its conventional meaning but rather with the (essential to all communication) first-in-order intended meaning on the part of the speaker. This important distinction will later help Davidson arrive to the conclusion that conventional meanings, although helpful for linguistic communication, are neither necessary nor sufficient for achieving successful communication.

Let us now look at the accepted view that Davidson wants to attack. As Davidson claims [Dav86, p. 436] it has been the belief of many philosophers and linguists that speaker and hearer share a theory which allows them to articulate logical relations between utterances and interpret them in an organized way. These theories require from first meaning in language the following [Dav86, p. 436]:

> 1. *First meaning is systematic.* A competent speaker or interpreter is able to interpret utterances, his own or those of others, on the basis of the semantic properties of the parts, or words, in the utterance, and the structure of the utterance. For this to be possible, there must be systematic relations between the meanings of utterances.
> 2. *First meanings are shared.* For speaker and interpreter to communicate successfully and regularly, they must share the method of interpretation of the sort described in 1.
> 3. *First meanings are governed by learned conventions or regularities.* The systematic knowledge or competence of the speaker or interpreter is learned in advance of occasions of interpretation and is conventional in character.

Davidson's objection is related to the third of the above requirements. If one accepts the third requirement as true, then first meanings must be identified with conventional meanings, and the important distinction drawn by Davidson earlier without doubt disappears. In a way, this requirement suggests that everything we need in order to interpret speakers, we are equipped with in advance of any specific case of discourse in the form of learned conventions. But if we accept this suggestion, several questions arise: what happens in cases where malapropisms occur in language but nevertheless we succeed in deriving the correct interpretations? How do the conventions that should have helped us interpret an utterance enter the picture then? Davidson claims that if we want to explain malapropisms and to accommodate them into our theory of how linguistic communication is actually achieved, we have to change in some way the third principle. In order to provide a solution to this problem, Davidson suggests an alternative view of how linguistic interpretation works.

How does interpretation work? ***Prior*** and ***Passing*** theories

What is taking place during interpretation in Davidson's words is the following [Dav86, p. 441]:

> Here is a highly simplified and idealized proposal about what goes on. An interpreter has, at any moment of a speech transaction, what I persist in calling a theory. [...] I assume that the interpreter's theory has been adjusted to the evidence so far available to him: knowledge of the character, dress, role, sex, of the speaker, and whatever else has been gained by observing the speaker's behaviour, linguistic or otherwise. As the speaker speaks his piece the interpreter alters his theory, entering hypotheses about new names, altering an interpretation of familiar predicates, and revising past interpretations of particular utterances in the light of new evidence.

And on the speaker's side, Davidson suggests the following [Dav86, p. 442]:

> The speaker wants to be understood, so he utters words he believes can and will be interpreted in a certain way. In order to judge how he will be interpreted, he forms, or uses, a picture of the interpreter's readiness to interpret along certain lines. Central to this picture is what the speaker believes is the starting theory of interpretation the interpreter has for him. The speaker does not necessarily speak in such a way as to prompt the interpreter to apply this prior theory; he may deliberately dispose the interpreter to modify his prior theory. But the speaker's view of the interpreter's prior theory is not irrelevant to what he says, nor to what he means by his words; it is an important part of what he has to go on if he wants to be understood.

Davidson thinks that both speaker and interpreter are equipped with a *prior theory*. For the interpreter, the prior theory 'expresses how he is prepared in advance to interpret an utterance of the speaker' (and of course this prior theory might prove to be insufficient for yielding the desired interpretation) while for the speaker, the prior theory 'is what he believes the interpreter's prior theory to be'. Davidson then brings into the game the central definition of a *passing theory*. For the interpreter, the passing theory expresses how the interpreter actually interprets the utterances of the speaker, while for the speaker the passing theory is 'the theory he intends the interpreter to use'.

After arriving at the above definitions, Davidson makes the following point. Although most of the times interpreter and speaker are not actually sharing prior theories (and even as he says 'an interpreter must be expected to have quite different prior theories for different speakers'), it is necessary to be sharing passing theories if they want to achieve successful communication. This is because the passing theory is the one used by the interpreter to interpret the speaker's utterances and at the same time the one that the speaker intends the interpreter to use.

Thus, having a passing theory means knowing how to interpret a particular utterance in a particular situation. It is stressed by Davidson that a passing theory is by no means something that you could have known before the occurrence of that particular utterance in that particular situation. You only arrive at it and make use of it on the spot by using what is provided to you by your prior theory. The fact that there is no explicit rules one can know in advance for arriving at a passing theory leads Davidson to the conclusion that conventions are not basic for interpretation of speech and therefore for linguistic communication.

2.2.2 Rationality: global or local?

Davidson in his later work, claims that people are not equipped with a single theory of meaning in advance of any instance of interpretation and communication. Although it might seem that Davidson's earlier and later work entirely contradict each other, there are still many similarities. For instance, there are significant similarities between Davidson's later work and his work on Radical Interpretation.

In Radical Interpretation Davidson suggested that rationality is a global characteristic of humans. We are all equipped with the same kind of rationality, and this presupposition is necessary for communication. This view about rationality lead Davidson to the conclusion that there must be a single theory of meaning in advance of any instance of interpretation. However, he later realized that such a single theory does not exist. Rather, he suggested that there exist many theories of meaning which are constantly updated in a local level. Does this mean that a global notion of rationality can no longer be held? Does the move from a global to a more local view affect the normative character of rationality in Davidson's theory? I believe not. Updating your theory of meaning in a single instance of communication depending on contextual parameters does not imply that you have a different kind of rationality from the people you communicate with. What it does seem to imply, however, is that the normative system that defines rationality cannot be a single logic, something that Davidson did not suggest in any of his writings examined here. In the account that van Lambalgen and Stenning offer, we will see that there is an attempt to combine this global character of rationality with the multiplicity of logics (at a local level) which would determine the normative system according to which one reasons given certain contextual parameters.

In discussing some of the main work of Davidson, we run across his claims about the central aim of his theory, that is, to look at what is needed in order to arrive at a theory of meaning and at an explanation of linguistic communication rather than to throw light on how we actually understand each other in everyday life. Davidson often claims that the aim of his theory is not to describe how we actually interpret (see for instance [Dav95, p. 128]) but to give an account of what makes thought and language interpretable. Does this mean he does not think that empirical investigation is necessary for the investigation of what rationality is? I think this is not what Davidson claims, and in the following section I will give evidence for this claim.

2.3 Empirical investigation of rationality and specification of its basic characteristics

We should make clear two points that will be of great importance in our discussion: first, Davidson's claims about the possibility of the empirical investigation of rationality and second, the degree of variation that Davidson allows in the specification of the basic characteristics of rationality. If Davidson does not think that empirical investigation is necessary in order to understand human communication and rationality, then we will have serious difficulties in our attempt to compare his theory to that of van Lambalgen and Stenning, a theory which supports the claim that empirical investigation is the way of arriving at a correct understanding of human rationality. Similarly, if Davidson's theory suggests that the characteristics of rationality are decided once and for all and there is no room for variation, then the comparison with a theory like that of van Lambalgen and Stenning would be, not impossible but, definitely less interesting since in their opinion—as we will see in the following chapter—the norms of rationality are not thought of as decided once and for all by classical logic. I will try to throw some light on Davidson's ideas about these two issues here.

The possibility of the empirical investigation of rationality

The method Davidson follows, in order to arrive at an understanding of what rationality is and of how human communication is achieved, is both conceptual and empirical.

In the essay 'Could there be a Science of Rationality?' [Dav95, p. 124], Davidson—while discussing Fechner's law—makes the following claim which I believe suggests that every scientific theory should be tested by empirical interpretations:

> Fechner had the right idea. If scientific methods can be applied to the mental, it is by proposing a solid theory and asking how it can be tested and interpreted empirically. Theories describe abstract structure; their empirical interpretations ask whether these structures can be discovered in the real world.

Even though the theory of meaning Davidson suggests does not say much about actual situations of interpretation and is not empirically tested, Davidson does not ignore the importance of empirical interpretation. In fact, he is aware that 'it is when we attend to the empirical interpretation of the theory that the basic questions and problems arise' [Dav95, p. 127], and he adds:

> Unofficially, one can admit that as living, working interpreters, we have never enough of the sort of evidence needed to follow the official route

> [the method of interpretation of the Unified Theory], and we always have a great deal of other sorts of evidence. We make endless assumptions about the people we meet, about what they want, what they are apt to mean by what they say, what they believe about the environment we share with them, and why they act as they do. Our skills as interpreters come into play mainly when one or another of these assumptions turn out to be false, and by then we have much more than the poverty-stricken evidence the Unified Theory depends on.

This is not a problem, Davidson insists; the actual purpose of the theory after all was only to speculate about what it is that makes thought and language interpretable, and this is why the theory he proposes has little to say on first language acquisition and the origins of speech.

Elsewhere, he expresses his doubts on whether empirical investigation can be used against the axioms of any scientific theory. We read in [Dav76, p. 273],

> I am skeptical that we have a clear idea what would, or should, show that decision theory is false; and I think that compared to attributions of desires, preferences or beliefs, the axioms of decision theory lend little empirical force to explanation of action. In this respect, decision theory is like a theory of measurement for length or mass, or Tarski's theory of truth. The theory in each case is so powerful and simple, and so constitutive of concepts assumed by further satisfactory theory (physical or linguistic) that we must strain to fit our findings, or our interpretations, to preserve the theory. If length is not transitive, what does it mean to use a number to measure a length at all? We could find or invent an answer, but unless or until we do we must strive to interpret 'longer than' so that it comes out transitive. Similarly for 'preferred to'.

This passage seems to suggest that we should adjust the findings of our experimental investigations of what rationality is to the standards provided by the scientific theory we use. In this sense, decision theory or logic cannot be false. They provide a conceptual backbone of a theory of rationality, and if the evidence we come across suggests that people do not actually follow the rules of rationality set by decision theory and logic we must not reject the theory as such but rather try to adjust the details of the theory to the evidence. For instance, errors in reasoning should not be translated as irrational moves. They should be translated either as misinterpretations of reasoning behaviour, in which case the details of the theory should be adjusted, or as performance errors in which case the predicted by the theory behaviour is not exhibited due to errors that may not threaten the reliability of the theory.

So we see that for Davidson, the main way to discovering a theory which adequately describes linguistic communication and rationality is conceptual, however, he acknowledges that empirical investigation is necessary for filling in the important details of the theory.

The specification of the basic characteristics of rationality

Since any empirical investigation of rationality would have to assume some kind of basic characterization of what rationality is, we need to understand here and be clear about Davidson's claims on what actually rationality is and about how much variation can be allowed to the characterization of rationality. The possibility of the Unified Theory is based on structures which are dictated by our concept of rationality, as Davidson rightly points out. Thus, the possibility and 'the entire structure of the theory depends on the standards and norms of rationality.' ([14], p. 126).

In the essay 'Incoherence and Irrationality', Davidson is trying to examine if irrationality is possible and actually what exactly irrationality is. Davidson concludes that the only clear case of irrationality is that of an agent holding beliefs (or any other propositional attitudes) that are inconsistent with other beliefs (or any other propositional attitudes respectively) based on his own principles, which suggests that there is an inner inconsistency. So, irrationality is defined as inconsistency within a set of beliefs. But then a problem arises: why must inconsistency be considered irrational?

Davidson thinks that the principles of decision theory and the basic principles of logic are shared by all creatures who have propositional attitudes and therefore there is no need for questioning whether or not one ought to subscribe to these principles. However, Davidson admits that decision theory and Tarskian theories of truth are true only of perfect logicians, which suggests that Davidson is aware of the fact that many times people fail to reason in accordance with the principles of decision theory and logic. But it seems that he would characterize such errors as performance errors rather than competence errors (see footnote 1 in the Introduction for a definition). For Davidson it is a given that we all share the same principles of rationality. He claims that unless we assume rationality, we cannot assign beliefs (or any other propositional attitudes) to human beings. Rationality here is thought of as entailing consistency within a set of beliefs. If we come across a person which seems to us to have contradictory beliefs, we should check whether we have interpreted him correctly rather than concluding that he is irrational. The thoughts themselves, according to Davidson, cannot be irrational.

Davidson's own words about the principles of rationality explain the above [Dav85, pp. 195–197]:

> These are principles shared by all creatures that have propositional attitudes or act intentionally; and since I am (I hope) one of those creatures, I can put it this way: all thinking creatures subscribe to *my* basic standards and norms of rationality. This sounds sweeping, even authoritarian, but it comes to no more than this, that it is a condition of having thoughts, judgments, and intentions that the basic standards of rationality have application.[. . .] I have greatly oversimplified by making it seem that there is a definite, and short, list of "basic principles of rationality." There is no such list. The kinds and degrees of deviation from the norms of rationality that we can understand or explain are not settled in advance. We

> make sense of aberrations when they are seen against a background of rationality; but the background can be constituted in various ways to make various forms of battiness comprehensible. [. . .] then it does not make sense to ask, concerning a creature with propositional attitudes, whether that creature is *in general* rational, whether its attitudes and intentional actions are in accord with the basic standards of rationality. Rationality, in this primitive sense, is a condition of having thoughts at all. The question whether a creature "subscribes" to the principle of continence, or to the logic of the sentential calculus, or to the principle of total evidence for inductive reasoning, is not an empirical question. For it is only by interpreting a creature as largely in accord with these principles that we can intelligibly attribute propositional attitudes to it, or that we can raise the question whether it is in some respect irrational. [. . .] Inner inconsistency is possible just because there are norms no agent can lack.

It is obvious in the above that Davidson considers rationality as a global characteristic which is shared by all creatures with propositional attitudes. This view appears in many forms in the rest of his work and, as we saw in the previous sections, in his work on Radical Interpretation this view plays a central role in the form of the Principle of Charity. However, for Davidson there is no definite list of general principles of rationality.

Summing up

Before starting the discussion on Radical interpretation I want to sum up the two significant points of Davidson's theory discussed in this section.

First, Davidson allows the possibility of empirical investigation of what rationality is, even though this is not the main aim of his work. Empirical as well as conceptual considerations are relevant in determining what rationality is; conceptual considerations lead to the conclusion that rationality is a normative concept and a presupposition of understanding human behaviour at all, and empirical investigation might lead to the determination of the basic characteristics of rationality. Davidson is mainly engaged with the task of building the foundations of a solid theory of meaning, which he is trying to achieve by answering the question of what is needed—and this is not available for empirical investigation—for such a theory to be possible. But he does believe that once a theory of meaning is achieved, the theory will—and should if it wants to be called scientific—be available for empirical investigation.

Second, Davidson allows enough room for variation in the specification of the basic characteristics of rationality. Davidson's view of rationality resembles the decision theoretic view; human agents are rational and they share the basic (global) normative characteristics of rationality. The entire structure and possibility of a unified theory of the kind Davidson seeks depends on these standards and norms of rationality. However, Davidson believes no list of norms would exhaust the basic characteristics of rationality as he understands it. Certainly, according to Davidson, an element of this list is consistency. Davidson conceives rationality as necessarily including preservation of logical consistency, therefore logical consistency should always be included in the list of norms of rationality.

