Studies in Belief and Belief Attribution

by

Miguel Hernando

Licenciado en Filosofía y Letras, sección Filosofía
Universitat Autònoma de Barcelona, 1993

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2001

Author .....
Department of Linguistics and Philosophy
October 11, 2000

Certified by .....
Robert Stalnaker
Laurance Rockefeller Professor of Philosophy
Thesis Supervisor

Accepted by .....
Vann McGee
Chairman, Department Committee on Graduate Studies

Studies in Belief and Belief Attribution

by

Miguel Hernando

Submitted to the Department of Linguistics and Philosophy
on October 11, 2000, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Philosophy

Abstract

My dissertation is about Frege's classic problem of the morning and the evening star. I distinguish two aspects of the problem. One aspect I call it *psychological*, and it consists in describing the content of the beliefs of people who are willing to assent to pairs like (1) 'Hesperus is nice' and (2) 'Phosphorus is not nice.' I assume an *interpretivist* account of belief content, according to which an agent has the beliefs that best explain her behavior, and I propose certain *principles of interpretation* to substantiate this view. I use this account to argue that the person who assents to (1) and (2) is not incoherent, but simply mistaken about the proposition expressed by those sentences. In my view, the subject who assents to (1) and (2) takes them to express propositions about different planets, but at least one of those planets cannot be a real planet. I propose that it is a fictional one, and appeal to Kendall Walton's account of prop-oriented make-believe to explain how to use propositions that are about fictional entities to describe the belief state of people who are confused about some identity.

The other aspect of the problem I call it *semantical*, and it consists in explaining how pairs of attributions like 'Charles believes that Hesperus is nice' and 'Charles does not believe that Phosphorus is nice' can be true at the same time. I offer a semantics based on the idea that, when we describe the belief state of people who are confused about some identity, we have to put ourselves in their shoes. We put ourselves in someone else's shoes by modifying our belief state to resemble the belief state of the other person; when we change our beliefs in this way, we acquire the beliefs necessary to talk of a single object as if it were two different ones. I argue that this *Simulation Semantics* can offer a satisfactory treatment of certain examples of belief attribution that cannot be handled by contemporary theories (examples in which the subject of the attribution is both confused about an identity, and is not familiar with the words that we use to attribute a belief to her). I also argue that this semantics has interesting applications to other problems in the philosophy of language, like for example the problem of the informativeness of identity statements.

Thesis Supervisor: Robert Stalnaker
Title: Laurance Rockefeller Professor of Philosophy

Acknowledgments

I am very grateful to the three members of my thesis committee. Alex Byrne was on my committee from the beginning, and offered me helpful advice and encouragement on the numerous drafts that I showed him. I knew I could always count on Alex to provide objections that would help me sharpen my theories, and indeed many parts in this dissertation sprung to life from conversations that I had with him. What is more, Alex always provided me with constant encouragement, even though I knew that his own views on these issues are quite different from mine.

Steve Yablo also gave me numerous comments about the dissertation. I am particularly grateful to him for helping me understand Kendall Walton's work on make-believe, and for making me acquainted with the singular 'they,' which is the topic of chapter 8.

To Bob Stalnaker I am indebted at several levels. As the chair of my thesis committee, he helped me appreciate the difficulties of the issues I was writing about. He was a constant source of challenges, as well as encouragement. I also thought that Bob was always very generous with his time and his ideas; I could always come to him with my worries and, sometimes, frustrations; he would always show me a way to overcome them. I will miss his guidance when I leave. At the same time, I am also indebted to Bob for his writings on belief and belief attribution. I think they are the most insightful treatment of the subject ever written, and this dissertation would not have been possible without them.

Besides the people in my committee, there were other people that also helped me to write this dissertation. Judy Thomson, as the placement officer of the department, read some parts of this dissertation that I used as writing samples. I always thought that Judy's comments both about the content and the form of my papers were incredibly helpful, and I think she will always be on my mind when I write a philosophy paper. Sylvain Bromberger was always ready to listen to my worries, and to provide some fresh ones that I wish I knew how to answer. Michael Glanzberg gave me the opportunity to present Bob's paper "Assertion" in his Fall 1999 seminar, which later would become chapter 6 of this dissertation. As I see things now, it was at that presentation that I first got the feeling that I could contribute something to the problem of belief attribution, and for that opportunity I am grateful to Michael. Finally, Josep Macià is, like me, a fan of Bob's paper "Assertion," and I am grateful to him for the many illuminating conversations we have had on this and other subjects, on both sides of the Atlantic.

I though that the graduate students at MIT, past and present, are exceptional, both as philosophers and as friends. I feel very fortunate to have been at MIT when they were around, and to share in the special atmosphere that they created. All of them helped me in my work, both through sessions of the MATTI group (a reading group in which graduate students read and commented on each other's work), and individually. I am grateful to them all. I have special debts to Gabriel Uzquiano, who always wanted to know what I was doing and gave me useful comments and encouragement; to Patrick Hawley, who read the whole manuscript, provided numerous comments, and was constantly interested in my work; and to Matti Eklund, Adam Elga, Ólafur Jónsson, and Agustín Rayo, with whom I had many interesting and fruitful conversations on this and other subjects over the years.

None of this would have been possible without the economic support I received from different institutions. From 1994 to 1996 I held a "la Caixa" fellowship to study in the US, which served finance my first two years at MIT, for which I am grateful. From 1996 to 1998 I had a fellowship from MIT, for which I am also grateful. During the Fall Semester of 1998 I worked part-time on-campus at the office of Professor Rosenberg, and the front desk at Ashdown House; I am grateful to all the people I met there, for making my job at pleasant one, and to Deans Milena Levak and James Colbert, for giving me a waiver of INS and MIT regulations to work on-campus while I was a non-resident student. After the Spring of 1999 I worked as Teaching Fellow at Harvard for Professors Jim Pryor, Alison Simmons, and Jane Mansbridge. I am grateful to all of them for hiring me, and also for how much I learned from them in their lectures and in conversation.

In these last few months I had to work at the Spanish Consulate in Boston, performing the *Prestación Social Substitutoria*, the civil equivalent of the Spanish military service. I am grateful to the Spanish Consul, Sr. Peidró, for granting me, on several occasions, time off to work on my dissertation.

I am also grateful to the many friends I had in Boston, which made my stay a happy one; I miss the ones that left, and I know I will miss the rest if and when I leave town. Thanks, Daniel, Sudhir, Sanjay, Rajesh,

Merche, Rosa, Amador, José Luis, Calixto, and all the philosophy people at MIT, past and present, for all the good times. Especial mention deserve my roommates, Hari, Gustavo, Avram and Massimo; it was so much fun to share cooking, TV, music and parties with you!

Two years and a few months ago I met Loida Aponte, and we have been together ever since. Without her love and support, I simply would not have been able to finish this dissertation. *Gracias, Loida.*

To my family in Spain I owe the greatest debt of all, for their patience, their love, and their support, in spite of the distance. I missed them all, and I dedicate this dissertation to them. *Para vosotros.*

Contents

Chapter 1

Interpretation, Propositional Attitudes, and Recognition Failure

1.1 The Problem

Let us say that someone suffering from *recognition failure* is someone who is confused about the truth value of an identity sentence. An example is Lex Luthor, who has not realized that Superman is Clark Kent. There is a good question about how to describe the belief state of agents suffering from recognition failure, like Luthor. For example, Luthor is willing to assent to pairs of sentences like the following:

(1) Superman can fly

(2) Clark Kent cannot fly

In all likelihood, the reason why Luthor assents to both (1) and (2) is because he is confused about the identity of Superman and Clark Kent, and also about the reference of 'Superman' and 'Clark Kent.' When he assents (1) and (2), he takes himself to have asserted two compatible propositions. The problem is that it seems very hard to say what those propositions could be. They cannot be singular propositions, propositions to the effect that Superman can fly, and that Superman cannot fly, for those propositions are inconsistent. If Frege's suggestion that the semantic value of proper names is a sense were true, then

perhaps we could claim that Luthor takes (1) and (2) to express two general propositions, propositions that describe Superman in different ways, but do not contain Superman as a constituent. But, of course, if Kripke and Donnellan are right in their criticisms of Frege's theories, competent speakers of English do not believe that proper names contribute a sense to the proposition expressed. Neither option appears attractive, and they seem to be all the options[1].

In this chapter I want to take up the problem of describing the belief state of people like Luthor. I find it useful to break the problem up into two parts:

(i) **Does Luthor associate compatible beliefs with (1) and (2)?**

(ii) **If he does, how should those beliefs be described?**

Most of this chapter will be dedicated to answer question (i). My thesis is that Luthor does associate compatible beliefs with (1) and (2). I will defend this thesis by appealing to an interpretivist account of belief content. Question (ii), in contrast, will be answered in this chapter and the next. In these two chapters, I will defend the thesis that belief is a relation to propositions, where propositions are understood as abstract entities which are individuated by their truth conditions, and I will illustrate how the principles discussed as part of the answer to (i) determine which are the propositions believed by Luthor, and by others like him.

To be sure, the theses I will defend are hardly new. Nevertheless, I expect this chapter and the next to contribute to the literature on this topic in at least two different ways. The first way is to draw attention to question (i), which, I think, has been overlooked in many discussions of our topic. When approaching the case of Luthor, many philosophers begin by drawing our attention to the plausibility of the claim that Luthor is rational, and that therefore he does not have contradictory beliefs. These philosophers then go on to concentrate on the question of how those beliefs should be described. Many of those philosophers defend a *fine-grained* view of the object of belief, in order to honor the claim that Luthor does associate compatible beliefs with (1) and (2)[2].

I agree with these philosophers that Luthor does associate compatible beliefs with (1) and (2), but what

[1]For Frege's views, see Frege (1892) and Frege (1918). For the criticisms, see Kripke (1980) and Donnellan (1970).

[2]Examples of the views that I have in mind here are the ones defended in Braun (1991), and Crimmins (1992), esp. chap. 2. Stephen Schiffer endorses a similar line of argument in chapter 3 of Schiffer (1987b), though his conclusion in the book is skeptical, and he does not quite endorse the view that belief should finely individuated.

I find missing in this strategy is that, by itself, the claims that Luthor is rational, and that the object of belief should be fine-grained, do not seem to me to be enough to justify the claim that Luthor does associate consistent beliefs with (1) and (2). Ideally, that claim would need to be justified by proposing a general theory of belief content, a theory that, when applied to the particular case of Luthor, yields the conclusion that he associates compatible objects of belief with (1) and (2), and thus that Luthor is indeed rational. This step is eschewed in much contemporary literature, and this seems to me unacceptable, at least for two reasons[3].

On the one hand, there are valid arguments that are, at least, moderately plausible, and whose conclusion is that Luthor does associate incompatible beliefs with (1) and (2). One of those arguments relies on the notion of linguistic competence: Because Luthor is a competent speaker of English, he knows what he says when he says (1) and (2), and therefore does express support for a pair of inconsistent beliefs. Another relies on causal considerations: Because Luthor's assertions of (1) and (2) are caused by Superman himself, the beliefs that Luthor expresses each time are beliefs about Superman himself, and therefore inconsistent. I will later examine each of these arguments in more depth, but it is clear, I think, that the force of these arguments cannot be ignored. It is no response to these arguments to simply say that Luthor is rational, for this is what the arguments put in question; and, as I will argue later, it is not clear that the suggestion that content is fine-grained, by itself, would help to illuminate what is wrong with these arguments.

On the other hand, basing one's description of Luthor's belief state on the principle that Luthor is rational, or on the auxiliary hypothesis that the object of belief is finely individuated, will leave on the dark other cases of recognition failure that are, intuitively, related to Luthor's case. For example, remember Keith Donnellan's celebrated Aston-Martin example[4]. In that example, we have one subject who takes two different people (a philosopher and a party goer) to be the same. Intuitively, the subject of this case suffers from the same affliction as Luthor, in that she too is confused about some identity; therefore, one would expect a solution to the problem of Luthor to also shed some light on the Aston-Martin example. But the problem is that here the claims that the subject is rational, or that the object of belief should be individuated

[3]This failure is particularly glaring in the works cited in the previous footnote. Though Braun and Crimmins devote a great amount of space to the discussion of the nature of the object of belief, they say very little about what makes it the case that an agent is in the belief relation to an object of belief.

[4]Presented originally in Donnellan (1970).

more finely than by truth conditions, do not help us in the least to understand what it is that our subject believes. It is something of an embarrasment, I think, that many contemporary discussions of recognition failure leave Donnellan's Aston-Martin example completely in the dark.

I hope to avoid these two defects by beginning with the defense of a general theory of belief content, a theory of what makes it the case that an agent has a belief with a certain content. My idea is that this theory will help us to see what is wrong in the arguments from linguistic competence and from causal considerations, and that it will help us to see what it is that people like the subject in the Aston-Martin example believe.

The second way in which I hope that the discussion in this chapter and the next will contribute to the literature is by providing new considerations in defense of the view that belief is a relation between a person and a proposition. In recent years, the claim that cases like Luthor are counterexamples to the traditional view that belief is a relation between a person and a proposition has become something of a commonplace in the literature. I think that this conclusion is unwarranted, and in this chapter and the next I will show why. Though I agree with the opponents of the traditional view that cases of recognition failure present a challenge for the traditional view, I do not think that the challenge has been identified precisely enough, and to do so will be one of my goals. My thesis will be that the challenge raised by cases of recognition failure is *the same* as the challenge raised by cases of belief in fictional entities. In the next chapter I will then explore a proposal about how the traditional view could solve this challenge by trying to make sense of the idea that there is people whose belief state can be described by means of propositions that are about fictional entities.

1.2 An Interpretivist Framework to Describe Recognition Failure

1.2.1 Quasi-Radical Interpretation

Most agents suffering from recognition failure manifest their mistakes by endorsing sentences that cannot be true (like 'Superman is not Clark Kent', or 'Superman can fly but Clark Kent cannot'). We want a description of the belief state of these agents, and I am going to approach this problem from the linguistic

side: I will begin by asking what these agents think that the semantic value of the problematic proper names is, and I will use that as a guide to figure out what else the agent believes.

To accomplish this goal, I will adopt the perspective of what I am going to call a *Quasi-Radical Interpreter*, or *QRI*, for short. The QRI is someone who tries to figure out what is the semantic value that a given subject *S* attributes to the proper names in *S*'s vocabulary. To figure that out, the QRI has the following evidence at her disposal:

- The agent's physical constitution
- The agent's environment (including the beliefs of the people around her)
- The agent's behavior (including verbal behavior)
- The agent's history
- The agent's beliefs about the semantic properties of her language, except for the semantic properties of proper names
- Those of the agent's beliefs and desires that are *general* beliefs and desires

Perhaps the notion of a *general* belief and desire bears a bit of explanation. The idea is that a general belief is one whose content is a general proposition, a proposition to the effect that the unique satisfier of a certain condition is thus-and-so. General beliefs and desires contrast with *singular* beliefs and desires, which are beliefs and desires whose content is a singular proposition, a proposition to the effect that a certain object is thus and so. General and singular propositions differ in that, while singular propositions are about an object and depend, for their very existence, on the existence of that object, general propositions are not about any object in particular, and do not depend on any object for their existence[5].

The notion of a QRI is of course intended to bring to mind Davidson's notion of a *radical interpreter*, but note well that the task of the QRI is somewhat different: The QRI has to figure out fewer things, and has more evidence at her disposal[6]. I think there are two reasons why it is a good idea to adopt a perspective similar to that of the radical interpreter, when approaching the problem of describing the belief state of people suffering from recognition failure. One is that, by presenting the problem in this way, we gain a clear perspective on the evidence available to us to solve our problem, without begging any important questions. The other reason is that, by approaching the problem from the perspective of a QRI, we gain a

[5] I suppose this distinction between singular and general propositions can be found in many places; one place where it is articulated is Neale (1990), chapter 2.

[6] For Davidson's notion of a radical interpreter, see essay 11 of Davidson (1980), and essays 9 to 11 of Davidson (1984).

certain *detachment* from the agent to be interpreted, which is, I think, beneficial. Tyler Burge once remarked:

> It is perhaps surprising that one needs to theorize about proper names. They seem to present a straightforward, uncomplicated examples of how language relates to the world... (Burge (1973), p. 425)

Indeed proper names seem so straightforward that when we are asked to theorize about how other subjects use proper names themselves, and especially proper names that we ourselves use, it is only too easy to project our own beliefs about those proper names onto those agents. When we then see that subjects suffering from recognition failure, like Luthor, assent to sentences that we know express contradictory propositions, we tend to interpret those sentences according to our beliefs, which would yield the premature conclusion that Luthor has contradictory beliefs. At this point, we cannot yet assume that Luthor does not have contradictory beliefs, but by approaching the problem from the perspective of the QRI, we can at least check our inclinations to describe the case in this way, and wait until we see some substantive argument in favor of that description.

1.2.2 Some General Principles of Interpretation

To figure out the belief state of people like Luthor, the QRI will have to rely on certain principles about how the process of interpretation has to proceed, which I am going to call *Principles of Interpretation*; it is thus important to specify what these principles say.

How should we choose the Principles of Interpretation? There are two considerations that will guide us in this task. One is that they be principles of which it is plausible to say that ordinary people would be willing use them, at one time or another, when trying to figure out what other people think. The other consideration is that, in some sense, the principles have to yield the right results. By this I do not mean that there is an antecedent agreement about how the belief state of people suffering from recognition failure ought to be described; that would be question-begging. What I mean is that there are some cases of subjects for which it is simply not controversial how their belief state ought to be described, and that the Principles of Interpretation should be chosen so that they yield the right results in these non-controversial cases. (Of course, none of these non-controversial cases will be a case of recognition failure.)

Let us then try to figure out which are the principles guiding the task of the QRI. Some of these principles will be general principles, principles that, in all likelihood, Davidson's radical interpreter would also use[7]. I will begin by presenting these more general principles.

The first principle reflects the fact that, in our ordinary interpretive practices, we tend to avoid attributing contradictory beliefs to the agents we are interpreting:

> COHERENCE: Other things being equal, rational agents do not believe a proposition and its negation

The second principle reflects the fact that we also tend to avoid attributing false beliefs:

> CHARITY: Other things being equal, rational agents tend to have true beliefs

Our third principle is motivated by the thought that the verbal behavior of rational agents tends to be a reliable indicator of what they believe. But because we cannot presume that all agents will speak our language, we will have to formulate this constraint as demanding that the agent's acts of assent and dissent express the agent's beliefs and disbeliefs, but only if the assents and dissents in question are interpreted according to the rules of the language our agent speaks:

> SINCERITY: Other things being equal, a rational agent A tends to act in a way such that A assents to a sentence S (in context C) if, and only if, there is a proposition P such that S expresses P (in C) according to the language A speaks and A believes P

Ideally, a description of an agent's belief state will try to satisfy all these three principles, but there will be many cases in which it will not be possible to satisfy all three of them. How should the radical interpreter react, in such a case?

It is true, I think, that in our ordinary practice we seek, first and foremost, not to attribute incoherent beliefs to others. What this suggests is that the principle of COHERENCE is more important than the others, and the radical interpreter should give it special attention. But this still leaves the interpreter too much

[7]See in particular Lewis (1974), who offers a very thorough discussion of all the constraints that one might reasonably impose on the process of radical interpretation.

latitude. For example, suppose that an otherwise normal speaker of English asserts a sentence S that, in English, expresses a false proposition. On the face of it, that will leave the interpreter with at least three options: Either say that the agent has false beliefs about the situation described by the sentence; or say that the agent has false beliefs about what the sentence means; or else say that the agent is not being sincere (or perhaps a combination of these three). There is no predetermined doctrine about which of these options we should pursue; rather, as interpreters, we will first investigate the details of the situation, before deciding.

To illustrate this, let me run through a few examples that are, I think, uncontroversial, and that help to make this point vivid. The first example is this:

> EXAMPLE 1: Suppose that an agent who has been a normal speaker of English for many years, takes a vacation in a place where she expects to find dinosaur fossils. After inspecting the ground, asserts 'There were dinosaurs here.' Moments before we have seen her inspecting what as a matter of fact is a whale bone, not of a fossil of dinosaur.

In this case, it seems clear that the agent has formed the mistaken belief that there were dinosaurs on the ground, because she mistakenly believes that the bone she is examining is a dinosaur bone, rather than a whale bone. There is nothing in the case that suggests that she may be mistaken about the meaning of 'There were dinosaurs here,' or that she is not saying what she thinks. In this case, we violate CHARITY, by attributing her false beliefs about the bone she is examining. At the same time we attribute to her the correct beliefs about her language, and maintain that the agent is sincere.

> EXAMPLE 2: Suppose a foreigner, recently arrived from a non-English-speaking country, tells one of his friends, in English, 'Johnny is a good freak.' We observe that she has had a friendly relationship with Johnny up to that point, and that Johnny is not a freak.