Chapter 3

Michiel van Lambalgen and Keith Stenning

In this chapter, I will present the ideas of van Lambalgen and Stenning on rationality, interpretation, and human reasoning, as introduced in their recent book *Human Reasoning and Cognitive Science* [SvL08]. In [SvL08], van Lambalgen and Stenning investigate the relation between psychology and logic. Specifically, they reconsider the role of logic in human reasoning by examining and interpreting experimental data and results in the area of the psychology of reasoning, with the help of logical tools. Their ideas are laid against a background (described here in section 1.1) which is considered standard in the area of philosophy and psychology of reasoning. In this context, rationality is viewed as the fundamental characteristic which distinguishes humans from any other form of life, and is taken to be a highly normative concept.

3.1 Rationality, logic, and normativity

In this section I will explain the views of van Lambalgen and Stenning on rationality and on the role of logic and normativity in rationality and human reasoning. As we will see, van Lambalgen and Stenning vigorously argue that humans are rational, rejecting Wason's thesis (see section 1.1 for Wason's thesis). We will also look at their theory and their point of view about descriptive and normative theories and the role of classical logic as a normative standard of human reasoning.

3.1.1 Logic and normativity

What is made clear to the reader of [SvL08] already from the introduction, is that the role of logic in the investigation of human reasoning and cognition should be reconsidered. Logic, and more specifically classical logic, was traditionally thought to be the basic normative system against which human reasoning has to be judged. The paradoxical results of Wason's selection task came to change this picture, leading to the hasty conclusion that logic needs to be divorced from cognition. In [SvL08] van Lambalgen and Stenning claim that, after a careful examination of the assumptions which led logic and cognition to be divorced, we can bring the two back together as they should be.

Specifically, van Lambalgen and Stenning object to Wason's conclusion that reasoning is rational if it follows rules from a given and fixed set since, according to them, this view implies that logic should be viewed as irrelevant for cognition. The view that they support instead, is that reasoning is everywhere and sometimes it is not consciously processed but automatic. What their investigations have led them realize is that people may some times reason in ways which are inconsistent with their chosen interpretation which means that evidence for the interpretation which is independent of the performance is necessary when one uses a particular interpretation to explain performance.

Thus, van Lambalgen and Stenning claim that we should reconsider the role logic plays in human reasoning and try to use logic in the correct way, rather than abandoning it. One of the main roles of their work is to present new ways towards the understanding of reasoning in which logic can still be thought of as normative, *but* [SvL08, p. 11]:

> norms apply to instances of reasoning only after the interpretation of the (logical and non-logical) expressions in the argument has been fixed, and, furthermore, there are in general multiple natural options for such interpretations, even for interpreting the logical expressions. Thus, the reasoning process inevitably involves also steps aimed at fixing an interpretation; once this has been achieved, the norms governing logical reasoning are also determined.

The foundations of their investigations can be tracked back in Husserl, for who we should reserve a couple of paragraphs here. Let's go back to one of the questions posed in the previous section: Is there a definite and exhaustive list of rules which define rationality? In other words, are there unassailable and objective rules that define rationality? If we assume that logic provides such rules we have to give a justification of why we take these rules to be unassailable and objective. Husserl, in his *Logische Untersuchungen*, made a completely innovative suggestion: logic should be viewed not as a *normative* but as a *theoretical* discipline. By that he meant that logic as a theoretical discipline can, via theoretical statements of the type 'only such and such arguments preserve truth' and in combination with normative arguments of the type 'truth is good', derive the conclusion that 'only such and such arguments are good'. Thus, truth is an *a priori* concept and theories such as logic can determine only which rules are valid within the theory.

In the above picture normativity still has a central role. Logic consists of laws which are indeed unassailable and objective, but not in the same sense as before; logical laws are not providing absolutely valid norms, they are just providing valid norms relative to a particular domain and thus they are unassailable only as mathematical consequences of the structure of the domain studied.

In a similar way, van Lambalgen and Stenning claim that normativity plays a role in reasoning about reasoning tasks, but normativity enters the picture via an appropriate formal theory only after a stipulation of the meaning of the reasoning task has been given by following some *a priori* constitutive norms. This distinction will become obvious in the next sections where we will talk about the processes of reasoning *to* and *from* an interpretation and about the distinction between *regulative* and *constitutive* norms.

Finally, we should say that van Lambalgen and Stenning view logic 'as the mathematics of information systems, of which people are one kind' [SvL08, p. 14] and they suggest that this view

> helps in that it makes clear from the start that one's choice is never between "doing psychology" and "doing logic". Understanding reasoning is always going to require both, simply because science does not proceed far without conceptual and mathematical apparatus. [SvL08, p. 14]

We see that van Lambalgen and Stenning believe that logic and psychology are two tightly connected disciplines. We could situate their view in the 'naturalized epistemology' category that was discussed in section 1.1. Normativity is still in the picture, even though in a new form, and empirical scientific investigation is considered as the basic strategy for arriving at the right understanding of reasoning and rationality. Therefore, the 'psychologism' of van Lambalgen and Stenning,

> requires an account of how logic in its modern guise [multiplicity of logics] as a mathematical system is related to psychology in its modern guise as experimental science. [SvL08, p. 15]

In the following section, we will see how this new view of logical reasoning sketched here can be used in the psychology of reasoning according always to the theory which van Lambalgen and Stenning develop in [SvL08].

3.1.2 Reasoning *to* and *from* an interpretation

It is argued in [SvL08], that the subjects' behaviour in the reasoning tasks of the experimental psychology of reasoning (such as Wason's selection task) has been misinterpreted in the past. The most common interpretation of the results of the tasks, as we saw, has been that because of the failure in applying correctly the normative rules which would lead them to the right answers, subjects are irrational. As we read in [SvL04, p. 128] 'the selection task has in fact been used as a weapon against the application of linguistic and semantic theory in understanding subjects' reasoning.'

It is claimed by van Lambalgen and Stenning that this has been a major mistake and, in a different interpretation of the selection task results, they make the following claim: if we are more attentive to semantics we will realize that it is not that *people are irrational*, but rather that they *fail to impose the right*[1]

[1] Here 'right' means 'expected by the experimenter.'

meaning to the tasks, what makes them, more often than not, fail in giving the correct answer. Let us see what this different interpretation of the results amounts to, since this is one of the basic innovations that the van Lambalgen and Stenning approach offers.

First of all, it needs to be clear *why* semantics is so important for making sense of human reasoning in general and of deductive reasoning in particular. The importance attributed to semantics in [SvL08] is due to the observation that the subjects of the reasoning tasks seem to have serious difficulties in determining the 'right'—or, better, the expected by the experimenter—logical form from the grammatical form of the premises (see also discussion in [Cou08, p. 177]). Thus far, the standard (e.g. Wason's) interpretation of the task experiments has been that, since the subjects do not arrive to the correct, according to classical logic, answer of the various reasoning tasks, the subjects are behaving irrationally. In [SvL08], a quite different approach is taken: logical form and grammatical form are separated from each other.[2] Thus, the issue of interpretation now becomes important. Whereas Wason never thought that the subjects might give different interpretations to the reasoning task at hand, in the van Lambalgen and Stenning picture subjects are no longer expected to derive a unique logical form from a grammatical form. This means that many of the cases in which the subjects would have been deemed irrational in the past are not interpreted as evidence of irrationality in this picture.

The second chapter of [SvL08] is reserved, in order to make this idea a bit more clear, for presenting the many semantic possibilities that exist when a subject of a reasoning task is trying to impose a meaning on it: classical propositional logic, Kleene 3-valued logic, Łukasiewicz logic (fuzzy logic), probabilities, intuitionistic logic and deontic logic. It is nowadays a common truth that the universality of logic is a lost hope, and this is made evident by the multiplicity of logics that are apparently necessary in various fields of current scientific research. This fact cannot be neglected when aiming at explaining human reasoning. For instance, the idea that the reasoning of the subjects in Wason's selection task is based on the classical interpretation of the conditional as material implication should be reconsidered, and in [SvL08] it is clearly abandoned.

In this framework, it is also suggested that it is necessary to reconsider what validity is. Syllogisms like the Aristotelian which we saw in the previous section are of the form:

All As are Bs.
All Bs are Cs.
Therefore all As are Cs.

It is usually taken for granted that whatever terms you substitute for A, B, and C in the premises of the above syllogism, the conclusion will still follow from the premises if the premises are true. This means in other words that

[2] In [Cou08], Counihan, aligning herself with van Lambalgen and Stenning, dedicates the first two chapters to investigating the considerations involved in imposing logical form on various reasoning tasks performed by subjects coming from various educational backgrounds.

the syllogism is *valid* independent of the subject-matter of the substitutes for A, B, and C. This *domain-independent* character of inference patterns and of logic in general, is now questioned by the authors of [SvL08]. Logic is no longer considered as domain-independent and it is claimed that all valid schemata depend on the domain in which one reasons, or in other words, on the purpose of one's reasoning.

In the new picture of reasoning that van Lambalgen and Stenning introduce, we have the two following stages:

1. **Reasoning *to* an interpretation:** the establishment of the domain (and its formal properties) about which one reasons, or in other words the setting of the parameters (establishing what things are actually in the domain). In this step, the set of parameters which characterize the logic to be used for reasoning about something is defined. This set of parameters consists of three subsets:

   (a) choice of formal language

   (b) a correlation of natural language expressions to the formal language expressions

   (c) choice of a semantics for the formal language

   (d) choice of a definition of valid arguments in the language

   For setting the parameters, one may adopt a *credulous* or a *skeptical* attitude to the discourse. In a credulous attitude the the authority of the speaker is a presupposition of the truth of what is uttered. On the other hand, in a skeptical attitude one does not trust the authority of the speaker and thus various interpretations of the premises of a task are possible. Depending on the attitude one adopts, a certain logical form is assigned to the task.

2. **Reasoning *from* an interpretation:** the selection and application of the appropriate formal laws which will provide us with the suitable kind of reasoning, and eventually with an answer to the task, in our established domain. Examples of 'kinds of reasoning' are again *credulous* and *skeptical* reasoning. In the first, the interpreter should look for conclusions that are based on the ideal interpretation which makes an utterance true (by assuming mutual general knowledge), thus the authority of the speaker is a presupposition for the truth of what is uttered (defeasible logic is an example). In the latter, the interpreter should look for conclusions which are true in all interpretations of the premises. Thus, interpreter and interpretee have the same authority, and a single interpretation of the premises in which the conclusion is false is enough for raising an objection (classical logic is an example).

According to van Lambalgen and Stenning, during the step of reasoning *to* an interpretation two things are taking place: first, the domain in which someone reasons and its formal properties are established, and second, the logical form according to which the interpretation of the premises of a task will be interpreted is decided. Therefore, it is not any more taken for granted that the logical form will be determined by classical logic. In fact, the logical form is going to be determined by closed world reasoning[3]. Indeed, closed world reasoning is encountered very often in instances of everyday life reasoning. Reasoners who adopt closed world reasoning consider only a subset of the set of all models of the premises in order to reach a conclusion.

After deciding the logical form, an appropriate model is constructed in which the subject can judge the truth or falsity of utterances. Moreover, the hearer normally adopts a credulous attitude to the discourse, at least for the purposes of interpreting it. Van Lambalgen and Stenning present a non-monotonic logical model of this process of model-construction.

More attention is now paid to the process of interpretation[4] rather than that of derivation—which is the one that the psychology of reasoning has been mainly focusing in. Interpretation tasks are thus separated from derivation tasks.

3.1.3 What happens to validity?

Instead of presupposing that there is a unique logical system which specifies the meaning of the premises from which a conclusion is derived, it is now claimed that deciding what the right logic is for the purposes of interpreting a discourse depends on one's notion of truth, semantic consequence and more. For example, the truth of the premises of the Aristotelian syllogism

All men are mortal.

Socrates is a man.

Therefore, Socrates is mortal.

as well as its validity may vary across interpretations. This means that, depending on the logical form assigned by the subjects to a reasoning task, the defintition of validity changes. For example, if the classical logical form is chosen, then classical validity is assumed, whereas if a different logic is chosen to describe the logical form, then the corresponding to that logic definition of validity will be used. Actually, in [SvL08] it is suggested that most often subjects use closed world reasoning to reason about the task, in which validity of an

[3] In closed world reasoning it is assumed that lack of knowledge implies falsity; thus, when in a model no information is given about whether a positive sentence p is true, we assume that $\neg p$ is the case. In other words, if a positive sentence p cannot be derived from the knowledge we have, we assume that $\neg p$ is the case.

[4] What is meant here by interpretation is understood as the mapping of representation systems onto the things in the world that are represented, or in other words, "all the structures and processes which connect language to the specifics of the current context" [SvL05, p. 921]. So, interpretation decides for: which things in the world correspond to which words, which of these things are specifically in the domain of interpretation of the current discourse, which structural description should be assigned to an utterance, which propositions are assumed and which derived, which notions of validity of argument are intended and so on.

argument is defined as 'truth of the conclusion in all *preferred* models' rather than simply 'truth of the conclusion in all models.' This seems quite natural if one considers that people without any logical training will, more often than not, use a lot of external information in order to solve a reasoning task, which suggests they are not applying the classical definition of validity.

With this new picture of the relevance of semantics and logic in the psychology of reasoning and of the definition of validity being relative to a domain, a whole new area of research unfolds. No longer are the results of Wason's selection task—which has so much affected the development of psychology of reasoning—interpreted as Wason suggested. That is, material implication is no longer the right competence model for the task. As we read in [SvL04], a new range of questions which have very little—if at all—occupied researchers of the field in the past is put forward ([SvL04], p. 133):

> [. . .] what makes the original task so hard? What range of explanations could distinguish the hard and the easy versions of the task? How do subjects construe this task? Do they all construe it the same way? What range of possible interpretations might subjects reasonably have of the rule? Is Wason's logical competence model a reasonable interpretation of what we know about the meaning of the rules used?

These questions seem to be natural if one considers that the 'if . . . then' rule of classical logic does not correspond exactly to the 'if . . . then' of natural language. The natural language 'if . . . then' seems to work in a rather paradoxical way if you judge it by the standards of the classical logic implication rule. For instance, in classical logic the implication 'If $1+1 = 3$ then the Pope is an astronaut' is true since both antecedent and consequent are false, whereas in natural language the implication is taken to be false since antecedent and consequent are not related to each other in any way which would make the implication true. Similar problems are faced with implications where the antecedent is false and the consequent is true. These examples suggest that it is a mistake to identify the conditional 'if . . . then' with material implication.[5]

3.1.4 Marr's three levels of inquiry

In the new picture of reasoning that is offered in [SvL08], David Marr's three levels of cognitive inquiry play an important role. These three levels are:

1. identification of the information processing task as an input-output function

2. specification of an algorithm which computes the function

3. neural implementation of the algorithm specified

[5] The reader who wishes to know more about the paradoxical nature of the material implication should refer to [AD97] and [RTtMF82].

Van Lambalgen and Stenning explain in [SvL08, p. 298] that logical reasoning is similar to information processing in such a way that the former can be better dealt with when the latter is understood and taken into consideration. Let us look at these three levels one by one, in order to understand their importance and their relevance to a better understanding of logical reasoning. We will start with an example taken from Marr, and we will then look at the similarities of this example with the case of logical reasoning.

The example, offered also in [SvL08] for illustration, is that of vision (vision is characterized by Marr as extracting 3D shapes from 2D retinal arrays). For our purposes imagine the following situation: you are looking at an object in front of you; what happens then is that a 2D image is reflected on your retinal array, and then your brain 'translates' this 2D image of the object into the 3D form that corresponds to that image; that is, from surface information you derive a 3D shape. In this situation the first level as suggested by Marr is about defining a function which will take as input the two-dimensional image (from the retinal array) and as output the corresponding 3D shape thus describing an ideal construct.