In this case, there is good reason to think that the agent has mistaken beliefs about the meaning of 'freak': The agent is learning English, and seems to have mistaken 'freak' for 'friend.' The right description of this case seems to be this: She believes that Johnny is a good friend of her, and that 'freak' means friend. In this case, we violate CHARITY, by attributing to our agent mistaken beliefs about language, and we maintain that our agent is sincere.

EXAMPLE 3: Suppose an agent, physically normal and raised in an English-speaking country, says: 'It is raining here today.' Moments before we have observed her peeking through the window, and we know that, as a matter of fact, today it is a very sunny day —not a cloud in the sky.

In this case, there is good reason to think that our agent is lying. There is no reason to think that she does not know the meaning of 'It is raining,' and plenty of reason to think that she believes that it is sunny. Therefore the only remaining possibility is that she is trying to deceive us. In this case, we violate SINCERITY at the expense of CHARITY.

These examples show that, in our interpretive practices, we react in different ways to agents who assert false sentences. The interesting question is what guides us in choosing one or another course of action. Here, my hypothesis is that, as interpreters of other people, we are equipped with some tacit theories about the normal way in which people get certain beliefs, and that these tacit theories guide us in the process of describing the beliefs of other people. Our three examples illustrate this point in two ways.

We all have some idea about how people come to have the correct beliefs about the meaning of the sentences in their language: It is by learning the language either from one's family and friends, as one grows up, or by going to language courses, if one is already grown up. When an agent has not accomplished either of these things in respect to language L, and we observe that the agent asserts a sentence of L which we know to express a false proposition, we, as interpreters, would be justified in attributing to her the belief that she does not know what the sentence in question means. This is why in examples 1 and 3 we expect our agents to have the correct beliefs about the language they speak, but we do not expect that in example 2.

We also have some idea about how people come to have beliefs about the weather, and about whether something is a whale bone. In the case of the weather, we think that perception is usually enough. This is why, in example 3, we expect the agent to have the belief that it is sunny today. That is why we feel justified in attributing to her the belief that it is sunny here today, in spite of what she says. In the case of the whale bone, we do not think that visual inspection of a whale bone alone will be enough to confer to a speaker the belief that the bone is a whale bone rather than a dinosaur bone. The right way to decide this question would be to take the bone to a lab and perform some tests on it. The agent has not done that, so that, from

our point of view, it would not be surprising if she failed to acquire the correct belief that the bone is a whale bone, rather than a dinosaur bone. That is why we feel justified in attributing to her the false belief that the bone she is examining is a dinosaur bone, even if the bone is really a whale bone.

I am not sure that there is an informative way of formulating this *normality* requirement, beyond saying the following:

> NORMALITY:
> For any belief B, there are two sets of conditions C, C' such that:
> (i) If an agent A is in C, it would be normal for A to have the belief that B
> (ii) If an agent A is in C', it would be normal for A *not* to have the belief that B

This constraint would then have to be developed, in connection with specific beliefs. The remarks above suggest how to do this in connection with specific beliefs about the weather, about whether a bone is a whale bone, and about the meaning of 'friend;' but they leave open the question about the normality conditions for other kinds of beliefs.

Let me emphasize that, as I conceive of this NORMALITY constraint, a radical interpreter cannot rely solely on it to interpret the beliefs of other people. The interpreter must surely allow, for example, for the possibility that a speaker might get a belief B in a way that is not normal, or for his failure to gain a belief B in a situation in which other agents normally would have belief B. As I conceive of it, this NORMALITY constraint is a *tiebreaker* that comes into action whenever the other principles of interpretation suggest more than one description of a belief state.

One may of course be disappointed that the principle of NORMALITY is so uninformative. Fortunately for us, we are only interested in the process of radical interpretation, insofar as it overlaps with the task of the QRI. And the task of the QRI is quite circumscribed: It is just to figure out what certain agents believe about the semantic value of proper names. From this point of view, the QRI only needs to get clear on what are the normal conditions under which one is expected to get the correct beliefs about the semantics of proper names, to put the NORMALITY constraint to work.

1.2.3 Proper Names and Interpretation

The QRI is concerned with figuring out what is the semantic value that certain confused people attribute to certain proper names. To decide this question, we need to settle on a hypothesis about what kind of semantic value proper names tend to have, and about what are the facts that determine which proper name gets which semantic value. I shall propose several principles that help the QRI to get an answer to these questions.

In the first place, the QRI can entertain two hypotheses about what kind of semantic value a subject attributes to a proper name: Either the subject thinks that the value of a proper name is its referent, or she thinks that it is something like a Fregean descriptive sense. Here I will assume that one moral of Kripke's and Donnellan's arguments in favor of *Direct Reference* is that competent speakers of English think (or tend to think) that the semantic value of a name is its referent, rather than a descriptive sense[8]. This suggests the following normality condition:

> NON-DESCRIPTIVISM: An agent who learns about English in the normal way would normally have a belief to the effect that the contribution of a proper name to the proposition expressed is its referent, rather than a descriptive sense or a mode of presentation

NON-DESCRIPTIVISM implies that, when the QRI has good evidence that an agent A under investigation is a competent speaker of the language, the task of the QRI reduces to the task of figuring out which are the objects (if any) which A takes to be the referents of the names in her vocabulary. In all of the examples that we will examine, the agents will be competent enough to know this, so that in what follows we can concentrate on this question.

Saul Kripke proposed an attractive picture of how proper names relate to their referents. Kripke emphasized that the reference of our names does not depend just on what we ourselves think about it, but also on how other people in our community use the name, and on the history of how the name reached us[9]. This suggests that the normal situation in which one gets to have the belief that N refers to O is by being part of a community where, as a matter of fact, N is used, and has been used, to refer to O. But, of course, there are

[8]For these arguments, see Kripke (1980) and Donnellan (1970).

[9]See Kripke (1980), p. 95. For interesting discussion of Kripke's proposal, see Evans (1982a).

circumstances in which this normal mechanism may malfunction. For example, if the agent picks up the name by hearing some random conversation on the street, that does not seem enough to justify attributing to him the correct beliefs about the reference of the name.

This raises the following question. Suppose that the QRI has to interpret an agent A who has in her vocabulary a name N, and that A belongs to a community which uses N to refer to an object O. The question is, which are the conditions under which the QRI would be justified in *not* attributing to A the belief that N refers to O? We will discuss three kinds of situations.

First, we have the case of the agent who picks up a name N from a random conversation in the street. Intuitively, that should not guarantee that the agent has the correct belief about the reference of N. What is wrong is that the agent has picked up N in a way that does not allow her to connect it with O, and as a result there are no causal dependencies between the agent's utterances containing N, and the object O itself. Accordingly, this suggests the following normality condition:

> CAUSALITY:
> For any predicate F and object O, if an agent A's assertions of sentences of the form $\ulcorner N$ is $F \urcorner$ are never caused by O's satisfying F, then it would not be normal for A to have the belief that N refers to O

The second case concerns ambiguous names, names that have more than one referent. Suppose for example that an agent belongs to a community that uses N to refer to two objects, O and O'. If the agent picks up N in a way that does not allow her to realize that the name is indeed ambiguous (for example, if no one tells her that the name is ambiguous, and she never hears things like $\ulcorner N$ is $F \urcorner$ and $\ulcorner N$ is not $F \urcorner$, or any other similar utterances that may suggest that the name is ambiguous), then the QRI cannot assume that the agent has realized that the name is ambiguous. As a result, the QRI will be justified in not attributing to the agent at least one of the beliefs that N refers to O and that N refers to O'. This suggests the following normality condition:

> AMBIGUITY:
> If an agent A has a name N in her vocabulary, A belongs to a community in which N is indeed ambiguous, and A has picked up N in a way that does not allow A to infer that N is indeed ambiguous, then A would not normally have the belief that A is ambiguous

The third case concerns pairs of names that have the same referent. Suppose for example that an agent belongs to a community that uses a pair of names N and M to refer to an object O. There are several ways in which the agent can pick up the names, while failing to realize that the names are coreferential. For example, if she picks up both of the names in a conversation, but there is nothing in the conversation to suggest that the names are coreferential (no one says $\ulcorner N$ is $M\urcorner$, or $\ulcorner N$ is the unique $F\urcorner$ and $\ulcorner M$ is the unique $F\urcorner$, or any other similar things that may suggest that the names are coreferential), then the QRI cannot assume that she has realized that the names are coreferential.

This can occur even if the agent is herself responsible for the introduction of the name into the language. If Kripke and Donnellan are right, names tend to be introduced in the language by means of some kind of "baptismal" ceremony, in which a certain object is demonstrated and the name is assigned to it. However, one can perform this ceremony with two different proper names, and fail to realize that one is, in effect, baptising the same object twice over. If, for example, the two ceremonies occur in circumstances in which the baptised object has different manifest properties, then chances are that the subject will not realize that those names she has introduced have the same referent.

These two points suggest the following normality condition:

> COREFERENTIALITY:
> If an agent A has names N and M in her vocabulary, and A belongs to a community in which N and M are coreferential, then:
> (i) If A has picked N and M in a way in which it is not obvious that N and M are coreferential, then A would not normally have the belief that N and M are coreferential
> (ii) If A is herself responsible for introducing N and M into the language, and has done so in situations in which the object baptised as N and M had different manifest properties each time, then A would not normally have the belief that N and M are coreferential

Let me now summarize. The idea is that the normal way to get the belief that proper name N refers to O is by belonging to a community in which N refers to O, and that this normal mechanism may malfunction in at least three different occasions:

NORMALITY FOR REFERENCE OF PROPER NAMES:

If an agent A has a name N in her vocabulary, A belongs to a community that uses N to refer to O, and A has picked up N from other members of the community, then A would normally have the belief that N refers to O, except, possibly, if any of the following occur:

- A's assertions of sentences of the form $\ulcorner N$ is $F \urcorner$ (for any predicate F) are never caused by O's satisfying F
- N is ambiguous, and A has picked up N in a way that does not allow A to infer that N is indeed ambiguous
- There is a name M that A has in her vocabulary and in A's community is coreferential with N, and either (i) A has picked N and M in a way in which it is not obvious that N and M are coreferential, or else (ii) A is herself responsible for introducing N and M into the language, and has done so in situations in which the object baptised as N and M had different manifest properties each time

Let me emphasize that this rule is what I have called a *normality* condition, and as such it cannot drive the interpretation process all by itself. On my view, normality conditions act as tiebreakers: Whenever the QRI is forced to violate some principle of interpretation to describe the belief state of an agent, and there are several possibilities about which principle to violate, and about how to do it, the QRI will appeal to this normality condition to decide which description to choose.