In the second level an algorithm must be specified, which will compute the function defined in the first level. For this it is necessary to choose the representation (mathematical) languages in which the entities that the algorithm will operate on are expressed. When these two levels have been dealt with, we have in our hands a competence model against which we can judge performance of individuals. An important thing to keep in mind is that the competence model can only be approximated. In more technical terminology this means that the algorithm will allow for *graceful degradation*. That is, the algorithm must be able to cope with small deviations; the closer the input is to the ideal input of the function defined in the first level, the closer the output will be to the ideal output.

The third level is about the neural implementation of the specified algorithm, that is, about how actually the information is processed in the brain. It is outside the purpose of this thesis to touch on this level, therefore we will leave any discussion on that out.

Let us compare now the example of vision to that of logical reasoning. The first level, that is, the definition of the input-output function, corresponds to the stipulation of a competence model—what we have earlier called logical form—against which performance of individuals will be judged. Remember that this competence model is no longer thought to be determined by classical logic. The input of the function will in this case be the premises and the output the conclusion. The competence model is expressed in a mathematical language and it defines the 'ideal' situation, that is, the situation in which there are no constraints about the world that would limit the available data. It is important to stress here that the determination of the competence model (the logical form) is not done empirically; it is through cognitive considerations that we arrive at it which means that it is determined *a priori*.

In the transition from the first to the second step of logical reasoning, there are various constraints that need further consideration. These constraints are hypotheses about the world and about the process generating the data which one has to arrive at after the processing of the available information. It is necessary here to take into consideration all constraints since it should be expected that the actual situation we are in deviates from the ideal situation described by the logical form as introduced in the first level. We arrive thus at the second step, where we have to specify an algorithm which will describe possible performance. This algorithm must allow graceful degradation, as we saw earlier, and in the case of logical reasoning this suggests that the closer the actual premises are to the input required by the competence model, the closer the conclusion will be. Obviously performance will only be considered optimal when it proceeds in accordance with such a 'graceful' algorithm. After the specification of the algorithm there occurs the extraction of information.

3.1.5 Constitutive and regulative norms

The ideas developed about logical reasoning in the previous section, can be explained by a distinction Searle drew between constitutive and regulative rules[6] in his *Speech Acts* [Sea70] and Kant's ideas on constitutive processes from his *Critique of Pure Reason* [Kan98].[7]

This important distinction between regulative and constitutive rules can be illustrated by the following examples. Imagine you are involved in a chess game. What is that makes a chess game a chess game? Its rules. If there were no chess rules, there would have been no chess game. These kind of rules are what Searle calls constitutive rules, which 'do not merely regulate' but 'create or define new forms of behaviour' [Sea70, p. 33]. In the case of chess, the rules of the game create the very possibility of playing the game, and for that reason these rules are called constitutive rules.

Now imagine a different situation. You are driving a car in a city so you have to follow the city's traffic rules. Are these rules constitutive for the act of driving the car? No, because the rules are not creating in this case the possibility of driving a car. One can drive a car even without the existence of such rules. The traffic rules simply regulate driving behaviour, thus they are called regulative rules. According to Searle, in cases 'where the rule is purely regulative, behaviour which is in accordance with the rule could be given the same description or specification (the same answer to the question "What did he do") whether or not the rule existed' [Sea70, p. 35].

How does logical reasoning and its steps as described in the previous section fit into this picture? In other words, how can we use the distinction between regulative and constitutive norms in explaining logical reasoning? And how is this all connected to rationality? The stipulation of the competence model during the first step of logical reasoning that we saw in the previous section is actually the creation of the reasoning problem by the constitutive norms. As mentioned earlier, the competence model built by the constitutive norms is

[6] Searle based his distinction on Rawls's respective distinction in the area of ethics [Raw55].

[7] See also [Ach07, Ch. 4] for a presentation of this distinction.

not determined empirically; the constitutive norms are given *a priori* and are necessary for the existence of the model.

Think of the following example. A reasoning task, say Wason's selection task, is presented to you in natural language and you are asked to solve it. How do you interpret the implication rule of the selection task as stated in natural language? It depends on the constitutive norms which actually build the task. According to van Lambalgen and Stenning the task allows for multiple interpretations of the rule, therefore various people may build it in a different way by using different constitutive norms. One interpretation might be the one suggested by classical logic, but there are many other possibilities. This procedure of deciding the logical form with the help of the constitutive norms is what the authors have called *reasoning to an interpretation.*

After the stipulation of the competence model, that is, after building with the help of the constitutive norms the logical form of the task, there comes the second step of reasoning: specifying which arguments are valid in that logical form. The second step is necessary for understanding which conclusions one is allowed to draw. Here is where regulative norms come into the picture. In the transition from the first to the second step, one has to consider the constraints placed by the knowledge we have about the world at that moment. This is a crucial point, since the actual situation will deviate from the ideal situation described by the competence model of the previous step. In choosing the appropriate set of regulative norms, we specify the algorithm which allows for graceful degradation and describes possible performance. Note that the set of regulative norms is not unique for each logical form; at least not when we are dealing with closed world reasoning where the set of regulative norms depends on the knowledge about the world one has at a certain moment when trying to make an inference. The procedure of this second step, that is the drawing of the inferences according to the regulative norms, is what the authors have called *reasoning from an interpretation.*

Thus, we see that in [SvL08] an innovative model of logical reasoning is proposed. Judgments about human rationality do not depend any more on pre-constructed competence models and inference rules, and the expectations one has from the reasoners are less constrained; multiple interpretations are allowed and therefore a more mindful and scrupulous judgment of human reasoning behaviour is possible.

3.2 Empirical investigation

We will now see how the above analysis fits in the empirical investigations of van Lambalgen and Stenning. In chapter 4 of [SvL08], van Lambalgen and Stenning unfold 'the mother of all reasoning tasks' as they call it, namely Wason's selection task, and they look at it from the different point of view which was discussed in the previous sections. It is claimed that it is not correct, as Wason did, to decide *a priori* which is the right logic for solving the selection task; more attention should be paid to the 'contrast between what different subjects do in the same version of the task' (p. 278) rather than focusing only on the universal element characterizing the behaviour of the subjects. Systematic differences in behaviour can be a source of precious information about how people reason and, according to van Lambalgen and Stenning, an analysis based in comparing reasoning processes might prove to be easier than providing an absolute one.

As we read in [SvL08, p. 123],

> An immediate corollary of the formal stance of taking the multiplicity of interpretations seriously is the empirical consequence of needing to take individual differences seriously. Subjects do different things in the experiments, and this has so far been treated as simply stringing them out along a dimension of intelligence. They are all deemed to be trying to do the same thing, but succeeding or failing in different degrees. If there are many interpretations and each poses a qualitatively different task, then subjects are not even trying to do the same thing (at least at any finer grain than understand what the hell they are being asked to do), and suddenly the data and the theoretical demands become far richer.

Following the above idea, the authors of [SvL08] present Wason's selection task to the subjects and they examine closely the data collected. The experimental data is in the form of recorded Socratic tutorial dialogues between the experimenter and the subject, and much attention is now paid to all the expressions used by the subjects in their struggle to impose meaning on the descriptive task, since in the new framework the process of interpretation is equally important as that of derivation. An attempt is made to understand how the subjects access the process involved in solving the task in order to draw a full theory of performance in the task which would be able to explain how reasoning is related to the individual experiences of the subjects.[8]

By examining these tutorial dialogues, van Lambalgen and Stenning are trying to support the semantically based predictions of their theory. Some of the issues which they are looking at are described below.[9]

[8] In addition to the selection task, the experiments include a two-rule task in which again, like in Wason's selection task, four cards are laid in front of the subject which have an 8 or a 3 on one of its sides and a *U* or an *I* on the other side, and they look like what you see below:

$\boxed{U}$ $\boxed{I}$ $\boxed{8}$ $\boxed{3}$

The subject is then provided with two rules and is told that these rules apply only to these four cards and that exactly one rule is true. The subject has to decide which, if any, of the four cards must be turned in order to decide which rule is true. Unnecessary cards should not be turned. The rules are given as follows:

1. if there is a U on one side, then there is an 8 on the other side.
2. if there is an I on one side, then there is an 8 on the other side.

The classical logical competence solution of the task is to turn just the 3 card.

[9] For a more detailed treatment see [SvL08, pp. 49–86].

Distinction between descriptive and deontic tasks

Deontic rules are usually expressed with shoulds, oughts, or musts, and cannot be falsified; they can only be violated. For instance the rule ‘If you are under 18 years old you must not drink alcohol’ is a rule that can be violated if there is some exception to the rule, however the rule is still valid even if there is an exception to it. If the rule of Wason’s selection task is interpreted as a deontic rule by a subject, then an exception to the rule is not enough to falsify it, and actually there are many cases in which the subjects of the task seem to interpret it deontically.

On the other hand, descriptive rules are of a different kind; they can be true or false. In the case of a descriptive rule, a single exception to the rule is enough for falsifying it. It is claimed by the authors of [SvL08] that the original descriptive task, as Wason proposed it, creates many more problems for the subjects trying to solve it than a corresponding deontic task would create. Therefore, in order to understand why people fail to succeed in the descriptive task, we should focus on versions of the task that are clearly descriptive. For a formal analysis of the differences between deontic and descriptive tasks one can refer to [SvL08, pp. 53–58].

What truth and falsity is (according to the subjects)

In both the selection task and the two-rule task (see footnote 8 in this section for an explanation of the two-rule task), many subjects seem to have a non-standard understanding of what truth and falsity is. For instance, non-falsity does not necessarily imply truth for all subjects. To give an illustration of this, in the case of the selection task, if the subject decided to turn the A and 7 cards then this would have been enough to show that the rule is not false. However, it is observed in the experiments that subjects are often not convinced that proving the rule non-false implies that it is true. This may happen because subjects sometimes think, although this does not seem to be suggested by the instructions of the task, that the rule should apply to a much larger population of cards. Thus, if a counter-example is not found among the 4 cards, it might be the case that there exists one in a ‘bigger’ model. For this, one would need information which comes outside of the four-card model. Similarly, in the two-rule task, some subjects believe that falsifying one of the two rules does not necessarily imply that the other rule is true.

The authority of the experimenter

Often, as the experimental data shows, subjects of the selection task take it for granted that the rule is true, since it is uttered by the experimenter whom they consider as an authority on the subject. Thus, the subjects do not try to judge whether the rule is true or not, but rather they try to find out which cards make the rule be what it is, namely true.

Cards as sample from a larger domain

It is often the case that the subjects of the experiment consider the sample of 4 cards not enough for judging the correctness of the rule. A rule for them is something that applies within a sufficiently wide range of cards. Even though subjects are told that the rule applies only to the four cards, some times they interpret this piece of information as 'the rule applies only to the four cards now, but these four cards belong to a bigger set of cards to which the same rule applies.'

All the above points show that the empirical data is richer than what had been assumed by Wason, and that there are various interesting ways to look at it. Although this thesis does not offer a sufficient elaboration of concrete examples of experimental data, two such examples are briefly examined from the view point of a combination of the Davidson and the van Lambalgen and Stenning account in appendix A.

3.3 Criticism and replies

The account presented in the present chapter has been standing against some criticism.[10] The criticism goes roughly as follows: if interpretation takes such an important role as it is given in [SvL08], then psychology of reasoning will cease to be of interest as a branch of psychology, because the methodology of psychology is to fix all variables except the one that is tested, in this case performance on the given reasoning problem. If interpretation cannot be fixed, the field ceases to be experimental. The interpretation not being fixed also raises a more general, though connected with the previous, point of criticism, namely the threat of relativism.

The source of the above criticism is the local-logics approach to rationality that van Lambalgen and Stenning suggest. In the new picture of human reasoning presented in [SvL08], the logical form of an argument is not directly interpreted in classical-logic terms; rather, there are various possible interpretations indexed to the specific circumstances of a reasoning task and to the reasoning abilities of the person trying to solve the task. Because of this picture of human reasoning, it has been claimed that the theory supporting it dismisses the experimental character of psychology and that it is relativistic.

For the first part of the criticism, the problems that the non-fixed interpretation brings to psychology of reasoning as a field, a reply is offered in [SvL08, p. 51]:

> Given the existence of interpretational variety, the right response is richer empirical methods aimed at producing convergent evidence for deeper theories which are more indirectly related to the stimuli observed. What the richness of interpretation does mean is that the

[10] The criticism has not yet appeared in print, but it comes out of many discussions the authors of [SvL08] are having with colleagues about their account.

> psychology of reasoning narrowly construed has less direct implications for the rationality of subjects reasoning. What was right about the earlier appeals to interpretational variation is that it indeed takes a lot of evidence to confidently convict subjects of irrationality. It is necessary to go to great lengths to make a charitable interpretation of what they are trying to do and how they understand what they are supposed to do, before one can be in a position to assert that they are irrational. Even when all this is done, the irrational element can only be interpreted against a background of rational effort.

It is suggested here that keeping the parameter of interpretation non-fixed should not be seen as a problem, but rather as a suggestion for psychology of reasoning to become less narrow; this means that psychology of reasoning should avoid conclusions regarding the irrationality of subjects before all the necessary evidence is collected that determines what interpretation the subjects give to a given task. What is the point of fixing the interpretation of a task when all evidence suggests that subjects often use alternative interpretations in order to reason about tasks? Van Lambalgen and Stenning have a point here. However, the more general point of criticism, that of the threat of relativism in such a picture of logical reasoning, is not yet dealt with to a sufficient extent.

Indeed, if relativism is true about reasoning, then there are no standards of rationality that characterize all humans, which means that rationality is not a normative concept. Each person can in that case decide whether a rule is correct for reasoning under some specific circumstances (in the process of reasoning *from* interpretation). And for each person, an experimenter may say that the person interprets the task using *this* rather than the *other* logical form (in the process of reasoning *to* an interpretation). In other words 'anything goes' in this picture. For instance, I can claim that a rule I come up with in a specific circumstance is appropriate for reasoning in that circumstance in the same way I can decide which colour is the nicest according to my personal aesthetic judgment. Or I can claim that an interpretation that a subject gives to a task is expressed by this rather than the other logical form but this claim might be insufficient in the light of restricted or misleading evidence.

In fact, if we assume that van Lambalgen and Stenning are indeed relativists, we will have to conclude that they support the view that what counts as rational is indexed to each individual and is therefore different for each human being. In that case, any rule of reasoning that a person comes up with is taken as 'correct' under the specific circumstances in which that person is reasoning. Without doubt, as will be explained in what follows, van Lambalgen and Stenning are not following this line of thought.

First of all, van Lambalgen and Stenning allow for the possibility of performance errors both in the process of reasoning *to* and reasoning *from* an interpretation on the part of the subjects. Specifically, they claim that 'It is possible [...] to make errors in reasoning: the parameter settings may be inconsistent, or a subject may draw inferences not consistent with the settings' [SvL08, p. 51]. These errors cannot exist in a relativistic picture of reasoning. It may be the case that logical reasoning is relative to a particular cognitive task, *but* the task imposes a norm upon the subject engaged in the task. Otherwise inconsistencies would not be considered as errors but as part of the rationality of individuals. As we read in a section of [SvL08] called ‘On why we are not postmodern relativists’ [SvL08, p. 352],

> it is a general norm of rationality that the subject should try to perform a task as well she can, consistent with reasonable costs. What this means for a concrete task has to be determined from the competence model for that task. Thus we end up with a picture of the normative status of logic which is very much like the one sketched by Husserl [. . .]: logic itself is a theoretical discipline, proving consequences of choices of parameters, and norms come from the outside, by the choice of a particular logical form.

Apparently, norms are still in the game, only they are placed in a different level, the level of reasoning *from* an interpretation or else the level where we derive conclusions from a task with a given logical form. This does not mean in any possible way that norms do not play any role in human reasoning. What the theory of van Lambalgen and Stenning offers is a different way of looking at normativity, rather than abandoning normativity. And what is definitely supported here is that classical logic should not be thought of as the logic that decides what are the norms of rationality. *Logic* is a more general theoretical field that should not be identified with *classical logic*. In this—wide—sense, logic is (and should be) indeed relevant to cognition and the psychology of reasoning. It is only the role of *classical logic* that is being reconsidered.

In support of the above, we read in [SvL04, pp. 144–145],

> And of course dismissing logic on the grounds of prescriptiveness is hopelessly simplistic. First, it is simplistic because *without some standard, description of behaviour is impossible*. Second, subjects themselves have prescriptive standards that they apply to their own reasoning processes. One has only to listen to the Socratic dialogues to hear subjects upbraiding themselves for what they judge to be lapses in their standards of reasoning. There is no way for psychology to escape the description of prescriptions. (emphasis added)

It seems that the account of van Lambalgen and Stenning supports a version of the rationality thesis which, not only is not weaker than the common rationality thesis—which says that all (normal) people are rational—but it is even stronger, since it presents humans not only as rational enough to be able to solve all tasks according to the norms, but also as able to assign a different logical form, according to the circumstances, to each reasoning task. Thus, not only are people able to reason according to the laws of classical logic when they decide this is necessary, but they are even able to reason according to other kinds of logic equally important for solving everyday reasoning tasks.

The basic difference of the van Lambalgen and Stenning account from the standard accounts of rationality and reasoning is that more possibilities are considered for assigning a logical form to a given task. So far, it has been assumed that classical logic is the only normative logical form. In this new picture, if tasks are stated in a way which allows various interpretations (like the Wason selection task), the subjects are free to decide for an appropriate interpretation of the task from within a range of choices. This range is determined by contextual parameters, by the assumption of rationality, and by the assumption of general mutual knowledge. If the tasks, on the other hand, are explicitly stated in a way that decides in advance the interpretation (and thus the logical from) of the task for the subject, then the subjects are able to solve the task according to the norms that the task imposes on them.

Presenting an explicit list of the general norms of rationality would lead to the determination of the appropriate logical form for any task, and would certainly be an immediate answer to the criticism about relativism. Such a list is indeed not offered in [SvL08], but as discussed also in the previous chapter on Davidson, to identify this list would be an impossible task. On the other hand, it is not impossible to identify the norms of rationality which are relative to a specific task; these norms would be the regulative norms, that is, the rules which, given a certain logical form, determine which inferences are valid given the premises of a task. For example, if the logical form chosen is classical logic, then the only valid inferences for solving the task would be those defined by classical logic. However, it is less clear in [SvL08] how the logical form is decided by the subjects.

This is apparently not enough for convincing the skeptical reader, and at this point Davidson's theory might be a way to support van Lambalgen and Stenning's theory. Some kind of a *Principle of Charity* might be appropriate for ensuring that human beings impose an 'appropriate' or else 'normative' logical form to reasoning tasks. The assumption of a shared rationality among humans implies that the choice of the logical from is (*a priori* and) not absolutely free.

Chapter 4

Comparison

The two accounts examined in this thesis apparently have much in common. However, there are also some fundamental differences between the two. Looking at these similarities and differences will facilitate our understanding of how the two accounts fit with each other and of how they could be combined in order to give a more complete picture of rationality and human reasoning. This chapter is reserved for this comparison.

4.1 The aims of the two accounts

The two accounts seem to have different aims although they share roughly the same (global) notion of rationality. Davidson, in the part of his work which is examined in this thesis, aims mainly at understanding what rationality is via conceptual considerations[1] (e.g. by building a theory, such as Radical Interpretation, of how-to-build a theory of meaning) whereas van Lambalgen and Stenning aim at making more explicit, via empirical considerations (experiments and interviews with subjects), which norms of rationality are relevant in specific instances of reasoning. This difference in focus does not set the two accounts apart. On the contrary, if the two accounts are combined, they might lead to a more complete picture of many aspects of rationality, human reasoning and linguistic communication.

4.2 What is rationality?

As discussed earlier in this thesis, any empirical investigation of rationality would have to assume some kind of basic characterization of rationality. Since both accounts discussed here include some kind of empirical investigation of rationality, we need to state explicitly what is the characterization of rationality

[1] Without denying the significance of empirical investigation.

provided by each account in order to create a basis for comparison. Is Davidson talking about 'rationality' in the same way van Lambalgen and Stenning do?

4.2.1 Classical logic, interpretation and rationality

The first point we should make concerns the status of classical logic in Davidson's account of rationality. As we already saw, Davidson maintains that a definite list of norms of rationality does not exist. He claims that in general we all share the same kind of rationality, without actually making explicit what principles define rationality. The only clear principle is that of overall coherence and consistency of beliefs. In fact, in both accounts under investigation the assumption of rationality trivially includes consistency and coherence. In addition, for Davidson, a basic characterization of rationality would include the principles of logic and decision theory. It is a given, for him, that the laws of classical logic are included in the norms of rationality.

In section 2.3, we noted that Davidson allows enough room for variation in the specification of the basic characteristics of rationality. Although there is no definite list of norms of rationality which we can use under all circumstances in order to understand whether a person is behaving rationally or not, there is a possibility to specify the basic characteristics of rationality necessary for making sense of each other under specific circumstances determined by the given context. Davidson allows some room for variation, nevertheless apparently classical logic has a central position in his picture of rationality, which is a fundamental difference from the account of van Lambalgen and Stenning, who put classical logic at the same level as other kinds of logic.

As was made obvious in the discussion of van Lambalgen and Stenning's account, logic plays an important role in the empirical investigation of rationality. Different reasoning behaviours can be locally explained by different logics. This local-logics approach seems to explain and add something to the conceptual considerations of Davidson about rationality. In a way, the local-logics approach confirms the claim of Davidson that there is no definite list of norms of rationality.[2] It seems that this approach provides some answers as to how norms define a task and its solution by introducing the distinction between constitutive and regulative norms. This distinction is used to introduce a new way of looking at human reasoning, namely looking at it as comprised by two parts: reasoning *to* and *from* an interpretation.

Reasoning *to* an interpretation, that is, the step where one has to assign logical form to a reasoning task, is the hardest part when trying to reason under some given premises. Once the logical form is determined, the logic which one will follow in order to reason according to the norms is settled. This means that even though for instance classical logic is not universal it can still be the competence model within a certain context which is constrained by the choice of interpretation that the logical form suggests.

Stenning and van Lambalgen [SvL08, p. 353] claim that

[2] 'I have greatly oversimplified by making it seem that there is a definite, and short, list of "basic principles of rationality." There is no such list' [Dav85, p. 196].

> it is the very indeterminacy of natural language meaning which makes imposition of logical form necessary: the assignment of logical form is constitutive of meaning. For example, the conditional [in Wason's selection task] can have many different meanings depending upon context; and the sparser the context is, the fewer clues one has about the intended meaning.

Here it is suggested that in order to assign meaning to a rule such as the one given in Wason's selection task, which is phrased in natural language, we have to impose a logical form on the rule. The choice of this logical form is inferred by the subject who is trying to solve a task given the contextual parameters and the specific wording of the task. We see that in this new picture of logical reasoning, interpretation has a central position.

As we saw in the previous chapter, van Lambalgen and Stenning suggest that the logic which best describes everyday human reasoning and interpretation should be non-monotonic. Davidson, although possibly without realizing it, seems to support a quite similar idea. Especially in his later work, and in particular in the essay 'A Nice Derangement of Epitaphs', Davidson suggests that during each step of linguistic communication—during each single act of interpreting a new utterance—the people involved update their 'theories' with any new information necessary in order to make the correct interpretations. This is exactly the basic assumption of non-monotonic logic: in view of new information one can change the interpretation chosen earlier and of course, if we take it one step further, one can retract any previously attained conclusions. Thus we can see here another connection between the two theories; van Lambalgen and Stenning's account uses non-monotonic logic as a logic which best describes human reasoning, and later Davidson describes what takes place during linguistic communication in a way which is very similar to non-monotonic logic.

Of course, Davidson was particularly interested in how linguistic communication, and in particular interpretation, can be explained, whereas van Lambalgen and Stenning go one step further and look at the inferences one may draw given a choice of an interpretation of a particular reasoning task. Therefore, Davidson's work can be compared to the first part of the reasoning process as van Lambalgen and Stenning describe it, namely reasoning *to* an interpretation. When it comes to the second part, reasoning *from* an interpretation, Davidson does not say much in the body of literature examined in this thesis.[3]

Reasoning *from* an interpretation is tightly related to the notion of validity. On the other hand, Davidson does not seem to care much about the notion of validity in the work examined here. He seems to accept the classical logical version of validity, and this would mean that given the chosen interpretation, all subjects follow the same rules of inference as they are defined by classical validity. Van Lambalgen and Stenning on the other hand, offer a new definition of validity which is different from the classical logical one. This new picture of

[3] Davidson's work on the explanation of action deals to some extent with the part of reasoning *from* an interpretation. Human actions are explained with reference to the attitudes and beliefs of humans, thus they are explained to an extent by reference to rationality.

validity was discussed in section 3.1.3. The regulative norms, which are chosen after the specification of the logical form of a task, define what are the valid inferences during the step of reasoning *from* an interpretation. We should note again that the set of regulative norms is unique for each logical form, but the selection of logical form depends (in closed world reasoning) on the knowledge one has about the world at a certain moment when trying to make an inference. Thus, the notion of classical validity is no longer unique in this new picture of logical reasoning. This is one of the most important innovations that van Lambalgen and Stenning's approach offers, and one that could make Davidson's account more complete.

4.2.2 Rationality as a presupposition of interpretation: The Principle of Charity

The second point worth noting is that both accounts suggest that there is no way to arrive through empirical investigation to the conclusion that humans are not rational. Rationality is a *presupposition* of linguistic communication, therefore one has to start from assuming that we all share the same kind of rationality, if we want to make sense of each other. There is an important distinction between the question 'Is this person rational?' and the question 'Does this person behave rationally under certain circumstances?' The first one implies that rationality is not a common characteristic of all humans (since we question whether one of them is rational), while the latter implies that a human might deviate from the standards of rationality under certain circumstances (and thus behave irrationally in that instance of reasoning) although he is generally rational. Davidson claims that the first question does not make sense to ask at all. What matters most for successful linguistic communication and for finding out what makes interpretation possible, is being aware of the necessity of sharing the same kind of rationality rather than trying to come up with the exact list of the norms of rationality—which does not exist in a definite form. All human beings share the same norms of rationality, although the specific background (list of norms) of rationality against which we may judge humans as being rational or irrational at certain instances of communication depends on the specific context. Recall Davidson's words:

> The kinds and degrees of deviation from the norms of rationality that we can understand or explain are not settled in advance. We make sense of aberrations when they are seen against a background of rationality; but the background can be constituted in various ways to make various forms of battiness comprehensible.

Just because there are norms of rationality which no human can lack we may say that, in certain situations, some humans behave irrationally (e.g. are not consistent in their reasoning) due to performance, or other kinds[4] of errors.

[4] Davidson mentions compartmentalization of the mind as a presupposition of explaining irrationality in [Dav85, p. 198].

There can be only momentary lapses of reason. Thus, Davidson would claim that there is no way we can conclude from experiments such as Wason's selection task, that the subjects of the experiment are irrational.

In a similar way, van Lambalgen and Stenning claim that all humans are rational. The only conclusion one could get from an empirical investigation, and the one that van Lambalgen and Stenning's investigation attempts to lead us to, is that errors in human reasoning can be explained in multiple ways; there is no single logic against which we can give such explanations. In other words, they claim—like Davidson—that there is no hope of defining explicitly a list of norms of rationality appropriate for explaining human reasoning under all circumstances. We do know that we all share the same kind of rationality but we need to narrow down the background of rationality against which our performance is judged given certain reasoning tasks.

The above discussion is immediately connected to the Principle of Charity which considers rationality to be a global characteristic of all human beings. Davidson [Dav82, p. 138] claimed that

> unless we can interpret others as sharing the vast amount of what makes up our common sense we will not be able to identify any of their beliefs and desires and intentions, any of their propositional attitudes. The reason is the holistic character of the mental.

Charity, we concluded, is about agreement (in general) between interpreter and speaker about their shared environment.

The Principle of Charity seems to be significant also in the picture of rationality that van Lambalgen and Stenning propose. For instance, in the account of van Lambalgen and Stenning, there is a presupposition made that the logical form of the selection task should not be determined *a priori* and thus subjects should be interpreted as rational beings which assign different interpretations than the expected ones, not only to different versions of reasoning tasks such as the selection task, but also to the same version. This means that according to van Lambalgen there is some kind of a Principle of Charity that we should apply if we want to really understand rational subjects' linguistic behaviour. They have claimed as we saw that 'without some standard, description of behaviour is impossible'; the Principle of Charity seems to be this 'standard'.

After having discussed the above, it seems quite obvious that van Lambalgen and Stenning's approach to logical reasoning has a significant global element, just like Davidson's. Namely, the presupposition that all humans are rational and that understanding human reasoning is impossible without having some kind of global standards on which one can base the description of human behaviour.

First meaning

There is another important similarity between the picture van Lambalgen and Stenning offer as an explanation of what is taking place during the reasoning process of the selection task subjects, and the picture Davidson offers in the essay ‘A Nice Derangement of Epitaphs’ as an explanation of linguistic communication and interpretation. In both cases, successful communication is thought to be achieved when the speaker’s intended meaning coincides with the hearer’s interpreted meaning.

Before explaining further this similarity, let us recall what ‘first meaning’ is for Davidson. First meaning includes the conventional literal meaning but goes beyond that; it is the first-in-order meaning according to the intentions of the speaker. If the conventional literal meaning is intended by the speaker, then first meaning and literal meaning coincide.

Since first meaning depends on the intentions of the speaker, Davidson claims that it cannot be governed by learned conventions or regularities. It is impossible to learn in advance all the potential intended meanings that a random speaker might come up with. According to Davidson, what is going on instead is that each time an utterance is made, the hearer is updating the theory that he has at that point—his prior theory as Davidson calls it—with the new intended meaning of the speaker as the hearer conceives it. If the intended meaning of the speaker coincides with the hearer’s interpretation, then successful communication is achieved.

Applying Davidson’s picture on the results of Wason’s selection task, we can conclude that successful communication is not achieved between Wason and the subjects of the task, since Wason’ intended meaning of the task does not coincide with the meaning assigned to the task by the subjects. It should by now be obvious that it is not appropriate to interpret this failure of communication as irrationality on the part of the subjects. Remember that it is only by assuming rationality that interpretation becomes possible.

4.3 Graceful degradation and logical reasoning

In [SvL08, p. 350], the following question is raised: How can we ‘define a notion of “graceful degradation” that makes performance an approximation of the ideal norm’? Recall that, according to van Lambalgen and Stenning’s account, the algorithm which describes possible performance during a reasoning task must be able to cope with small deviations, or in other words, the algorithm must have the property of graceful degradation: the closer the input is to the ideal input of the function which defines the logical form of the task, the closer the output will be to the ideal output. Or if you prefer, the closer the actual premises of a reasoning task are to the input required by the competence model, the closer the conclusion will be. Performance will be considered optimal only when it proceeds in accordance with such a ‘graceful’ algorithm.