Let me now show how this interpretivist framework can be applied to the description of subjects suffering from recognition failure.

1.3 Donnellan Cases

The first case of recognition failure that we will examine can intuitively be described as a case in which an agent takes *two* different people to be the same. Because Keith Donnellan was the first in drawing philosophical attention to these cases, I am going to call them *Donnellan Cases*. Here is an example, inspired in Donnellan's own celebrated Aston-Martin example[10]:

ASTON-MARTIN: Suppose that Jones has heard of the philosopher Aston-Martin, but has never met him before. One day at a party, he finds a man that calls himself 'Aston-Martin,' and Jones

[10]For which, see Donnellan (1970), esp. pp. 370–372.

believes that he is in front of the great philosopher. Unfortunately for Jones, he is not really in front of the philosopher, but of a different man who is not a philosopher but also happens to have the name 'Aston-Martin.' Jones talks to this man for some time, but the conversation proceeds in a way such that Jones never realizes that he is not in front of the philosopher.

After the party, Jones is inclined to assent to the following sentences containing the name 'Aston-Martin:'

- The great philosopher and the man I met at the party are the same person
- There is just one person called 'Aston-Martin'
- Aston-Martin is a great philosopher
- I met Aston-Martin at a party
- Aston-Martin drinks bourbon
- Aston-Martin has three daughters
- Aston-Martin lives in Boston

For the sake of clarification, let us distinguish between the two men who are called 'Aston-Martin' by means of subscripts: Let us use 'Aston-Martin$_{philosopher}$' for the philosopher Aston-Martin, and 'Aston-Martin$_{partygoer}$' for the Aston-Martin who was at the party. And let us suppose that, as a matter of fact, Aston-Martin$_{philosopher}$ is a great philosopher, but that he does not drink bourbon, does not have three daughters, and does not live in Boston; and also that Aston-Martin$_{partygoer}$ is not a great philosopher, but drinks bourbon, has three daughters, and lives in Boston.

What makes this case difficult is that there is evidence in favor of both the hypothesis that Jones takes 'Aston-Martin' to refer to Aston-Martin$_{philosopher}$, and of the hypothesis that he takes 'Aston-Martin' to refer Aston-Martin$_{partygoer}$. For example, Jones' disposition to assert 'Aston-Martin is a great philosopher' suggests that he takes 'Aston-Martin' to refer to Aston-Martin$_{philosopher}$; while his disposition to assert 'I met Aston-Martin at a party the other day' suggests that he takes 'Aston-Martin' to refer to Aston-Martin$_{partygoer}$. Jones' causal history certainly does not settle the matter, since some of his assertions are caused by Aston-Martin$_{philosopher}$ and others by Aston-Martin$_{partygoer}$.

This is a case in which, to settle the matter, we will have to appeal to the principle of CHARITY. Here the relevant fact here is that, of all the sentences that Jones is willing to assent to after the party, more of those would come out true if interpreted so that 'Aston-Martin' refers to Aston-Martin$_{partygoer}$, than if interpreted so that 'Aston-Martin' refers to Aston-Martin$_{philosopher}$. To see this, consider the following chart, which displays the sentences that Jones is willing to assert, and the propositions that he would take those

sentences to express, on the hypothesis that he takes 'Aston-Martin' to refer to Aston-Martin$_{partygoer}$:

SENTENCES JONES ACCEPTS	PROPOSITIONS JONES BELIEVES
•"The great philosopher and the man I met at the party are the same person"	•The great philosopher and the man Jones met at the party are the same person
•"There is just one person called 'Aston-Martin' "	•There is just one person called 'Aston-Martin'
•"Aston-Martin is a great philosopher"	•Aston-Martin$_{partygoer}$ is a great philosopher
•"I met Aston-Martin at a party"	•Jones met Aston-Martin$_{partygoer}$ at a party
•"Aston-Martin drinks bourbon"	•Aston-Martin$_{partygoer}$ drinks bourbon
•"Aston-Martin has three daughters"	•Aston-Martin$_{partygoer}$ has three daughters
•"Aston-Martin lives in Boston"	•Aston-Martin$_{partygoer}$ lives in Boston

This hypothesis yields the result that Jones has four true beliefs, and three false beliefs. It is easy to see that the hypothesis that Jones takes 'Aston-Martin' to refer to Aston-Martin$_{philosopher}$ would attribute to Jones more false beliefs (in particular, six false beliefs, and only one true belief). Therefore, CHARITY supports the description of Jones' belief state displayed in the chart.

Our description of the case assumes that Jones has not realized that the name 'Aston-Martin' is ambiguous. One could wonder what would happen if we did not make this assumption. If one could make the claim that Jones has realized that the name 'Aston-Martin' has two different referents, Aston-Martin$_{philosopher}$ and Aston-Martin$_{partygoer}$, then one could assign more true beliefs to Jones. For example, we could assign to Jones the belief that Aston-Martin$_{philosopher}$ is a great philosopher and that Jones met Aston-Martin$_{partygoer}$ at a party, both of which are true; and so on. Thus, on the face of it, CHARITY seems to give us a reason to describe Jones as believing that the name 'Aston-Martin' has two referents.

The problem that this description would have to face is the problem of making sense of Jones' willingness to assert that "There is only one person named 'Aston Martin'." This option would have to say that, when Jones says this, either he is not being sincere, or he does not know what this sentence means. However, there is no evidence that he is not being sincere (why would he lie?), nor is there any reason to expect that he does not know what this sentence means. On the other hand, we know that normal speakers of a

language can fail to realize that a proper name of their language has more than one referent. The normal ways in which people come to realize that a name N is ambiguous is by being told that N is indeed ambiguous, or by deducing from the way other speakers talk that, as a matter of fact, N is used to talk about two different people. There is nothing in Jones' history to suggest that he has accomplished either of these two things; therefore, the NORMALITY constraint forces the QRI to reject this description in favor of the one proposed above.

It is important to note that other versions of the ASTON-MARTIN story might yield different results. If Jones has an extensive body of information about the philosopher, and his interaction with the man at the party is short, then CHARITY would probably enjoin the QRI to say that Jones takes 'Aston-Martin' to refer to Aston-Martin$_{philosopher}$, rather than to Aston-Martin$_{partygoer}$. Also of interest is a version of the case (mentioned by Donnellan) in which everything goes more or less as explained in ASTON-MARTIN, except for the fact that the man at the party actively tries to *impersonate* the great philosopher. If this is the case, even though Jones has been in more direct contact with Aston-Martin$_{partygoer}$ than with Aston-Martin$_{philosopher}$, CHARITY would probably enjoin the QRI to say that Jones takes 'Aston-Martin' to refer to Aston-Martin$_{philosopher}$, since he would have more information about this man.

It is also worth noting that our example was intentionally arranged so that the evidence in favor of the hypothesis that Jones takes 'Aston-Martin' to refer to Aston-Martin$_{partygoer}$ outweighs the evidence in favor of the hypothesis that he takes it to refer to Aston-Martin$_{philosopher}$, but we could easily imagine a case in which there is the same amount of evidence in favor of each hypothesis. Example: Like in ASTON-MARTIN, except that Jones never asserts, or becomes inclined to assent to, 'Aston-Martin drinks bourbon,' 'Aston-Martin has three daughters,' and 'Aston-Martin lives in Boston.' In this case, and from the perspective of the QRI, there will be two equally plausible hypotheses about who Jones believes to be the referent of 'Aston-Martin,' but the QRI will not have evidence to decide between them. Nevertheless, the QRI can still say that, even in this case, Jones believes that the name 'Aston-Martin' refers to just one person, and that it is either to Aston-Martin$_{philosopher}$ or to Aston-Martin$_{partygoer}$, but that he does not know which[11].

[11] Michael Devitt suggests some cases related to Donnellan's Aston-Martin example in Devitt (1981), pp. 138–152. Devitt suggests that, when our uses of a name are causally grounded in more than one object, it is indeterminate which of those objects is the referent of the name. While this is surely true of the version of the Aston-Martin example in which there is the same amount of evidence in

My general proposal about Donnellan cases is thus this. Donnellan cases can be intuitively described as cases in which an agent takes two different objects to be the same, and has in her vocabulary a name N which her community uses to refer to both of those objects. The idea is to appeal to the principle of CHARITY to decide which of those two objects is the one which the agent takes to be the referent of N.

1.3.1 Slow Switching

It is worth noticing a curious consequence of our interpretivist framework, a consequence that Donnellan cases help to bring to light. As we have explained the Aston-Martin example, Jones believes that the name 'Aston-Martin' refers to Aston-Martin$_{partygoer}$, and not to Aston-Martin$_{philosopher}$. Let us agree that the story occurs at a time t_1, and that Jones has this belief at t_1. Let us now suppose that after t_1, Jones goes on to meet Aston-Martin$_{philosopher}$, and that he gets to know him in depth. At the same time, Jones does not meet Aston-Martin$_{partygoer}$ ever again. Suppose also that, in addition, Jones does not have any reason to suspect that Aston-Martin$_{philosopher}$ was not the man he met at the party. At that moment —call it t_2—, what does Jones believe? Well, since at t_2 has interacted more with Aston-Martin$_{philosopher}$ than with Aston-Martin$_{partygoer}$, and since probably most of his assertions containing the name 'Aston-Martin' would come out true, if interpreted so that he takes 'Aston-Martin' to refer to Aston-Martin$_{philosopher}$, the QRI will conclude that at t_2 Jones believes that 'Aston-Martin' refers to Aston-Martin$_{philosopher}$. Thus, the interpretivist framework that I am assuming implies that Jones' beliefs about the reference of 'Aston-Martin' changed between t_1 and t_2, even if Jones himself did not notice it.

One may find this consequence surprising, but the truth is that any theory with externalist commitments is committed to the existence of agents whose belief state changes in this way; let me explain. Externalism is the thesis that the content of at least some of our beliefs depends on our environment, and this thesis is defended by appealing to thought experiments in which physically identical people grow up in different environments and get to have beliefs with different content. Recent externalist literature has drawn our attention to the phenomenon of *slow-switching*, illustrated by thought experiments in which we imagine an

favor of each hypothesis, it seems too radical for other cases in which there is more evidence in favor of the hypothesis that one of those objects is the referent, like the main version of the case discussed in the text.

agent who is switched back and forth between different environments. The intuition is that, if the agent remains in each environment long enough, the content of her beliefs will change with each switch, without the agent noticing it[12].

From our point of view, what we see in the version of the Aston-Martin case summarized two paragraphs ago is the same phenomenon that we observe in Slow Switching cases. Thus, this consequence of our interpretivist framework is more common than one may have thought, since any theory that is committed to externalism will have to deal with similar kinds of belief change.

In any case, the fate of the interpretivist framework that I have been defending is tied to the fate of some form of externalist thesis, in more ways than one. One way is by being committed to the existence of belief changes of the same sort that externalists are committed to. Another is more direct, and is by being committed to some sort of externalist thesis. The crucial principle, in this respect, is CHARITY. We could have two agents who are physically alike in all respect, and whose behavior (non-intentionally described) is the same; but if their environments are different, they many nonetheless have beliefs with different contents, since the CHARITY principle will be satisfied in relation to different environments in each case. Whether this is a fatal objection is something that I will leave for another occasion.

1.4 Frege Cases

Surely the most celebrated kind of recognition failure is the case which an agent could be described as taking *one* single object to be *two* different ones. Because Gottlob Frege was probably the philosopher that did more to bring philosophical attention to such cases, I am going to call these cases *Frege Cases*[13]. In this section I will discuss what the QRI should say about Frege cases. One of the points I am going to make is that, from the point of view of the QRI, there are some interesting similarities between Frege cases and cases of belief about fictional entities. As we will see later, this point will be important in my defense of the view that belief is a relation between a person and a proposition. I will therefore begin by explaining how the QRI should approach the case of subjects who believe in fictional entities, and then move on to discuss

[12] To my knowledge, Tyler Burge was the first to draw attention to slow-switching cases. See Burge (1988).
[13] Frege discussed these cases in Frege (1892) and Frege (1918).

Frege cases proper.

1.4.1 Belief About Fiction

Some people are, or at least seem to be, unfortunate enough to believe in entities that do not exist. In some cases, the beliefs in question are related to some story or myth, as in the following example:

> SANTA CLAUS: Johnny is a four-year old who says that he expects Santa Claus to bring him a truck. He also says that Santa Claus wears a red suit, has a white beard, and rides a flying sleigh; and that, on Christmas Eve, he brings gifts to all (well-behaved) children of the world. He claims to have seen Santa Claus in no less than four department stores, and again claims to be in front of Santa Claus when his father, dressed as Santa Claus, comes down the chimney to bring him gifts on Christmas Eve. As a matter of fact, each department store has different people posing as Santa Claus, and there exists no one who rides a flying sleigh and delivers gifts to all children.

Intuitively, the problem with Johnny is that he has been fooled into taking the Santa Claus myth as a fact. However, it is possible to believe in a fictional entity that does not play any role in any fiction, as in the following example:

> VULCAN: Suppose Smith discovers certain perturbations in the orbit of Mercury. She attributes the phenomenon to the existence of a planet that she dubs 'Vulcan.' On her view, Vulcan is different from the other nine planets of the Solar System. She says, for example, that there are ten planets, and Vulcan is one of them. As a matter of fact, Smith is mistaken: The perturbations in the orbit of Mercury are due to some relativistic phenomenon, and not to the existence of a tenth planet. Vulcan does not exist.

Smith believes that Vulcan is a real planet, and she is wrong. But what leads her to make this mistake is not a myth, but rather an inductive failure: She infers the existence of Vulcan from certain perturbations in Mercury's orbit, but the perturbations are not really due to any planet.

From the point of view of the QRI, both cases are very similar. In the case of Johnny, his causal history suggests too many candidates to be the object which he takes to be the referent of 'Santa Claus': His father, and four different people that pose as Santa Claus at four department stores. What is more, Johnny attributes to the referent of 'Santa Claus' properties that no one, as a matter of fact, has. From the point of view of the QRI, there simply is no real object of which it could plausibly be said that Johnny takes that object to be the referent of 'Santa Claus.'

In the case of Smith, her causal history does not suggest any particular object as the one which she takes as the referent of 'Vulcan.' In particular, the perturbations of the orbit of Mercury which she takes as indicative of the existence of Vulcan are not caused by any planet at all, but rather by some relativistic effect. Furthermore, she attributes to the referent of 'Vulcan' properties that no real object has: Being the tenth planet, and being a planet responsible for Mercury's orbital perturbations. From the point of view of the QRI, there simply is no real object of which it could plausibly be said that Smith takes that object to be the referent of 'Vulcan.'

How should the QRI describe the belief state of these agents? Well, on the face of it, Johnny believes, or seems to believe, the proposition that 'Santa Claus' refers to Santa Claus; and Smith believes, or seems to believe the proposition that 'Vulcan' refers to Vulcan. I am going to assume that, to handle cases like Johnny's or Smith's, the QRI will appeal to a special sort of propositions, propositions that are about fictional entities. On this view, the QRI will be able to say that Johnny believes the proposition that 'Santa Claus' refers to Santa Claus, understanding that Santa Claus is a fictional entity; and also that Smith believes the proposition that 'Vulcan' refers to Vulcan, understanding that Vulcan is a fictional entity.

Of course, the questions whether there are propositions that are about objects that do not exist, of what would such propositions be like, and of whether we can meaningfully use those propositions to describe the belief state of other people, are very controversial questions, whose force cannot be ignored. For the sake of convenience, I am going to postpone the discussion of these questions until the next chapter. For now, I will assume that the QRI can use propositions that are about fictional entities, and I will see where this leads us. Then, in the next chapter, I will try to justify this assumption.

1.4.2 Hesperus and Phosphorus

Let us finally take up a real Frege case. Consider the following example:

> HESPERUS AND PHOSPHORUS: Charles is an amateur astronomer. He has studied, with his telescope, most of the planets in the Solar System, and knows many things about them. Indeed, he has studied all of them but Venus, which he does not even know exists.
>
> One good evening he notices a bright planet that appears at sunset, and he immediately decides that the planet is different from all the other planets he has studied so far. He names the planet

in question 'Hesperus.' He studies this planet intensively, and during many nights he tries to figure out its exact position, relative to the Sun, and its brightness, relative to other stars and planets.

After a few months, he notices, just by chance, a bright planet visible at dawn. Again he decides that that planet is different from all the planets he has studied before, the one he calls 'Hesperus' included. He names this planet 'Phosphorus.' Because Charles is busy studying the planet that he calls 'Hesperus,' he decides that he is not going to study the one he calls 'Phosphorus,' at least for the time being.

As a matter of fact, Charles has named the same planet twice, the planet Venus. But he does not realize it, and after these episodes, he becomes inclined to assent to the following sentences:

- There are ten planets
- Hesperus and Phosphorus are different planets
- Hesperus is visible at sunset
- Phosphorus is not visible at sunset
- Hesperus is not visible at dawn
- Phosphorus is visible at dawn
- Hesperus always appears close to the Sun
- Hesperus is very bright, but not as bright as Jupiter

What should the QRI say about this case? The fact that Charles is inclined to assent to the sentence 'Hesperus and Phosphorus are different planets' suggests that Carlos takes 'Hesperus' and 'Phosphorus' to refer to two different planets. But it is not easy to say what those planets could be. When Charles uses sentences containing the names 'Hesperus' and 'Phosphorus,' his utterances, in both cases, are caused by the planet Venus. The QRI needs *two different planets* to characterize the beliefs of Charles, but the problem is that she has *only one* candidate.

However, this shortage of candidates does not mean that the QRI cannot say which are the planets that Charles takes to be the referent of 'Hesperus' *and* 'Phosphorus.' The QRI has one candidate, the planet Venus, to play those two roles, and the QRI can certainly determine whether Venus fits one of those roles better than the other. As we have arranged the case, it turns out that Charles has studied Venus more intensively in its evening appearances than in its morning appearances. As a result, he is willing to assent to more true sentences containing the name 'Hesperus,' than containing the name 'Phosphorus.' For example, the QRI asserts 'Hesperus always appears close to the Sun,' and 'Hesperus is very bright, but not as bright as Jupiter,' both of which are true, on the hypothesis that 'Hesperus' refers to Venus. Therefore, CHARITY would compel us to attribute to Charles the belief that 'Hesperus' refers to Venus.

The QRI is left with the problem of explaining which is the object that Charles takes the name 'Phosphorus' to refer to. The problem here is that there is no planet left to play this role, since the other nine have already been used in the description of Charles' belief state. More importantly, this situation should seem familiar. From the point of view of the QRI, the situation of Charles *vis à vis* 'Phosphorus' is very similar to the situation of Johnny *vis a vis* 'Santa Claus,' and to the situation of Smith, *vis à vis* 'Vulcan.' Johnny, Smith and Jones all have names in their vocabulary which they use as if they really referred to something, but the QRI cannot identify any real object as the one that those agents take to be the referent of those names.

As I explained above, I am going to assume that the belief state of people like Johnny, Smith, and Charles is to be described by appealing to propositions that are about fictional entities. In the case of Charles, we will assume that there is a fictional planet that we will call 'Phosphoria,' and that Charles believes, among other things, the proposition that 'Phosphorus' refers to *Phosphoria*. The following chart then describes our proposal about Charles' belief state:

SENTENCES CHARLES ASSENTS TO	PROPOSITIONS CHARLES BELIEVES
•"There are ten planets"	•There are ten planets
•"Hesperus and Phosphorus are different planets"	•Venus and Phosphoria are different planets
•"Hesperus is visible at sunset"	•Venus is visible at sunset
•"Phosphorus is not visible at sunset"	•Phosphoria is not visible at sunset
•"Hesperus is not visible at dawn"	•Venus is not visible at dawn
•"Phosphorus is visible at dawn"	•Phosphoria is visible at dawn
•"Hesperus always appears close to the Sun"	•Venus always appears close to the Sun
•"Hesperus is very bright, but not as bright as Jupiter"	•Venus is very bright, but not as bright as Jupiter

On this view, Charles has many false beliefs (all those that are about the fictional planet Phosphoria, and also the beliefs that there are ten planets, and that Venus is not visible in the morning), but also has some true beliefs (namely, the belief that Venus is visible at sunset, that it is close to the Sun, and that it is very bright). It is easy to see that the alternative hypothesis that Charles takes 'Phosphorus' to refer to Venus, and 'Hesperus' to a fictional planet, would not attribute as many true beliefs to Charles, since, as the example

has been set up, Charles asserts more sentences containing the name 'Hesperus' than containing the name 'Phosphorus.'