This picture of logical reasoning resembles the picture of linguistic communication as presented in later Davidson. In [Dav86, p. 442] we read:

> The passing theory is where, accident aside, agreement is greatest. As speaker and interpreter talk, their prior theories become more alike; so do their passing theories. The asymptote of agreement and understanding is reached when passing theories coincide.

The convergence of passing theories described in the above passage, can be compared to the first part of the algorithm as described by van Lambalgen and Stenning. That is, how close the input of the algorithm is to the ideal input is determined by the extent to which the passing theories of speaker and hearer coincide. It seems that linguistic communication, in the way Davidson perceives it, has the property of graceful degradation.

By making an analogy with Davidson's prior and passing theories for speaker and hearer/interpreter,[5] we can express prior and passing theories of experimenter and subject during Wason's selection task. Of course, during linguistic communication speaker and hearer exchange their roles. Sometimes the speaker becomes the hearer and vice versa. For our purposes, hearer will be the subject (trying to interpret the reasoning task) and speaker will be the experimenter. By analogy to hearer's and speaker's prior and passing theories, we have the subject's and the experimenter's prior and passing theories:

Prior and passing theories as defined by Davidson

1. *Prior theory for the hearer:* The theory with which the hearer is equipped in advance of the conversation in order to interpret the speaker's utterances. This theory is adjusted to evidence prior to communication such as character, gender, role etc. of the speaker. The prior theory might not be sufficient for yielding the desired (by the speaker) interpretation.

2. *Prior theory for the speaker:* The theory which the speaker expects the hearer to be equipped with before any conversation begins. This theory of course may be different to the actual prior theory of the hearer.

3. *Passing theory for the hearer:* The theory which expresses how the hearer actually interprets the speaker. In other words it is the prior theory updated with any elements that were necessary for the correct, according to the hearer, interpretation of the speaker's utterances.

4. *Passing theory for the speaker:* The theory which the speaker intends the hearer to use in order to interpret the speaker.

[5] We discussed Davidson's later theory in section 2.2.1. To summarize it here, both speaker and interpreter (hearer) are equipped, at the beginning of a conversation, with a *prior theory*. For the interpreter, the prior theory 'expresses how he is prepared in advance to interpret an utterance of the speaker' while for the speaker, the prior theory 'is what he believes the interpreter's prior theory to be'. There is also a passing theory. For the interpreter, the passing theory expresses how the interpreter actually interprets the utterances of the speaker, while for the speaker the passing theory is 'the theory he intends the interpreter to use'. Even though most of the times interpreter and speaker are not actually sharing prior theories, it is necessary for them to share passing theories if they want to communicate successfully, since the passing theory is the one used by the interpreter to interpret the speaker's utterances and at the same time the one that the speaker intends the interpreter to use.

Prior and passing theories for experimenter and subject

1. *Prior theory for the subject:* the theory with which the subject is equipped in advance in order to interpret a particular expression of the selection task. This theory is adjusted to the evidence such as character, gender, role etc. of the experimenter. This theory might not be sufficient for yielding the desired (by the experimenter) interpretation. In the case of the selection task this means that the prior theory of the subject might not be sufficient for the subject to assign the classical logical form to the 'If [...] then' clause of the task. (This might be the case, for example, when the task is described in natural language which often allows for multiple interpretations of certain phrases and expressions. Thus, it is not surprising if the subject assigns a different meaning to the task from the one intended by the experimenter.)

2. *Prior theory for the experimenter:* The theory which the experimenter expects the subject of the selection task to be equipped with before any conversation begins. This theory of course may be different of the actual prior theory of the subject.

   In fact, van Lambalgen and Stenning claim that Wason's prior theory is far from the prior theories of most of the subjects of the task. The prior theory of the experimenter should be more open to include many possibilities for different prior theories for each subject, which can all yield rational interpretations of the task, though possibly ones deviating from those intended by the experimenter.

3. *Passing theory for the subject:* The theory with which the subject actually interprets the experimenter. In other words it is the prior theory of the subject updated with any new information which might help for the correct, according to the subject, interpretation of the selection task. In the interviews between experimenter and subject of the selection task the building process of the subject's passing theory is obvious; the subject updates his prior theory during the conversation with the experimenter and is constantly adjusting it to the new evidence taken from the conversation.

4. *Passing theory for the experimenter:* The theory which the experimenter intends the subject to use in order to interpret the task, that is (for Wason) classical logic.

Applications of the two accounts

The accounts of Davidson, and of van Lambalgen and Stenning, can be applied on specific examples of discourses encountered in reasoning experiments or in educational contexts. Van Lambalgen and Stenning offer an illustration of how their account is applied on subjects of the Wason task in [SvL08]. In appendix A, I offer a brief illustration of two discourse examples interpreted according to a combination of the two accounts discussed here.

Chapter 5

Conclusion

We started out our investigation by posing the following questions:

(a) What are the common elements of, and differences between, the two accounts?

(b) What is the innovation that van Lambalgen and Stenning's account offers?

(c) How can van Lambalgen and Stenning's work escape the threat of relativism (of which it is accused)? How can their theory be improved in that respect?

(d) What kind of criticism can be brought against a Davidsonian account from the van Lambalgen and Stenning perspective and vice versa?

(e) Would combining the two approaches result in a more 'complete' explanation of human reasoning and rationality?

(f) How can the two approaches be used in interpreting specific instances of discourse?

In this chapter I will look at how the above questions were answered. The following section gives a summary of the answers that came out of this thesis. In section 5.2 some further plans will be discussed.

5.1 Suggested answers

Similarities and differences between the two accounts

The similarities and differences between the two accounts were discussed in chapter 4. Let us summarize first the similarities. As we saw, although the two accounts seem to be radically different at first glance, a more careful examination shows that they have much in common. At first glance Davidson (especially in his early work) adopts a strictly global view of rationality while van Lambalgen and Stenning adopt a strictly local one. However, both accounts are to an extent global, since they assume rationality to be a presupposition of communication and thus a global characteristic shared by all humans. This idea is expressed in Davidson's work by the Principle of Charity, a principle which is implicitly used also in the van Lambalgen and Stenning account. Moreover, we saw that both accounts allow enough room for variation in the specification of the norms of rationality, which means that they both have room for a local element in the characterization of rationality and the explanation of human reasoning. In later Davidson (post 1984), this is expressed with the prior and passing theories of speaker and hearer, while van Lambalgen and Stenning express the local element by introducing a multiplicity of logics available for describing the process of reasoning *to* an interpretation.

Another similarity between the two accounts is that the non-monotonic logic which van Lambalgen and Stenning suggest as best describing logical reasoning and linguistic communication seems to match in certain respects with Davidson's account of linguistic communication. Davidson's explanation, in his later work, of how successful linguistic communication is achieved has the property of graceful degradation (which is also discussed by van Lambalgen and Stenning in order to describe optimal performance during logical reasoning): the more the passing theories of speaker and hearer approach each other the higher the level of successful communication; if the passing theories do not coincide, we still have communication although we might not have agreement between speaker and hearer.

The main difference between the two accounts, apart from the fact that they have slightly different aims, is the status they give to classical logic. In the Davidsonian account, classical logic plays a central role in the explanation of rationality and linguistic communication. Classical logic seems to be for him the main normative system of rationality. On the other hand, van Lambalgen and Stenning introduce a whole new range of logics which sit at the same level as classical logic when it comes to explaining logical reasoning. In particular, during the process of reasoning *to* an interpretation classical logic sits at the same level as other logics and any of those logics can be used to describe how a person assigns a logical form to a task. During the process of reasoning *from* an interpretation, conclusions are derived from the premises of the task. Here, the notion of validity is again, in the van Lambalgen and Stenning account, not necessarily that of classical logic, whereas in the Davidsonian account the notion of validity remains the classical logical one.

The innovation of the van Lambalgen and Stenning account

There are three main innovations offered by van Lambalgen and Stenning. They reconsider the universality of classical validity, they give to semantics an important role in the understanding of logical reasoning, and they describe the process of logical reasoning as having two equally significant parts: reasoning *to* and *from* an interpretation.

The first and most significant innovation, namely that the classical logical notion of validity is no longer considered to be universal, changes dramatically the until now established view of rationality. As we saw, it has been a common belief that classical logic is the normative competence model when it comes to logical reasoning. In this view, all reasoning tasks should be assigned the logical form determined by classical logic, since it is assumed that classical logic defines the norms of rationality. For van Lambalgen and Stenning this is a wrong assumption, and there is much more going on before a logical form is assigned to a task. Depending on the semantics one may choose, the logical form of a task may be different to the classical logical one, which means that classical logic is no longer thought to be universal.

This brings us to the second important innovation, namely the importance of interpretation. Disregarding the importance of interpretation in the previous centuries has lead to the conclusion that classical logic is the normative system of rationality and that any subject that fails to reason according to classical logic is irrational (we discussed Wason's example in this thesis). As we saw, Davidson also adopts a view of classical logic as universal, although he thinks that even if one fails to reason according to classical logic this does not imply that one is irrational. Van Lambalgen and Stenning go beyond that, and insist on the importance of interpretation. A task stated in natural language may have different interpretations which correspond to different logical forms. The process of determining the logical form of the task is called reasoning *to* an interpretation.

The division of the process of logical reasoning into two parts, reasoning *from* an interpretation and reasoning *to* an interpretation, is the third innovation, tightly connected with the previous two. After a logical form is assigned to a task by a subject, the regulative norms within the chosen logic will determine the normative solution for the task. Thus, the interpretation and the derivation processes are separated. The normative element is apparent during the process of reasoning *from* an interpretation.

How van Lambalgen and Stenning's work can escape the threat of relativism

We discussed how one may reply to the criticism about relativism in the work of van Lambalgen and Stenning in section 3.3. As we saw, van Lambalgen and Stenning describe logical reasoning as comprised by two processes: reasoning *to* and reasoning *from* an interpretation. In the case of reasoning *from* an interpretation, the reply to the criticism is quite straightforward. Once a logical form is selected the rules that will lead to the right, normative conclusions are determined by the logical form and any deviation from the norms would indicate a performance error in reasoning. However, in the case of reasoning *to* an interpretation, the reply to the criticism is not given in [SvL08]. Van Lambalgen and Stenning do not offer a clear and normative picture on how a logical form is assigned to a task or, more generally, to natural language expressions. That is, the threat of relativism still exists in the process of reasoning *to* an interpretation.[1]

A move that would possibly help van Lambalgen and Stenning escape the threat of relativism would be to explain further the process of reasoning *to* an interpretation and how the subjects come to choose a logical form for a task. If the choice is absolutely free, then the threat of relativism still exists. Using Davidson's account of rationality might be of help at this point. The Principle of Charity may be used as a standard which determines (because of the common characteristics we have as humans—e.g. our discriminatory abilities—, our mutual knowledge and beliefs, various contextual parameters etc.) a fixed range of options for selecting a logical form for a reasoning task.

Criticism of the Davidsonian account from the van Lambalgen and Stenning perspective and vice versa

From the Davidsonian perspective, a criticism that can be brought against the van Lambalgen and Stenning account is the issue of normativity in the process of reasoning *to* an interpretation. As indicated above, the process of reasoning *to* an interpretation is still vulnerable to criticism about relativism, since the choice of logical form is considered to be in some vague way *a priori* by van Lambalgen and Stenning. Davidson's account uses the Principle of Charity to tackle this problem, and this could be used by van Lambalgen and Stenning in order to improve their own account of logical reasoning.

From the van Lambalgen and Stenning perspective, the most significant criticism would be about the status of classical logic in the understanding of everyday logical reasoning within Davidson's account. For Davidson classical logic seems to be the normative competence system, an idea which is obviously abandoned in the van Lambalgen and Stenning account. Davidson's account could benefit from considering a multiplicity of logics for explaining rationality and logical reasoning, and from reconsidering the notion of classical validity.

Combining the two approaches to form a more 'complete' explanation of human reasoning and rationality

Having said the above, it seems that a combination of the two accounts would help in forming a more 'complete' explanation of human reasoning, rationality and interpretation. The Davidsonian account has much to offer in the understanding of what is rationality and why humans tend to give similar interpretations to similar words and expressions under exactly the same circumstances (Principle of Charity) and may give an explanation of how during reasoning *to* an interpretation subjects tend to give certain logical forms to reasoning tasks. The van Lambalgen and Stenning account has much to offer in the explanation of the variety of normative answers to a single expression of a reasoning task. The issue of interpretation is stressed, and the rationality behind reasoning is

[1] In [SvL08, p. 352], van Lambalgen and Stenning give a short explanation of why they are not postmodern relativists. In their explanation the process of reasoning *to* an interpretation is still left undefended.

revealed. The two accounts seem to complement one another, although at first glance they might seem conflicting.[2]

Using the two accounts in interpreting specific instances of discourse

In appendix A I examine briefly how the two accounts combined can be applied to two specific examples. The first example is an interview with a subject who has solved the Wason selection task, and the second example is a discourse from a science classroom in a high school. The two accounts can explain the misunderstandings that may occur during linguistic communication and how subjects' answers, although probably deviating from those expected by the person who built the task, can still be considered normative and thus rational.

5.2 Further plans and discussion

The two accounts combined may have interesting implications in the area of education. Within a classroom, linguistic interaction is the most common way that teachers and students communicate with each other, and reasoning tasks are very common. The richness in reasoning that is revealed in the Socratic dialogues of the kind examined by van Lambalgen and Stenning, reveals the importance of discussions within educational contexts. Lecturing should always be combined with discussions where the teacher has the opportunity to recognize the way students interpret the teacher's words or the tasks they have to solve, as well as the way students reason. An example of a linguistic interaction within a classroom and of what it may reveal is given in appendix A.2.

In the area of research in education, researchers are often confronted with interviews which they need to interpret in order to check students' understanding of certain tasks or their conceptual understanding. An example of such an interview is given in appendix A.1. An analysis similar to the one offered for the two examples of appendix A may reveal what is going on during students' reasoning, in order to avoid hasty conclusions such as 'students are irrational' or 'students have serious misconceptions about the world.' The way educational researchers interpret students' reasoning determines the way new educational methods are built. For this reason, investing in understanding how students interpret teachers and textbooks is worth the effort.

I am convinced, and I hope I have managed to convince you, that philosophy of language and logic may significantly help in understanding human reasoning and rationality. As a consequence, both philosophy of language and logic are necessary tools in the area of educational research at the center of which lies linguistic interaction between humans. Soon I will be commencing my research in the MSc in Mathematics and Science Education, in which I will attempt to apply the inspiring ideas of Davidson and the promising innovations of van Lambalgen and Stenning within the context of mathematics education.

[2] See also the discussion in chapter 4.

Appendix A

Applications of the two accounts

In what follows I will apply the accounts of Davidson, and van Lambalgen and Stenning to two examples: first, an interview between experimenter and subject on the solution of Wason's selection task, and second, an instance of (unsuccessful) linguistic communication between teacher and student in a secondary science classroom setting. I will use empirical data taken from the research of Counihan [Cou08], and from Klaassen and Lijnse [KL96]. The first example reveals the kind of reasoning that goes on in the mind of a subject when solving a reasoning task. I will illustrate how the two accounts discussed in this thesis can explain this reasoning procedure. The second example does not involve a reasoning task, but rather an instance of linguistic communication between teacher and student. I will use this example to illustrate how the ideas in both early and later Davidson can be used in order to explain the miscommunication which occurs, and to suggest ideas about how van Lambalgen and Stenning's account might be of help in interpreting similar instances of communication in the classroom.