What about the hypothesis that Charles takes both 'Hesperus' and 'Phosphorus' to refer to Venus? That would imply that, when Charles asserts pairs like 'Hesperus is nice' and 'Phosphorus is not nice,' he believes that those sentences express contradictory propositions. The QRI cannot plausibly say that Charles is unsincere in this situation (why would he lie?), so we must assume that Charles means what he says. But if what Charles means what he says, then this hypothesis cannot avoid attributing to Charles contradictory beliefs (via SINCERITY). The problem with this strategy is that, as explained above, the QRI should always be very reluctant to attribute contradictory beliefs to an agent; she should always check whether there is some more plausible alternative.

Indeed, there is a more attractive alternative, and that is the hypothesis that Charles has not realized that the names 'Hesperus' and 'Phosphorus' are coreferential. Charles introduces 'Hesperus' by demonstrating Venus in the evening sky, when it occupies a certain celestial position P; and he introduces 'Phosphorus' by demonstrating Venus in the morning sky, when it occupies position P'. It is not obvious that the object that occupies position P should be the same as the object that occupies position P', and there is nothing in Charles' history that suggests that he has worked it out. This shows that Charles is not in a position to realize that the names 'Hesperus' and 'Phosphorus' are coreferential. Therefore, if the QRI has to choose between describing Charles as incoherent, and describing him as mistaken about 'Hesperus' and 'Phosphorus,' the NORMALITY constraint gives her a reason to opt for the latter option.

Let me summarize my proposal about Frege cases. In Frege cases, the agent suffering from recognition failure is mistaken about the truth value of some statement of the form $\ulcorner N \text{ is } M \urcorner$, for some proper names N and M. Normally, the QRI will be able to identify only one candidate O to be both the object which the subject takes to be the referent of N, and to be the object which the subject takes to be the referent of M. The QRI should then say that our agent takes either N or M to refer to O, depending on which option satisfies CHARITY best; and that our agent takes the other name to refer to a fictional entity, an entity which as a matter of fact does not exist.

1.5 Mixed Cases

It is interesting to note that our interpretivist framework leaves us room to characterize some cases of recognition failure as *mixed cases* which share the features of both Frege and Donnellan cases. Consider, for example, the following example:

> LUTHOR AND SUPERMAN: Because Superman is always interfering with his plans, Lex Luthor plans to eliminate him. He has guessed correctly that Superman works under cover at the Daily Planet, because Superman's appearances always occur near the Daily Planet. He thinks it would be easier to eliminate him when he is assuming his undercover identity, so he tries to figure out who, of all the people working in the newspaper, is Superman. He considers the possibility that Clark Kent may be Superman, but he quickly dismisses it.
>
> At last, Luthor sets his sights on Perry White, the chief editor of the *Daily Planet*. Being the chief editor of the *Daily Planet*, Perry White has at his disposal a great deal of information, something that Luthor thinks is necessary to be a superhero like Superman. Luthor has also tried to figure out Superman's origins, and he concludes that Superman grew up in Montana. In this, Luthor is mistaken, since Superman grew up in Kansas; but he nevertheless believes that he has found another valuable clue when he learns that Perry White did grow up in Montana. At that point he tells to himself: 'I figured it out, Superman is Perry White.' After this, Luthor begins to use the names 'Superman' and 'Perry White' interchangeably. In particular, he becomes inclined to assent to the following sentences:
>
> - Superman can fly
> - Clark Kent cannot fly
> - Perry White can fly
> - Perry White is the editor of the Daily Planet
> - Superman is the editor of the Daily Planet
> - Clark Kent is not the editor of the Daily Planet
> - Perry White is from Montana
> - Superman is from Montana
> - Clark Kent is not from Montana

There are two sides to this story. On the one hand, it is clear that Luthor is confused about the identity of Superman and Clark Kent: They are the same person, but Luthor thinks that they are different people. This part of the story makes it a Frege case. On the other hand, it is clear that Luthor is confused about the identity of Superman and Perry White: He takes them to be the same person, but as a matter of fact they are two different people. This part of the story makes it a Donnellan case too. How should the QRI describe Luthor's belief state?

Let us focus first on the Donnellan case involved in the problem. It seems clear that Luthor takes the names 'Superman' and 'Perry White' to refer to the same person. The causal history of Luthor suggests two candidates for that role, Superman and Perry White, but it seems clear that Perry White is the better fit, for reasons having to do with CHARITY: Most of the sentences containing the names 'Superman' and 'Perry White' to which Luthor is inclined to assent would be true, if interpreted so that 'Perry White' and 'Superman' refer to Perry White; but this would not be so, if they were interpreted so that they refer to Superman. This part of the problem suggests that Luthor takes 'Superman' and 'Perry White' to refer to Perry White.

Let us next tackle the Frege case. Luthor takes the names 'Superman' and 'Clark Kent' to refer to different people, and the task of the QRI is to figure out who those people could be. Luthor's causal history suggests two candidates to be the person which Luthor takes to be the referent of 'Superman,' namely Superman and Perry White; and only one candidate to be the person which Luthor takes to be the referent of 'Clark Kent,' namely Superman himself. But because Luthor is willing to assent to pairs like 'Superman can fly' and 'Clark Kent cannot fly,' the QRI is forced to look for an assignment that does not make Luthor incoherent. That would be accomplished by the hypothesis that Luthor takes 'Superman' to refer to Perry White, and 'Clark Kent' to refer to Superman.

Both solutions fit together, and suggest this description of Luthor's belief state:

SENTENCES LUTHOR ACCEPTS	PROPOSITIONS LUTHOR BELIEVES
•"Superman can fly"	•Perry White can fly
•"Perry White can fly"	•Perry White can fly
•"Clark Kent cannot fly"	•Clark Kent cannot fly
•"Perry White is the editor of the Daily Planet"	•Perry White is the editor of the Daily Planet
•"Superman is the editor of the Daily Planet"	•Perry White is the editor of the Daily Planet
•"Clark Kent is not the editor of the Daily Planet"	•Superman is not the editor of the Daily Planet
•"Perry White is from Montana"	•Perry White is from Montana
•"Superman is from Montana"	•Perry White is from Montana
•"Clark Kent is not from Montana"	•Superman is not from Montana

According to this description, Luthor has a few false beliefs, but his beliefs are not incoherent. If we had attributed to Luthor the belief that both 'Superman' and 'Clark Kent' refer to Clark Kent, that would imply that Luthor had contradictory beliefs (since he assents to the sentences 'Superman can fly' and 'Clark Kent cannot fly,' and there is no reason to think that he is not sincere). And the option of describing Luthor as believing that 'Superman' and 'Perry White' both refer to Superman, and 'Clark Kent' to someone else, would attribute to him more false beliefs than our proposal. It thus seems clear that CHARITY supports our description of Luthor's belief state.

1.6 The Argument from Linguistic Competence

On our view, when Luthor asserts:

(1) Superman can fly

(2) Clark Kent cannot fly

he is not incoherent, because he takes himself to have expressed a pair of compatible propositions. What we have done in the preceding sections is to defend the claim that he associates a pair of compatible propositions with (1) and (2), and to say a bit (though not much) about which propositions Luthor takes himself

to be asserting. In this section and the next, I will discuss two different arguments against my claim that Luthor associates a pair of compatible propositions with (1) and (2).

The first argument that I will examine begins by drawing our attention to three principles, all of which appear initially plausible:

> COMPETENCE: Luthor is a competent speaker of English
>
> SEMANTICS: (1) and (2) express contradictory propositions
>
> DISQUOTATION: Suppose English sentence S expresses proposition P with respect to context C; and that A is a competent speaker of English who in C is speaking sincerely. Then:
>
> (i) If A assents to S in C, then A is in the belief relation to P
>
> (ii) If A dissents from S in C, then A is not in the belief relation to P

But the conjunction of all these three principles implies that Luthor is incoherent, when he asserts (1) and (2). First, according to COMPETENCE, Luthor knows the propositions expressed by (1) and (2). By SEMANTICS, he knows that those sentences express contradictory propositions. Because he is sincere, by DISQUOTATION it follows that he believes those contradictory propositions[14].

I do not think that this argument can be dispelled just by appealing to the claim that Luthor is rational, since that is, precisely, the negation of the conclusion of the argument. And I do not think that this argument can be dispelled either by appealing to some fine-grained conception of the object of belief. For those of us who are committed to *Direct Reference*, the conclusion that (1) and (2) express contradictory objects of belief seems unavoidable, however finely the object of belief is individuated; and once this premise is granted, the argument can go through equally well.

From my point of view, the problem is not caused by SEMANTICS, which I accept as an implication of the doctrine of Direct Reference. The problem must lie either with COMPETENCE or DISQUOTATION; and which of these theses should go will depend on which is the conception of *linguistic competence* that one operates with. Once we settle on one such conception, the interpretivist framework presented in §2 will give us a reason to reject either COMPETENCE or DISQUOTATION.

[14]One source of this argument is to be found in Kripke (1976). It is worth noting, however, that Kripke does not conclude that agents like Luthor are incoherent; rather, he regards that conclusion as paradoxical.

I prefer to say that Luthor is a competent speaker of English, and that therefore he is a counterexample to DISQUOTATION. The reason for this is that there is a natural conception of linguistic competence which implies that Luthor *is* a competent speaker of English. This conception begins by drawing our attention to a distinction between two kinds of expressions. On the one hand, there are expressions whose contribution to the proposition expressed is determined by *convention*. A paradigm of this is surely the conjunction 'and,' whose meaning does not seem to depend on any features of the context in which it appears. On the other hand, there are expressions whose contribution to the proposition expressed is determined, at least in part, by *contingent, non-conventional facts*. A paradigm of this kind of expressions are demonstratives. Though, in all likelihood, there is a part of the meaning of demonstratives that is conventional, everyone can agree, I think, that the denotation of an utterance of a demonstrative ultimately depends on the contingent circumstances surrounding the utterance of the demonstrative. With this distinction at hand, we can formulate the preferred conception of linguistic competence as follows:

> LINGUISTIC COMPETENCE (I): Competent speakers of a language *L* must know:
> (i) For expressions whose semantic value is determined in virtue of conventional facts, their semantic value
> (ii) For expressions whose semantic value is determined, at least in part, in virtue of non-conventional, contingent facts, the semantic rule that determines the contribution of the expression to the proposition expressed, relative to a context of utterance

It does not seem controversial that proper names fall on the non-conventional side of this distinction. Fregeans and Kripkeans alike can agree that the reference of a name is determined, at least in part, in virtue of contingent, non-conventional facts. For example, take the name 'Hesperus.' If Frege is right, the referent of this name is the object that satisfies the description 'The evening star,' and this is clearly a contingent fact. On the other hand, if Kripke is right, then the referent of 'Hesperus' is the object that is at the end of a certain causal chain connecting uses of 'Hesperus' with a baptismal ceremony in which the name was first introduced. Again, these are clearly contingent, non-conventional facts.

If this is right, then our interpretivist framework implies that competent speakers of English do not need to know the reference of all names in English. Luthor is a competent speaker of English, in spite of his mistake about 'Superman' and 'Clark Kent.' On this view, DISQUOTATION fails because it does

not take into account that competence with a language does not guarantee that a speaker will know the denotation of every proper name in her language. From this point of view, the intuition behind the principle of DISQUOTATION can be more plausibly expressed as follows:

> REVISED DISQUOTATION: Suppose English sentence *S* expresses proposition *P* with respect to context C; and that *A* is a competent speaker of English who in *C* is speaking sincerely, *and that moreover knows all the relevant contingent facts that determine the proposition expressed by* S *in* C. Then:
>
> (i) If *A* assents to *S* in *C*, then *A* is in the belief relation to *P*
>
> (ii) If *A* dissents from *S* in *C*, then *A* is not in the belief relation to *P*

Luthor is not a counterexample to REVISED DISQUOTATION, because he is not aware of all the contingent facts that determine the reference of 'Superman' and 'Clark Kent' in the context in which he asserts (1) and (2).

It would be possible, however, to argue that the right answer is the opposite one: That Luthor, and people like him, are not linguistically competent, and that the original formulation of DISQUOTATION is right. If one holds this view, then one should have an alternative to the notion of linguistic competence described above. One way of drawing this alternative is to appeal to the difference between *context-independent* expressions, whose contribution to the proposition expressed is *independent* of the context in which they are used, and *context-dependent* expressions, whose contribution to the proposition can vary with the context of utterance. This distinction is a familiar one; the claim then would be[15]:

> LINGUISTIC COMPETENCE (II): Competent speakers of a language *L* must know:
>
> (i) For context-independent expressions, their context-independent semantic value
>
> (ii) For context-dependent expressions, the semantic rule that determines the contribution of the expression to the proposition expressed, relative to a context

It seems clear that proper names fall in the context-independent category, because their denotation does not change from context of utterance to context of utterance. Therefore, if this account of linguistic competence is right, then the interpretivist framework presented in §2 gives us a reason to think that Luthor is not

[15]This view is found, for example, in Larson and Segal (1995), esp. chapter 2.

a competent speaker of English, since he does not know the referents of the names 'Superman' and 'Clark Kent.'

Now, this point of view is coherent, but I do not think that it is attractive: This conception of linguistic competence implies that everyone who is mistaken about the reference of a proper name is automatically a non-competent speaker of her language. The problem is that, in all likelihood, most of us are ignorant or mistaken about the reference of some proper name or other, and therefore would turn out to be non-competent speakers of our language, according to this definition. This seems to me quite implausible. Other things being equal, I would prefer a conception of linguistic competence that does not imply that most of us are non-competent speakers. Our first definition of linguistic competence is free from such implausible consequences, and is surely to be preferred.

However you define the notion of linguistic competence, the conclusion is that the argument from linguistic competence fails in any case, and fails because the interpretivist framework that we have been assuming implies that COMPETENCE and DISQUOTATION cannot be held at the same time.

1.7 The Argument from Perception

Probably most philosophers who entertain the idea that when Luthor asserts (1) and (2):

(1) Superman can fly
(2) Clark Kent cannot fly,

he expresses contradictory beliefs, do so because they are attracted to some version of a *causal theory* of belief content. If there is one lesson that has emerged from the philosophies of mind and language in the last half of the twentieth century is that causal considerations are important to content. And one may feel tempted to think that our theory is wrong, because it does not pay enough attention to causal considerations.

One way in which one could use causal considerations to argue against our theory is by relying on the relation between perception and belief. Ordinarily, one would expect that a perception of an object O would give rise to a *de re* belief about O, but our theory seems to be at odds with this principle. To make this point vivid, consider the following variation on the case of Luthor:

LUTHOR AND SUPERMAN (Continued): Suppose that all the facts are as explained in the story LUTHOR AND SUPERMAN, except for the following.

First, one day Luthor is sitting in his room, and all of a sudden he sees Superman flying off to rescue someone in danger. Luthor sees all this, and as a result he says:

(1) Superman can fly

Second, suppose that Luthor visits the *Daily Planet*, and that there he sees the reporter Clark Kent, who is on his way to Perry White's office. In the way, Clark Kent characteristically slips on the wet floor, and falls to the ground. With a chuckle, Luthor says:

(2) Clark Kent cannot fly

It seems clear that each of Luthor's utterances expresses a belief that is at least partly caused by the perception he was having moments before, and it also seems clear that those perceptions were caused by Superman himself. Since the content of the beliefs that Luthor gains each time is determined by the perception that, each time, gave rise to that belief, and those perceptions were caused by Superman himself, the conclusion that both of those beliefs are about Superman seems hard to avoid. But then, we will be committed to the claim that when Luthor asserts (1) he expresses the belief that Superman can fly, and when he asserts (2) he expresses the belief that Superman cannot fly, which are contradictory. According to our theory, Luthor does not have contradictory beliefs, so that the causal theory of content seems to be in tension with our treatment of cases like Luthor[16].

However, I do not think that this argument is convincing. To begin with, notice that the argument relies on a view about how belief content is determined by the causes of the belief that perhaps could be put as follows:

> PERCEPTUAL CONTENT: If an agent has a perception P of an object O and P is directly caused by O, then the agent will gain as a result the *de re* belief about O

The problem is that this causal account of perceptual content is too crude to be plausible. Everyone who has ever endorsed a causal account of mental content (or perceptual content, for that matter) has been quick to emphasize that the causal condition needs to be qualified, in order to meet certain well-known counterexamples[17]. There are of course many proposals about what the required qualification is. One

[16] I am not sure that anyone has explicitly committed to this argument, though, in conversation, Nathan Salmon seemed to be attracted to it.

[17] For example, see the initial discussion in Goldman (1978).

strategy that many philosophers endorse begins by amending the principle of PERCEPTUAL CONTENT so that it says something like the following:

> If an agent has a perception P of an object O, P is directly caused by O, *and the conditions under which* P *takes place are normal*, then the agent will gain as a result a *de re* belief about O

The remaining task is then to spell out what it means for conditions to be normal.

Ideally, we should expect that a solution to the problem of describing the belief state of people suffering from recognition failure should clarify this question. I do not think that the claim that the object of belief is finely grained, by itself, can do this, since the question of how finely individuated the object of belief should be is logically independent of the question of what are the conditions under which perception of an object O would normally cause a *de re* belief about O.

On the other hand, an attractive feature of the interpretivist framework defended in this chapter is that it suggests some constraints on when the causal conditions in which a *de re* belief is formed are normal. In particular, our treatment of the cases of recognition failure suggests two such constraints, having to do with the history of the agent undergoing the perception.

The first constraint is suggested by our treatment of Donnellan's Aston-Martin example. We proposed that the normal case in which a person gets the belief that a proper name N refers to O is if the agent belongs to a community which, as a matter of fact, takes N to refer to O. The Donnellan case suggested that this normal mechanism may malfunction in case of ambiguous proper names: If the agent picks up what as a matter of fact is an ambiguous name, and the agent fails to realize that the name is ambiguous, the QRI will have good reason to describe the agent as having the wrong beliefs about the referent of the name.

The analogy with the case of perception is that, though in general the normal mechanism for an agent A to acquire a *de re* belief about O is by having a perception P directly caused by O, this normal mechanism may malfunction if A has had many perceptions qualitatively similar to P in the relevant respects, but directly caused by objects different from O.

Here is an illustration. Suppose that Jones is at a party in which there are several men with the same hair color, who are of the same height, and who wear exactly the same kind of electric blue jacket. Jones

gets to see all those people from the back only, so that he cannot tell them apart. Further, Jones never gets to see more than one of them at the same time, so that he believes that there is just one such man at the party. (He believes, say, that electric blue jackets are unfashionable, and that it is already pretty unlikely that there will be one person at the party wearing one, and even more unlikely that there is more than one such person.) In this case, Jones has several perceptions of men wearing blue jackets, all of them are qualitatively undistinguishable in the relevant respects, and they are all caused by different people. My point is that, in this situation, any of Jones' perceptions of one of those men may fail (and, in all likelihood, will fail) to elicit a *de re* belief about the person causing the perception in question, because Jones is not in a position to tell that each of the perceptions is caused by a different person.

The second constraint is suggested by our treatment of Frege cases, and speaks more directly to the argument summarized above. Frege cases suggest yet another way in which the normal mechanism to acquire the correct beliefs about the reference of proper names may malfunction, this time in connection with pairs of coreferential proper names: If an agent picks up a pair of names that as a matter of fact are coreferential, but is not in a position to realize that the names are in fact coreferential, then the QRI will have good reason to describe the agent as having the wrong beliefs about the reference of at least one of the names.

The analogy is that, though in general the normal mechanism for an agent A to acquire a *de re* belief about O is by having a perception P directly caused by O, this normal mechanism may malfunction if the agent undergoes two different perceptions P and P', each of which was directly caused by O, but A was not in a position to tell that P and P' were caused by the same object.

The story of SUPERMAN AND LUTHOR summarized above illustrates that point. In that example, Luthor has two perceptions of Superman, one of Superman dressed in his superhero outfit, flying, and another of Superman dressed in his reporter garb, slipping on the floor. My point is that, though both perceptions are caused by Superman, one of them will fail to elicit a *de re* belief about Superman because Luthor cannot tell that both perceptions were caused by Superman.

The following summarizes our two proposals about when perceptual conditions are normal for the formation of *de re* belief:

PERCEPTION AND NORMAL CONDITIONS: A situation in which an agent has a perception P directly caused by an object O is *not normal* if (i) or (ii):

(i) The agent has had other perceptions qualitatively similar to P in the relevant respects, but caused by objects different from O

(ii) The agent has had other perceptions qualitatively different from P which have been directly caused by O, but the agent was not in a position to tell that those other perceptions and P were all caused by the same object

From this point of view, the moral that friends of the causal theory of content should extract from cases of recognition failure is that, in order for a perception P of an object O to elicit in an agent A a *de re* belief about O, the *history* of the agent has to be just right: Too many perceptions similar to P but caused by other objects, or too many perceptions different from P but caused by O, and the perception may fail to elicit the corresponding *de re* belief.