A.1 Example 1 - Wason's selection task

The first example is taken from Counihan's research data. It is an interview (for the full transcript see appendix B) with a subject that has, before the interview, given a "wrong"[1] written response to the following version of Wason's selection task.[2]

> Below is depicted a set of four cards, of which you can see only the exposed face but not the hidden back. On each card, there is a letter on one of

[1] The subject's response was $A, 4$ whereas that expected by Wason is $A, 7$.

[2] Quoted from [Cou08, p. 127].

> its sides and a number on the other side. The letters are *A* and *K*; the numbers are 4 and 7.
>
> Also below, there is a rule which applies only to the four cards. Your task is to decide which (if any) of these four cards you must turn in order to decide if the rule is true. Don't turn unnecessary cards. Tick the cards you want to turn.
>
> $\boxed{A}$ $\boxed{K}$ $\boxed{4}$ $\boxed{7}$
>
> **Rule:** *If there is an A on one side of the card, then there is a 4 on the other side.*

Since the subject has given a wrong response to the written task, that is, a response which is not in accordance with the interpretation of the task rule as the material implication, Wason's conclusion would be that the subject is irrational. However, both the accounts discussed in this thesis would reject Wason's conclusion and would start by making the assumption that the subject *is* rational. This assumption follows from the Principle of Charity, and it is the first step in trying to understand the reasoning of the subject. Recall Davidson's claim that without assuming rationality interpretation is impossible.

The second step, immediately related to the Principle of Charity, would be to assume that the subject will most likely adopt a credulous attitude for interpreting the task. That is, the subject will try to interpret the task based on a model which will make the rule true. Any conclusions derived by the subject will be true only in this specific model in which the rule is true (as opposed to a skeptical attitude, in which one would try to derive conclusions true under all interpretations of the task). During the procedure of finding an appropriate model for interpreting the rule, the subject will use contextual information that would be provided by the task description, the experimenter etc. A kind of credulous attitude is closed world reasoning, in which interpretation and inferences are based on the positive knowledge we have at a specific moment about the world. The subject takes into consideration all (and only) the contextual information that is given by the task rule and the description of the task. As we will see, this subject indeed adopts a credulous attitude.

If one interviews the subjects after the written task, the reasoning behind the written response can be revealed. This is the purpose of the interviews conducted by Counihan, as well as by van Lambalgen and Stenning: to give subjects a voice in order to understand the logic behind their reasoning rather than concluding that the subjects are irrational.

In this case the interview starts like this:

> [chat about AK47 as cards are laid out]
>
> S: Surely the easiest way would be to turn *A* and 4, but 4 first.
>
> E: OK. 4 first - why?
>
> S: Because I can already see that there's an *A* on this side [pointing to the *A*], and I can already see that there's a 4 on this side [pointing to the

> 4 card] so if I turn over 4 there should be an A. Just like if I turn over A there should be a 4.

It is obvious here that the logical form which the subject gives to the 'If ... then' rule of the task does not coincide with its grammatical form, as one might have expected. Is this subject irrational? Well, no. The grammatical form of the rule does not imply that its logical form should be the material implication, although we usually express the logical formula $p \rightarrow q$ with the natural language phrase 'If p then q.' As van Lambalgen and Stenning claim, the indeterminacy of natural language meaning makes imposition of logical form necessary. There is no meaningful task to solve, until we have assigned a logical form to the task. In other words, the assignment of logical form to the task is constitutive of the meaning of the task. As the following passage from the interview clearly suggests, for some people 'If ... then' might mean something different than the conditional $p \rightarrow q$.

> E: Then let's think about what could be behind the 4. The 4 could have an A or a K behind it.
>
> S: Right.
>
> E: Let's suppose it has a K behind it.
>
> S: I'd be surprised.
>
> E: OK why?
>
> S: Just because, the way the statement is, if there's an A on one side there's a 4 on the other side, I, don't know about other people, but I just tend to assume that the reverse applies as well. That if A has 4 behind it then 4 should have A behind it. Well not should but is more likely that it would. [...]

It seems as if the subject is interpreting the rule as the material biconditional, but this is a false impression. If we assume that the rule is interpreted as the material biconditional, the normative answer (defined by the regulative norms within classical logic) would be to turn all the cards. Remember that the material biconditional is true when either both antecedent and consequent are true or when they are both false. So one would have to check all cards to see whether the rule is true or false. However, we know that the subject chose cards A and 4 in his written response. This fact, together with the justification of this selection by the subject during the interview and with the assumption that the subject is rational (Principle of Charity), suggest a certain prior theory of the subject before the interview.

Before describing the prior theory of the subject I want to define some notation for prior and passing theories. I will use $(X)^Y_Z$, where X characterizes the theory and can take the values Pr (prior) or Pa (passing), Y indicates the person to whom the theory belongs and can take the values E (experimenter) or S (subject), and Z is a natural number indicating the version of the theory (0 will be used to indicate the version of the prior theory that one has before the beginning of a conversation).

- $(Pr)_0^S$: The logical form of the sentence *If there is an A on one side of the card, then there is a 4 on the other side*, is a biconditional of the form $A \leftrightarrow 4$, but with the following addition: the phrase 'on one side of the card' means 'on the front/visible side of the card'. Moreover, for the $4 \rightarrow A$ direction, the rule is interpreted as 'If there is a 4 on the front/visible side of the card, then there is an A on the other side.'

That the subject has this prior theory is apparent in the interview. The subject clearly interprets the rule as a kind of a biconditional: the rule is true when both antecedent and consequent are true or when they are both false, but with the restriction of the expression 'on one side' meaning 'on the front/visible side.' We will later illustrate this with some passages from the interview. Let's look now at the prior theory for the experimenter.

- $(Pr)_0^E$: There is a large set of possibilities of theories that the subject might use in order to interpret the selection task rule.

In this case the experimenter, judging by the written response of the subject to the selection task, assumes that the subject must not have interpreted the rule as the material implication, since the selection of the cards was not the one that corresponds to that interpretation. Of course, it might be the case that the subject did interpret the rule as the material implication, but due to some performance (or other) error, he failed to give the right corresponding answer. Since the intention of this interview is to find out what is the kind of interpretation the subject gives to the task, the experimenter's prior theory is comprised by a large set of possibilities. This is a different attitude than the one that Wason would have. Wason expected that the subjects would give the logical form of the conditional (the material implication) to the rule, wrongly assuming that the grammatical form of the task is identical to the logical form of the conditional as defined in classical logic.

What about the passing theories of experimenter and subject? For the subject, the building process of the passing theory is obvious in the interview; the subject frequently updates his prior theory and adjusts it to the new evidence taken from the conversation with the experimenter. We may look at some suggestive passages from the interview:

> E: OK. And if there's not an A behind the 4?
>
> S: Then the rule is wrong. Actually no because it doesn't say that if there's 4 on one side then there's an A on the other side. Does it? Cause it says if there's an A on one side then there's a 4 on the other side. So strictly speaking I should turn over A shouldn't I? Now that I think about it.
>
> E: It true, it doesn't say that if there's a 4 there should be an A.
>
> [...]
>
> E: So just the A then?
>
> S: Yeah.

> E: So why would you turn the 4 then, in the other case?
>
> S: Out of pure curiosity. To see if it was a trick question, as in..... 4s won't necessarily have As on the other side. Do you see what I mean?

In this passage we see that the subject has abandoned the idea that the rule should be interpreted as a biconditional. 'It doesn't say that if there's a 4 on the one side then there's an A on the other side.' The subject seems to believe that, after all, it is only one direction of the biconditional that he probably should be interested in, namely the $A \rightarrow 4$. But we still see that, at least out of curiosity, he would like to turn the 4 card to see if there is an A behind it, although if there wasn't this would not falsify the rule. So at this moment we have the following passing theory for the subject.

- $(Pa)^S_1$: The logical form assigned to the task rule is the following. Whenever there is an A on the front/visible side of the card, there is a 4 on the back side of it.

Note here that, as mentioned earlier, the subject seems to think that the only cards that are relevant for deciding whether the rule is true, are those which have an A on the front/visible side, which is also obvious in the following passages:

> E: And if you had to say which ones you're going to turn over beforehand, before you turn any of them, you'd say A and 4 - is that right?
>
> S: Yeah. These [indicating K and 7] are the unnecessary ones, these don't relate to the [gestures to rule]. Well I mean it does say A, K, 4 and 7 but it doesn't talk about turning over K or 7.
>
> [. . .]
>
> E: OK so it [the K] could have, could only have a 4 or a 7. . . Let's say it had a 4 on the back, would that mean anything for the rule?
>
> S: If it had a 4 on the back? (ja) It wouldn't mean that anything was wrong, because the statement there in bold [the rule] is talking about the card with an A on it, not the card with a 4 on it. Yeah?

Since the cards K and 7 obviously do not have any As or 4s on their visible side, the subject thinks there is no reason to check them. For him, the only relevant cards are the ones that have on the visible side either As or 4s.

Another interesting point is that the subject expects the rule to be true before checking whether this is indeed the case. In many places in the interview, we see that the subject seems to give a deontic interpretation to the rule, that is, interprets the rule as: 'If there is an A on the front/visible side of the card then there *should* be a 4 on the back of it' and at the same time 'If there is a 4 on the front/visible side of the cards then there *should* be an A on the back of it'. He often expresses that he would be surprised if something else other than a 4 appears behind the A or if something other than an A appears behind the 4:

> E: So why would you turn the 4 then, in the other case?
>
> S: Out of pure curiosity. To see if it was a trick question, as in . . . 4s won't necessarily have *A*s on the other side. Do you see what I mean?

In the above we see that if something different than an A is found behind the 4, then the subject would assume that the task was a trick question.

> E: So let's think about what could be on the back of the A. On the back of the A there could be a 4 or a 7.
>
> S: There should be a 4.
>
> E: There should be a 4?
>
> S: Yes, because it says if there's an A on one side then there is a 4 on the other side.
>
> E: Ja but you have to check whether that's true or not.
>
> S: Ah! OK fine. If it is true that's what it should be. But yeah you're right, it could be a 4 or a 7.
>
> E: So say you turned it over and there's a 7. What would that mean?
>
> S: The rule is wrong. Well it's, it's a lie!
>
> E: Ja. So that's basically your task here - check whether or not this [indicating the rule] is a lie. [. . .]

Here, the subject clearly says that there *should* be a 4 behind the A. Otherwise, the rule would be a lie. It could be the case that the authority of the experimenter makes the subject think that the rule should be true, so if it turns out to be false it would be a real surprise. Only after the experimenter explains how the subject should interpret the rule, and what his task actually is, the subject changes his prior theory accordingly. However, later in the conversation, the subject again returns to his earlier prior theory, insisting that he would be surprised if a K was found behind the 4. We can see him straggling in his attempt to change his interpretation of the task and accept the experimenter's.

> E: Then let's think about what could be behind the 4. The 4 could have an A or a K behind it.
>
> S: Right.
>
> E: Let's suppose it has a K behind it.
>
> S: I'd be surprised.
>
> E: OK why?
>
> S: Just because, the way the statement is, if there's an A on one side there's a 4 on the other side, I, don't know about other people, but I just tend to assume that the reverse applies as well. That if A has 4 behind it then 4 should have A behind it. Well not should but is more likely that it would. But then, if this was a trick question then. . . I mean you have to think about it before you make a finite decision. I would expect. . . I know it's 50-50, but I would expect 4 to have an A behind it.

But what about the passing theory, at this stage, of the experimenter?

- $(Pa)^E_1$: The subject should interpret the 'If. . . then' expression as some kind of conditional (rather than a biconditional). The subject should not interpret the 'one side', 'other side' expression like 'visible side' and 'invisible side.' The subject should not expect the rule to be true without checking it first.

The experimenter, although holding a relatively neutral position throughout the whole interview, does at times have certain expectations from the subject. One is that the subject interprets the rule as not already true, a fact which supposedly is implied by the way the task is phrased. We see that the experimenter points out to the subject what the actual task is, namely to check whether the rule is true or false: 'So that's basically your task here — check whether or not this [indicating the rule] is a lie.'

Another expectation on the part of the experimenter is that the subject does not interpret the rule as a biconditional. Thus, when the subject realizes this for a moment ('it doesn't say that if there's 4 on one side then there's an A on the other side. Does it? Cause it says if there's an A on one side then there's a 4 on the other side. So strictly speaking I should turn over A shouldn't I? Now that I think about it') the experiment confirms this ('It true, it doesn't say that if there's a 4 there should be an A').

In the last part of the interview, the subject again changes passing theory. Given that the subject accepts the fact that it is not necessary to find an A behind the 4 or a 4 behind the A, as the experimenter indicates, the subject gives an answer as to whether the rule would be falsified if certain combinations of letters and numbers suggested by the experimenter were the case:

> E: OK. And what if you found a K behind that (the 4)?
>
> S: Behind the 4?
>
> E; Yes.
>
> S: Like I said, I'd be surprised. Because I'd expect an A to be there.
>
> E: But if you did find it, would it tell you anything about this rule?
>
> S: It wouldn't mean that the rule was false, no, because that rule is still talking about A.
>
> E: And if there was an A, would that tell you something about the rule?
>
> S: If there was an A behind the 4? Then the statement is true.
>
> E: OK and the 7? What could be behind the 7?
>
> S: A K or an A.
>
> E: OK and say there was a K behind the 7 what would that mean?
>
> S: Nothing. Because the rule has nothing to do with K and 7. At least, not yet. Right there, it doesn't say anything about K and 7, so
>
> E: And if there was an A on the other side of the 7?
>
> S: Then the rule is false.

The last two lines of the interview, seem to suggest that the subject realizes that, in case there is an A behind the 7, then this would falsify the rule. Does this mean that the subject's final passing theory coincides with that of the experimenter? Most probably not. The two passing theories of subject and experimenter approach each other but do not actually coincide. The final[3] passing theories would be something like this:

- $(Pa)_2^S$: Assuming that the 7 card is relevant, if there is an A behind the 7 then the rule is false. However, the rule does not imply in the way it is stated that the 7 card is relevant. (Plus the assumptions of $(Pa)_1^S$)
- $(Pa)_2^E$: Given that the 7 card is relevant, if there is an A behind the 7 then the rule is false. The rule does not speak only for the cards with an A or a 4 on them, so the 7 card is also relevant. (Plus the assumptions of $(Pa)_1^E$)

In fact, the rule as intended by Wason, and as the experimenter in this case understands it, does not give any hint that would suggest that the 7 card is irrelevant. The experimenter needs to make the subject accept this assumption, and given this assumption the subject gives the 'right' answer to the task. However, the subject still believes that checking the 7 is not relevant, since the rule only talks about As and 4s. That is, the subject still gives a different interpretation to the task than that of the experimenter, and does not really want to update this interpretation to the one suggested by the experimenter. So, successful communication is not fully achieved, since experimenter and subject do not assign the same meaning to the task. However, there seems to be some kind of common ground established. Experimenter and subject might not agree on the interpretation of the task, but it seems that they are much closer now to understanding exactly what meaning each of them assigns to the task rule.

A thing that remains to check is whether in the process of reasoning *from* an interpretation the subject is consistent with his own interpretation. As we discussed in the previous chapters, an inconsistency would suggest not that the subject is irrational, but rather that the subject has performed some kind of reasoning error. The subject chose A and 4 on the written task. According to $(Pr)_0^S$, this selection is justified. During the interview we see that at some point the subject abandons the idea that turning the 4 card would help in deciding whether the rule is true or false. Although the subject would still be surprised if he found something else than an A behind the 4, he thinks he misinterpreted the rule as being a kind of biconditional. According to this later interpretation, he decides temporarily that he would turn only the A card, which is again consistent with this later interpretation of the rule (made under $(Pa)_1^S$). Finally, answering that an A behind the 7 would falsify the rule is a consistent answer given the interpretation of the rule under $(Pa)_2^S$. So in all cases we

[3] There are possibly more passing theories that one can find within this interview with further analysis. Thus, I do not suggest by the numbering which follows that the second passing theory is necessarily the final one.

see that the subject reasons consistently from the chosen interpretation which shows how wrong Wason was in assuming that the subjects are irrational.

A.2 Example 2 - Discourse in a Science Classroom

Classroom interactions can offer examples of communication failure between teacher and students. Teachers often conclude that their students are irrational, or that they have some serious misconceptions which prevent them from learning and understanding how the world really works. In this thesis we saw that the way people solve reasoning tasks depends mainly on the interpretation they give to such tasks. It is thus not right to conclude that people have weird beliefs or are irrational when they give answers to tasks which deviate from the expected ones. Rather, we saw that a task does not really exist before one imposes a meaning on it. Words, sentences, and reasoning tasks do not have a fixed interpretation, so we should not take it for granted that students will assign the same meaning to them that teachers do.

Explaining reasoning and linguistic communication in the way Davidson and van Lambalgen & Stenning do, may have interesting and useful implications on the way teachers interpret the reasoning of their students and therefore on the linguistic interaction among students or between teacher and students. Klaassen and Lijnse [KL96] have already applied the ideas of Davidson's radical interpretation to an instance of classroom linguistic interaction, presented below, which has been discussed a lot in the literature. Let's look at the example [KL96, pp. 116–118], and then discuss the ways it has been commonly analyzed, the way Klaassen and Lijnse analyze it, and what this thesis adds to this analysis.

> In the previous lesson, the students watched a specially developed video about forces that act when cycling. The following transcript begins with the teacher, who intends to summarize and elaborate on the video by means of the well-known air track. His introductory question, in which he asks for the forces acting on the glider when it rests on the not-yet-operating track, is meant simply to remind the students of the supposedly well-known static forces that are acting in that situation. Then the following discussion occurred, which took about 20 min.
>
> *Teacher*: The video has been about forces that act when cycling. Well, here [points to the glider on the track] I have a kind of bicycle. Let me now first ask what forces are acting on it. Just try: What forces do you think are acting at this moment? Are there any forces acting?
>
> *Eric*: Gravity.
>
> *Teacher*: Gravity, Eric says. What if gravity were the only force, what would happen then?
>
> *Eric*: Then it would go down.

Teacher: Then it would go down. Ernie, what other forces could be acting?

Ernie: Eh... well...

Teacher: What prevents it from falling down?

Ernie: The track.

Teacher: Right, the track. So the track has to supply a counterforce to prevent the glider from falling down. Just for the sake of completeness: Eric, which direction has gravity?

?: [joking] Upwards

Eric: No, downwards.

Teacher: So, Orson, the force of the track is upwards. Right?

Jane: Hows that?

Orson: Well, otherwise it would fall down.

Teacher: Otherwise it would fall down, he says. So, if it did not rest on the track and I dropped it, then only gravity would act and it would fall down. If the track wants to stop it, then it will have to push the glider upward.

Jane: But the track does not push, does it?

Teacher: The track does not push.

Jane: No....

Orson: Well, the track is just there.

Jane: ...It's just there.

[Some students are mumbling things such as, "Dont make such a fuss. Just accept it."]

Teacher: If you drop it, it will fall down; a force will act upon it.

Jane: Sure, if the track is not there.

Teacher: Okay. If you put it on your fingers... I can't take it off. [The teacher cannot get the glider off the track, and takes a small weight instead.] It's the same with this thing [the weight], isn't it? If you drop it, it will fall down. Now I want to stop it [places the weight on the tips of his fingers]. Since it is such a small weight, you don't feel much. But if you put a heavy weight on your fingers, you will feel it.

Jane: Okay.

Teacher: That is because you will have to exert a counterpressure. So you do have to...

Jane: Sure, if you're doing that yourself.

Teacher: If I place a heavy weight here, then my fingers will go down. If I want to keep it in place, I will have to push it upward. The track will do that too, it's just that we don't notice that. We don't notice that the track does it, the track doesn't move....

Carl: Yes, but the track can't push upward, can it?

Teacher: . . . But the track in fact does it as well.

Carl: Yes, but the track can't do that, can it?

Teacher: Oh yes, it can do just that.

Carl: You can push upward with your fingers, but the track can't.

Teacher: Let me take something else, something more flexible than metal. [Fetches a piece of foam rubber and puts it in front of him.] Here goes. So I will now try to convince you that the track really exerts an upward force. That is, I did agree with Orson, Jane did not; lets see whether we can come to an agreement. [Puts the small weight on the foam rubber, which gets pushed in a bit.] If I put this thing here, the foam rubber gets pushed in, doesnt it? Well, actually I need something a bit heavier. . . .

Jane: Oh, well I do believe you as it is.

Teacher: Do you? So you do actually believe that. [Laughter.] So, the foam rubber will get pushed in if you put something heavy on it. And if we don't put something heavy on it, but push it in and let go [does so with a finger], what will happen then?

Jane: Then it will come up again.

Teacher: Then it will come up again? Why's that?

Jane: Well, because theres nothing on it.

Teacher: Sure, but what does it do then, when it comes up? Then it pushes upward, doesn't it?

Jane: What?

Teacher: [Somewhat more pressing.] Then it pushes upward, doesn't it?

Jane: No, then it just gets back to its original state.

[Some students seem to suggest that Jane is just being stubborn.]

Jane: No, I don't think that has got anything to do with it.

Teacher: Don't you? I push the foam rubber in, put something on it, and the foam rubber pushes it upward. Then that is an upward force.

Jane: Well, I think that's really very strange.

Teacher: Do you?

Jane: Yes. That is not. . . well. . . no, that is not a force. I don't think it is really a force.

Teacher: If you want to push something up, then for that purpose you will have to exert a force. And now [pushes the weight into the foam rubber and then lets the foam rubber spring back] it is pushed in and it pushes the weight back up.

Jane: Okay.

Teacher: But you don't think that's a force.

Jane: Right.

Teacher: You don't think that's a force. For it is the same, isn't it? And do you consider this to be a force, when it falls down?

> *Jane*: Sure, that's gravity.
>
> *Teacher*: So, the downward motion is due to a force, but if it moves up [lets the weight again move up from the foam rubber] then that is not due to a force?
>
> *Jane*: Right.
>
> [Laughter from the class. The teacher remains serious.]
>
> *Teacher*: What if I now... I throw it upward, like this.
>
> [Jane also begins to laugh about the awkwardness of the whole situation.]
>
> *Teacher*: Is that a force or not?
>
> *Jane*: [Laughing.] It is, of your hand it is.
>
> *Teacher*: Of my hand it is. And now I let the foam rubber do it [again does so] and then it is no longer a force.
>
> *Jane*: [Still laughing a bit.] Right.
>
> *Teacher*: What, then, is the difference?
>
> *Jane*: [Serious again.] Well, that motion just goes all by itself. That's just the way things go. [Laughter.] Well, I really do think that's strange.
>
> *Teacher*: So because it goes all by itself, that is why according to you it is no force. If it now of itself gives a slap, then that will be a force.
>
> *Jane*: Yes.
>
> *Teacher*: I see. Well, so it seems that we haven't been making much progress. I do think there will be a force if you push it in, and Jane still doesnt think that that is a force. I'll leave it at that for a while. For the time being, everybody may think about it as he wishes. I would like to know, however, what the others do think about it.
>
> [Of the others, most indicate that they agree with the teacher, while no one indicates agreement with Jane. Some students, among which are Orson and Carl, are in doubt.]
>
> *Teacher*: Alright. Let's leave it at that for now. Perhaps I will be able to convince you at a later time. According to me, the difference between the foam rubber and the metal is that it can't be noticed that well that the metal is springy. But also the metal has got some spring that allows it to push back. So the metal is harder and—but now I speak for myself—it gets pushed in, but it does spring back and thus exerts a counterforce. Okay. It is sort of funny, though, that we still dont agree.

A first analysis of Jane's behaviour in the above passage is that Jane has not studied carefully the laws of Newtonian physics and that she has not studied the lesson very well. This is the most common analysis on the part of the teachers [KL96, p. 119]. A second analysis is that of students' misconceptions [KL96, p. 119]. According to this, students often have the wrong impression about how the world works, and hold beliefs that contradict reality. This assumption takes for granted that students have many misconceptions, and it is the task of the teacher to bring these misconceptions to the fore and replace them with the right conceptions. Finally, a common analysis of the above passage is that of alternative conceptions [KL96, p. 120]. The basic difference with the misconceptions analysis is that what was then called 'misconceptions' is now called 'preconceptions' or 'alternative conceptions.' So it is not actually stressed that the students conceptions are bluntly wrong, on the contrary it is claimed that the conceptions of students are simply different from scientific ones and have been created in a reasonable and intuitive way in their everyday life experiences. An attitude that one might adopt about student's theories from a scientific point of view, given this analysis, is the following (in [KL96, p. 121], quoted from [CKA80]):

> Of course, these [the students'] theories are often incomplete, incoherent and misguided.

According to Klaassen and Lijnse, all the above analyses are weak in that they do not represent what is actually happening in the minds of students during linguistic communication between teacher and student, and in that such analyses completely disregard the teacher's role in the outcome of communication. The issue of language, Klaassen and Lijnse claim, is of great importance and something that the other analyses disregard [KL96, p. 121].

The alternative analysis proposed by Klaassen and Lijnse focuses on the language used by teacher and students, and specifically on the—often different—meanings that teacher and students give to the same words. In doing that, the authors borrow some of Davidson's ideas on linguistic communication, and in particular his theory of Radical Interpretation and the, central to Davidson's work, Principle of Charity. According to this analysis it is assumed that teacher and students share general knowledge and beliefs about the world. Actually without this assumption we cannot interpret the words of one another (recall the discussion earlier in this thesis about the Principle of Charity as a presupposition of interpretation). So Jane and the teacher, in our case, do not actually have different beliefs about how an object will behave if certain forces are applied to it. There is not much difference in opinion between the teacher and Jane. They simply assign different meanings to the expression 'to exert a force' [KL96, p. 125] without realizing it, and this is why they fail to communicate.

The main conclusion in the article of Klaassen and Lijnse [KL96, p. 131] is that

> [. . .] all interpretation depends on our ability to find *common ground*. Finding the common ground is not subsequent to understanding, but a condition of it. Everything rests on sharing, and knowing one shares, a world, many reactions to its major features, and a way of thinking with someone else.

We see that interpretation has a central role in the analysis offered by Klaassen and Lijnse, and that building a common ground (i.e. sharing the meanings given to the same words by both speaker and hearer) is important in achieving successful communication. I align myself with Klaassen and Lijnse and I will show in what follows how the combination of the two accounts discussed in this thesis supports their analysis.

I will analyze the above dialogue in a way similar to the analysis of the interview in Example 1 of the previous section. The first step will be to assume, by using the principle of Charity, that Jane (and all the other students) are in general rational. Therefore, before concluding that any of the students is behaving irrationally, we should make sure we understand the way the students interpret the terms used by the teacher. In our example, what seems to cause the miscommunication between Jane and the teacher is the term 'force', as well as the phrase 'to exert a force.'

The second step is to assume that the students will most probably adopt a credulous attitude for interpreting the task posed by the teacher, which is to find out what forces are exerted on an object and who exerts these forces. During the procedure of finding an appropriate model for interpreting the expression 'exerts a force', students will use contextual information that would be provided by the teacher's description of the situation. As we saw earlier in this thesis, a credulous attitude is best described by closed world reasoning, in which interpretation and inferences are based on the knowledge we have at a specific moment about the world. The subject takes into consideration all (and only) the contextual information that is given by the task in question and the description of the task provided by the teacher.

It is worth mentioning at this point that, if the students were not discussing with the teacher the forces that act on the glider but they were merely asked to give an answer in a written form after the teacher having lectured about static forces, the answer that Jane would give would probably deviate from the one expected by the teacher. By initiating a discussion, it is much easier to understand what the students' interpretation is of a task in question, and what hides behind their answers. In this case, the teacher did not manage to understand Jane's interpretation, but surely the discussion plays a significant role which can be compared to the role of the interview with the Wason's selection task subjects. Interviews and discussions give subjects a voice which helps us understand the logic behind their reasoning rather than concluding that the subjects are irrational.

To continue our analysis of the dialogue, we should look at the Jane's process of reasoning *to* an interpretation. In terms of Davidson's prior and passing theories, it would be helpful to look at what the teacher's (T) and Jane's (J) prior theories are before the beginning of the dialogue.

- $(Pr)^J_0$: 'force' is something that causes an object to move. A force can be applied to an object by another object when they come in contact, but also at a distance (e.g. gravity). In the case of a contact force we say that the first object exerts a force on the second object only when movement is caused. If no movement is caused (as in the case of the track preventing the glider from falling down) when two objects are in contact, then we cannot say that a force is exerted from one to the other object. When an object causes itself to move (like the foam rubber returning to its initial position), then we cannot say that a force is exerted. 'Pushing' does not mean 'exerting a force.' A soulless object cannot push on its own. Only an agent can push.

- $(Pr)_0^T$: 'force' is a term used by physics, and it should be used in the classroom in that very sense defined by physics.[4]

The teacher's prior theory includes an ordinary definition of 'force' given by the science of physics, and a similar kind of interpretation for the expression 'to exert a force.' Usually teachers assume that students use such terms and expressions in the classroom with the same meaning the teachers do. But this is most often not true. Surely one of the aims of science teaching is to show to students the established meanings of such terms by scientists, but until something like this is achieved, teachers should not assume that students share the same meanings with teachers.

It is obvious through the dialogue that Jane is thinking of exerting a force as a *cause-effect* relation between *two* distinct objects. The effect of this relation should be a movement of one of the two objects. This is obvious in the dialogue:

> *Teacher*: [. . .] So, the foam rubber will get pushed in if you put something heavy on it. And if we don't put something heavy on it, but push it in and let go [does so with a finger], what will happen then?
>
> *Jane*: Then it will come up again. [. . .]
>
> *Jane*: [. . .] **it just gets back to its original state.** [. . .]
>
> *Teacher*: [. . .] I push the foam rubber in, put something on it, and the foam rubber pushes it upward. Then that is an upward force. [. . .]
>
> *Jane*: [..] That is not. . . well. . . no, that is not a force. **I don't think it is really a force.**
>
> *Teacher*: If you want to push something up, then for that purpose you will have to exert a force. And now [pushes the weight into the foam rubber and then lets the foam rubber spring back] it is pushed in and it pushes the weight back up.
>
> *Jane*: Okay.
>
> *Teacher*: But you don't think that's a force.
>
> *Jane*: Right.
>
> *Teacher*: You don't think that's a force. For it is the same, isn't it? And do you consider this to be a force, when it falls down?
>
> *Jane*: Sure, that's gravity.
>
> *Teacher*: So, the downward motion is due to a force, but if it moves up [lets the weight again move up from the foam rubber] then that is not due to a force?
>
> *Jane*: Right.

[4] It is easy to find a definition of force in a physics textbook.

Teacher: **What if I now. . . I throw it upward, like this.**

Teacher: **Is that a force or not?**

Jane: [Laughing.] **It is, of your hand it is.**

Teacher: **Of my hand it is. And now I let the foam rubber do it [again does so] and then it is no longer a force.**

Jane: [Still laughing a bit.] **Right.**

Teacher: **What, then, is the difference?**

Jane: [Serious again.] **Well, that motion just goes all by itself. That's just the way things go.** [Laughter.] Well, I really do think that's strange.

Teacher: **So because it goes all by itself, that is why according to you it is no force. If it now of itself gives a slap, then that will be a force.**

Jane: **Yes.** [. . .]

It is easy to notice that the passing theories of both teacher and Jane after the end of the dialogue remain the same as their prior theories. They are both trying to convince each other but no one succeeds. Although the teacher seems to get closer to understanding that the term 'force' means something different for Jane, at the end he remains puzzled with the situation, and assumes that Jane uses the term 'force' in the same sense he does, and this is why he thinks that Jane has a strange way of thinking about the world. Jane does not update her prior theory by any of the teacher's examples. For her 'force' has a very different meaning than it has for the teacher, so she also seems to feel puzzled and thinks it is very strange that the teacher perceives the world in such a different way from her own. So the passing theories of Jane and the teacher are not close enough to each other, which explains why there is no successful communication at the end.

Let us now look at the process of reasoning *from* an interpretation. Is Jane right in making the conclusions she makes, given the interpretation of the word 'force' by her? Yes, she is. Given her interpretation of the term 'force' and the expression 'exert a force', it does not make sense to say that the track exerts a force on the glider and thus it prevents it from falling. The track is in contact with the glider, yet the glider does not move. When no movement is caused, according to Jane's interpretation, there is no contact force exerted. What about the foam rubber? The foam rubber moves, but there is no other distinct object that causes this movement. So, there is no force acting upon the foam rubber.

The teacher is also correct in his process of reasoning *from* an interpretation, given the meaning he assigns to 'force' and 'exert a force.' So, who is right? The teacher or Jane? Both. They are simply talking about two completely different things by using the same words, and this causes them to fail in communicating successfully.

Appendix B

Interview Transcript

The interview was conducted by Marian Counihan (an analysis of parts of this interview is offered in [Cou08]).

Interview transcript

'S' = subject, 'E' = experimenter

[chat about $AK47$ as cards are laid out]

S: Surely the easiest way would be to turn A and 4, but 4 first.

E: OK. 4 first - why?

S: Because I can already see that there's an A on this side [pointing to the A], and I can already see that there's a 4 on this side [pointing to the 4 card] so if I turn over 4 there should be an A. Just like if I turn over A there should be a 4.

E: OK. And if there's not an A behind the 4?

S: Then the rule is wrong. Actually no because it doesn't say that if there's 4 on one side then there's an A on the other side. Does it? Cause it says if there's an A on one side then there's a 4 on the other side. So strictly speaking I should turn over A shouldn't I? Now that I think about it.

E: It true, it doesn't say that if there's a 4 there should be an A.

S: It's silly though, cause on the written one I ticked A and 4.

E: And do you think that is just because you didn't sort of, think it through?

S: No, I did think it through, but then I thought... if you don't read it into it too much then A has 4 on one side 4 has A on the other side. So I would actually turn both of them over, but probably this one [the 4] first. So shall I go ahead and turn it over?

E: Not yet, I want to think about all the ... options. OK so the A - you also want to turn it over?

S: I would turn this one [4] over first, just

E: OK and if you had to decide before you turned any of them, what you had to turn over, so that you couldn't check what was behind them before you turned anything else over, what would you choose then?

S: *A*. Because that's what the question says. If there's an *A* on one side, then there should be a 4 on the other side.

E: So just the *A* then?

S: Yeah.

E: So why would you turn the 4 then, in the other case?

S: Out of pure curiosity. To see if it was a trick question, as in . . . 4s won't necessarily have *A*s on the other side. Do you see what I mean?

E: Yup. So if there was say a *K* on the other side of the 4?

S: I wouldn't think that was wrong, because it doesn't say here if there's a 4, cause I mean the thing does relate to the front of the card, at least it did on the thing [pointing at the written work] it said you can only see the front of the card.

E: Ja.

S: Right. So that's the. . . that would be the basis upon which I decided which card to turn over. It would be . . . I would choose *A* first because that's what it says. But then. . . if I could turn over two at the same time I'd do *A* and 4 at the same time. But that's not allowed.

E: No that would be allowed.

S: Oh really? Then I'd turn them. . . . No I'd still do *A* first.

E: And if you had to say which ones you're going to turn over beforehand, before you turn any of them, you'd say *A* and 4 - is that right?

S: Yeah. These [indicating *K* and 7] are the unnecessary ones, these don't relate to the [gestures to rule]. Well I mean it does say *A*, *K*, 4 and 7 but it doesn't talk about turning over *K* or 7.

E: So let's think about what could be on the back of the *A*. On the back of the *A* there could be a 4 or a 7.

S: There should be a 4.

E: There should be a 4?

S: Yes, because it says if there's an *A* on one side then there is a 4 on the other side.

E: Ja but you have to check whether that's true or not.

S: Ah! OK fine. If it is true that's what it should be. But yeah you're right, it could be a 4 or a 7.

E: So say you turned it over and there's a 7. What would that mean?

S: The rule is wrong. Well it's, it's a lie!

E: Ja. So that's basically your task here - check whether or not this [indicating the rule] is a lie. So the *K* could have a 4 or a 7- I am going to follow this procedure all the way through, that we go through the cards, just to get explicit on the reasons for turning and not turning. Then we'll actually turn them.

S: Fair enough.

E: OK so the *K* could have a 4 or a 7.

S: Yeah. It could even have an *A* on the back of it. But I don't know that because you haven't given me enough information.

E: Well there is a background assumption that each card has a letter and a number.

S: OK fine so forget what I just said.

E: OK so it [the *K*] could have, could only have a 4 or a 7... Let's say it had a 4 on the back, would that mean anything for the rule?

S: If it had a 4 on the back? (ja) It wouldn't mean that anything was wrong, because the statement there in bold [the rule] is talking about the card with an *A* on it, not the card with a 4 on it. Yeah?

E: Ja.

S: So *K* could have a 4 behind it, just like 7 could have an *A* behind it, you know?

E: So if you turned over 7 it also wouldn't matter if it had an *A* behind it?

S: Nope.

E: In the same way it wouldn't matter if the *K* had a 4 behind it?

S: Nope.

E: OK. And if the *K* had a 7 behind it? That still wouldn't matter?

S: No.

E: OK.

S: I'm still, I'm relating everything to this statement, right? (ja) and they're talking about *A*. So yeah.

E: Then let's think about what could be behind the 4. The 4 could have an *A* or a *K* behind it.

S: Right.

E: Let's suppose it has a K behind it.

S: I'd be surprised.

E: OK why?

S: Just because, the way the statement is, if there's an *A* on one side there's a 4 on the other side, I, don't know about other people, but I just tend to assume that the reverse applies as well. That if *A* has 4 behind it then 4 should have *A* behind it. Well not should but is more likely that it would. But then, if this was a trick question then... I mean you have to think about it before you make a finite decision. I would expect... I know it's 50-50, but I would expect 4 to have an *A* behind it.

E: OK. And if you haven't turned any cards, and you don't know whether this statement is true or not, would you still expect the 4 to have an *A* behind it?... Say you haven't found anything yet, and you have to check whether or not this is true [pointing at the rule] would you have any expectations about what's behind the 4?

S: Based on that statement?

E: Uh, ja.

S: Yeah I'd still think there was an *A* behind it (the 4).

E: OK. And what if you found a *K* behind that (the 4)?

S: Behind the 4?

E; Yes.

S: Like I said, I'd be surprised. Because I'd expect an *A* to be there.

E: But if you did find it, would it tell you anything about this rule?

S: It wouldn't mean that the rule was false, no, because that rule is still talking about A.

E: And if there was an A, would that tell you something about the rule?

S: If there was an A behind the 4? Then the statement is true.

E: OK and the 7? What could be behind the 7?

S: A K or an A.

E: OK and say there was a K behind the 7 what would that mean?

S: Nothing. Because the rule has nothing to do with K and 7. At least, not yet. Right there, it doesn't say anything about K and 7, so

E: And if there was an A on the other side of the 7?

S: Then the rule is false.

Bibliography

[Ach07] Theodora Achourioti. Logic, normativity and the a priori. Master's thesis, MSc in Logic, University of Amsterdam, 2007.

[AD97] Angeliki Athanasiadou and Rene Dirven. *On conditionals again.* John Benjamins, 1997.

[CKA80] A. B. Champagne, L. E. Klopfer, and J. H. Anderson. Factors influencing the learning of classical mechanics. *American Journal of Physics*, 48:1074–1079, 1980.

[Cou08] Marian Counihan. *Looking for logic in all the wrong places: an investigation of language, literacy and logic in reasoning.* PhD thesis, Institute for Logic, Language and Computation, University of Amsterdam, 2008.

[Dav67] Donald Davidson. Truth and meaning. *Synthese*, 17:304–323, 1967. Reprinted in [Dav01b, pp. 17–36]; page numbers refer to this edition.

[Dav70] Donald Davidson. Mental events. In Lawrence Foster and J. W. Swanson, editors, *Experience and Theory*. The University of Massachusetts Press and Duckworth, 1970. Reprinted in [Dav01a, pp. 207–227]; page numbers refer to this edition.

[Dav73] Donald Davidson. Radical interpretation. *Dialectica*, 27:314–328, 1973. Reprinted in [Dav01b, pp. 125–139]; page numbers refer to this edition.

[Dav74] Donald Davidson. Belief and the basis of meaning. *Synthese*, 27:309–323, 1974. Reprinted in [Dav01b, pp. 141–154]; page numbers refer to this edition.

[Dav75] Donald Davidson. Thought and talk. In S. Guttenplan, editor, *Mind and Language*. Oxford University Press, 1975. Reprinted in [Dav01b, pp. 155–170]; page numbers refer to this edition.

[Dav76] Donald Davidson. Hempel on explaining action. *Erkenntnis*, 10:239–253, 1976. Reprinted in [Dav01a, pp. 261–275]; page numbers refer to this edition.

[Dav80] Donald Davidson. Toward a unified theory of thought, meaning, and action. *Grazer Philosophische Studien*, 11:1–12, 1980. Reprinted as 'A Unified Theory of Thought, Meaning, and Action' in [Dav04, pp. 151–166]; page numbers refer to this edition.

[Dav82] Donald Davidson. Paradoxes of irrationality. In R. Wollheim and J. Hopkins, editors, *Philosophical Essays on Freud*. Cambridge University Press, 1982. Reprinted in [Dav04, pp. 169–188]; page numbers refer to this edition.

[Dav85] Donald Davidson. Incoherence and irrationality. *Dialectica*, 39:345–354, 1985. Reprinted in [Dav04, pp. 189–198]; page numbers refer to this edition.

[Dav86] Donald Davidson. A nice derangement of epitaphs. In Ernest LePore, editor, *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, pages 433–446. Blackwell, 1986.

[Dav87] Donald Davidson. Problems in the explanation of action. In *Metaphysics and Morality: Essays in Honour of J. J. C. Smart*, pages 35–49. Oxford: Blackwell, 1987. Reprinted in [Dav04, pp. 101–116]; page numbers refer to this edition.

[Dav94] Donald Davidson. Radical interpretation interpreted. *Philosophical Perspectives*, 8:121–128, 1994.

[Dav95] Donald Davidson. Could there be a science of rationality? *International Journal of Philosophical Studies*, 3:1–16, 1995. Reprinted in [Dav04, pp. 117–134]; page numbers refer to this edition.

[Dav01a] Donald Davidson. *Essays on Actions and Events*. Oxford University Press, 2001.

[Dav01b] Donald Davidson. *Inquiries into Truth and Interpretation*. Oxford University Press, 2001.

[Dav01c] Donald Davidson. What thought requires. In *The Foundations of Cognitive Science*. Oxford University Press, 2001. Reprinted in [Dav04, pp. 135–150]; page numbers refer to this edition.

[Dav04] Donald Davidson. *Problems of Rationality*. Oxford University Press, 2004.

[EO96] Jonathan St. B. T. Evans and David E. Over. *Rationality and Reasoning*. Psychology Press, 1996.

[Jef83] Richard C. Jeffrey. *The Logic of Decision*. The University of Chicago Press, second edition, 1983.

[Kan98] Immanuel Kant. *Critique of pure reason (The Cambridge edition of the works of Immanuel Kant)*. Cambridge University Press, 1998.

[KL96] C. W. J. M Klaassen and P. L. Lijnse. Interpreting students' and teachers' discourse in science classes: An underestimated problem? *Journal of Research in Science Teaching*, 33:115–134, 1996.

[Lew74] David Lewis. Radical interpretation. *Synthese*, 23:331–344, 1974.

[LL05] Ernie Lepore and Kirk Ludwig. *Donald Davidson: Meaning, Truth, Language, and Reality.* Oxford University Press, 2005.

[Noz93] Robert Nozick. *The Nature of Rationality.* Princeton University Press, 1993.

[Pia53] Jean Piaget. *Logic and Psychology.* Manchester University Press, 1953.

[Pia08] Jean Piaget. Intellectual evolution from adolescence to adulthood. *Human Development*, 51:40–47, 2008.

[Ram89] Bjorn Ramberg. *Donald Davidson's Philosophy of Language.* Blackwell, Oxford, 1989.

[Raw55] John Rawls. Two concepts of rules. *The Philosophical Review*, 64:3–32, 1955.

[RTtMF82] Judy S. Reilly, Elizabeth C. Traugott, Alice ter Meulen, and Charles A. Ferguson. *On conditionals.* Cambridge University Press, 1982.

[Sea70] John R. Searle. *Speech Acts: An essay in the philosophy of language.* Cambridge University Press, 1970.

[Ste96] Edward Stein. *Without good reason: The rationality debate in philosophy and cognitive science.* Clarendon Press, Oxford, 1996.

[SvL04] Keith Stenning and Michiel van Lambalgen. The natural history of hypotheses about the selection task: towards a philosophy of science for investigating human reasoning. In Ken Manktelow and Man Chung, editors, *Psychology of reasoning: historical and theoretical perspectives.* Psychology Press, 2004.

[SvL05] Keith Stenning and Michiel van Lambalgen. Semantic interpretation as reasoning in nonmonotonic logic: the real meaning of the suppression task. *Cognitive Science*, 29:919–960, 2005.

[SvL08] Keith Stenning and Michiel van Lambalgen. *Human Reasoning and Cognitive Science.* Cambridge, MA: MIT Press, 2008.

[Was66] Peter Wason. Reasoning. In Brian M. Foss, editor, *New Horizons in Psychology.* Penguin, 1966.

[Was68] Peter Wason. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3):273–281, 1968.

[Weted] Linda Wetzel. Types and tokens. *The Stanford Encyclopedia of Philosophy*, Winter 2007, Edward N. Zalta (ed.).