1.8 The Argument Against Propositions

The preceding sections are, in part, an argument to show that, when Luthor asserts 'Superman can fly' and 'Clark Kent cannot fly,' he expresses a pair of consistent beliefs. To this argument, it does not matter very much whether belief is a relation to a proposition, a mode of presentation, or a sentence of the language of thought. What matters, for this part of the argument, is that Luthor expresses a pair of consistent beliefs, whatever the object of belief turns out to be.

But I also want to defend a different thesis, a thesis of which I have given a few hints, and that is the thesis that belief is a relation between a person and a proposition. I suspect that many philosophers are willing to accept that Frege cases show that the traditional view that belief is a relation between a person and a proposition is false. I am going to argue that, to the contrary, Frege cases do not show that the traditional view is mistaken. I do think that Frege cases present a challenge, but a challenge that can be met, once it is properly identified. The interpretivist framework that I have defended here helps us to identify the precise nature of this challenge.

To begin with, the preceding considerations allow us to distinguish two different lines of argument that one may take against the traditional view. In the first place, there is what I call the *Irrationality* argument. This line of argument purports to show that the traditional view is committed to describing people like

Charles or Luthor as believing inconsistent propositions, and that this flies in the face of the observation that Charles and Luthor are coherent agents. I do not think that this version of the argument can succeed. In the preceding two sections I have discussed two versions of this argument, and I have deflected them by appealing to the interpretivist account of belief presented earlier.

A different line of argument, which I find much more challenging, is what I call the *Scarcity* argument. This argument purports to show that there are *not enough* propositions to describe the belief state of people suffering from recognition failure. Here the crucial example is the one of HESPERUS AND PHOSPHORUS, discussed in section §1.4.2. As we saw in that case, to describe the belief state of Charles we need to distinguish between the belief that he expresses when he says 'Hesperus is visible,' and the one he expresses when he says 'Phosphorus is visible.' But we seem to have only one proposition to play both roles, the proposition that Venus is visible. Thus, in this example, we seem to run out of propositions. The preceding considerations allow us to make three points in response to this argument.

In the first place, it is normally assumed that all -+Frege cases are counterexamples to the traditional view, but our discussion of what I called *Mixed Cases* shows that this is not true. In our example of LUTHOR AND SUPERMAN, we argued that when Luthor asserts 'Superman can fly,' he expresses the proposition that Perry White can fly, and that when he asserts 'Clark Kent cannot fly,' he expresses the proposition that Superman cannot fly, and there is no reason why there should not be enough propositions to describe that. If there is a challenge to the traditional view, the challenge comes from those Frege cases that are not Donnellan cases, like the case of HESPERUS AND PHOSPHORUS.

In the second place, it is worth noting that the preceding actually shows that there has been a *reduction* in the number of challenges that the traditional theory has to face. Most philosophers who oppose the traditional view emphasize the difficulties that Frege cases raise for it, but a few others also draw attention to the fact that the traditional view has trouble accounting for belief in fictional entities[18]. One could then infer that the traditional view faces at least *two* different challenges. However, the considerations presented in this chapter show that the problem of accounting for belief in fictional entities, and the problem of explaining what people like Charles believe, are one and the same, at least from the point of view of the QRI.

[18]See for example Braun (1993).

Thus the traditional view does not have to face two challenges, but just one.

The second point suggests a *new* way in which the friend of the traditional view might try to meet the challenge raised by the problematic Frege cases, and that is by trying to make sense of the idea that people like Charles believe in propositions that are about fictional entities. If we could show that this idea could be made to work, then the traditional view of belief would have an answer to the challenge raised by cases like HESPERUS AND PHOSPHORUS.

The third and final consideration is that the question of whether the traditional view of belief can make sense of the idea of propositions that are about fictional entities has not been discussed in sufficient depth, in particular by those interested in the phenomenon of recognition failure. And I think that here there is an interesting battle to be fought. In the next chapter I am going to present an account of belief about fictional entities which, I think, can help the propositionalist meet the challenge raised by cases like HESPERUS AND PHOSPHORUS.

Chapter 2

Belief About Nothing

2.1 Propositions and Belief About Nothing

A singular proposition is a proposition of which it can be said that it is about an object. As one philosopher has put it, singular propositions are object-dependent: The very existence of the proposition depends on the existence of the object that the proposition is about. Therefore, if the object *O* does not exist, there is no such thing as a singular proposition about *O*[1].

Typically, singular propositions are the propositions that people express when they assert a sentence containing a proper name. This beautiful correlation between singular propositions and sentences with proper names breaks down, however, when it comes to sentences that contain *empty* names, names that do not have a referent. Take for instance (1):

(1) Santa Claus does not exist

(1) is surely meaningful, which means that there must be some proposition that it conveys. But it cannot be the singular proposition that *Santa Claus* does not exist, for there is no Santa Claus, and therefore there is no such proposition.

Keith Donnellan explained how to get over this difficulty. According to the so-called *Causal Theory*

[1]See Neale (1990), chapter 2.

of naming, an object is the referent of a proper name if, and only if, the object has received the name in an appropriate baptismal ceremony and the current uses of the name in question are connected to the previous uses leading up to the baptism by means of an appropriate causal chain. If the causal chain determining the reference of a proper name does not end in a baptismal ceremony, it is said to be *blocked*. Donnellan then proposed that the proposition expressed by (1) is more or less equivalent to the claim that in the causal chain determining the reference of the name 'Santa Claus' there is a block[2].

Donnellan's proposal is certainly attractive, but it is not general enough. For example, take the case of Johnny. Johnny is a kid who, like most kids, believes that Santa Claus will come on Christmas' Eve to bring him gifts. Among other things, Johnny is inclined to assent to (2):

(2) Santa Claus will bring videogames

The traditional view says that, when Johnny asserts (2), he expresses a belief, and the content of that belief can be captured by means of a proposition. Well, which is the proposition that Johnny believes? We already know that it cannot be the singular proposition that Santa Claus will bring videogames, and it is not clear how Donnellan's strategy could help here. Since Johnny believes in Santa Claus, the proposition that gives the content of his belief cannot imply that the name 'Santa Claus' does not refer, for Johnny does believe that the name refers.

Philosophers committed to the Russellian view of propositions, on which propositions are understood as sequences of objects, have proposed that cases like Johnny could be analyzed by appeal to "gappy" propositions, propositions that miss an element. On the Russellian view, one expects simple propositions to be made up of an object and a property; on this proposal, the proposition that Santa Claus will bring videogames would represented as having a hole in the object position:

$$<<\ >,\ <\textit{the property of bringing videogames}>>$$

Because Santa Claus does not exist, there is nothing to play the role of the object of this proposition.

The problem with this proposal is that it fails to make some intuitive distinctions between believers. For example, take the case of Carlitos, a Spanish child who, like most Spanish children, believes that on the

[2]For more on Donnellan's account, see his classic Donnellan (1974). See also Stalnaker (1978), which explains how to derive compositionally this proposition from the assertion of (1).

night from January 5 to January 6, Melchor, Gaspar and Baltasar, the three Kings of Orient, come to bring gifts to all (well-behaved) children. Carlitos fancies that Baltasar, in particular, was the one who read his letter, and that he will take special care to bring him what he asked for—namely videogames. So he says (3):

(3) Baltasar will bring videogames

Presumably, the belief that Carlitos expresses when he says (3) is different from the belief that Johnny expresses when he says (2). The problem is that the proposal that there are gappy propositions cannot account for this difference: Since Baltasar too is a fictional character, the proposal is committed to saying that the proposition Carlitos is expressed is:

$$<<\ >,\ <\textit{the property of bringing videogames}>>$$

which would be the same that Johnny expresses when he says (2). This is unacceptable, since Johnny and Carlitos quite obviously express different beliefs each time[3].

So, to represent the beliefs that people like Johnny and Carlitos express by asserting sentences containing empty names, we cannot appeal to singular propositions, or to propositions to the effect that the causal chain determining the reference of the name is blocked, or to gappy propositions. A natural thought at this point is to revisit Frege and Russell's idea that sentences containing proper names express *general* propositions. General propositions are those that merely describe objects, and do not depend on any particular object for their existence. Let us take a look at it.

2.2 Frege and Russell's Song of the Syren

2.2.1 Descriptivism and Empty Names

Frege and Russell held a doctrine according to which proper names contribute something like descriptive senses to the proposition expressed[4]. For the reasons that we will shortly come to, the view of Frege and

[3]David Braun has suggested the idea of gappy propositions in Braun (1993), though it is only fair to say that he does not take it too seriously.

[4]See Frege (1892), Frege (1918), Russell (1904).

Russell cannot be maintained. But we nevertheless can get an inspiration from Frege and Russell to solve our current difficulties. The proposal is the following:

> DESCRIPTIVISM ABOUT EMPTY NAMES: When an agent A asserts a sentence $\ulcorner N \text{ is } F \urcorner$, for some empty proper name N, he expresses a belief whose content is the general proposition that D is G, for some definite description D, and some property G.

This doctrine yields some attractive results with respect to the cases of Johnny and Carlitos. For example, take the case of Johnny. We could now say that, when Johnny asserts (2), he expresses a general proposition that is equivalent to (4):

(4) The red-suited, white bearded man who rides a flying sleigh will bring videogames

Because the existence of general propositions does not depend on the existence of any particular object, there is no reason why there should not be a proposition like the one expressed by (4). And there is more good news: We can say that Johnny believes (4) without risking committing ourselves to the existence of Santa Claus, since in (4) the name 'Santa Claus' is not used. True, in (4) there is a description of Santa Claus, but, as Russell explained to us, we all know that using a description does not commit us to the existence of something that satisfies it[5].

Further, this doctrine can also distinguish what Johnny believes from what Carlitos believes. On their view, the proposition that Carlitos expresses when he asserts (3) is equivalent to (5):

(5) The black king of Orient will bring videogames,

Since (4) and (5) are different propositions, this proposal would manage to distinguish the propositions that Johnny and Carlitos express by means of (2) and (3).

2.2.2 The Modal Argument and Empty Names

There is another consideration in favor of DESCRIPTIVISM ABOUT EMPTY NAMES. As is well-known, Saul Kripke and Keith Donnellan presented very convincing arguments against the general Frege-Russell treat-

[5]See Russell (1904).

ment of proper names[6]. One of these arguments exploited the fact that the Frege-Russell view, applied to proper names in general, gets wrong the truth value of certain modal sentences. For example, suppose that one were to say that the name 'Aristotle' contributes to truth conditions something like the description 'The teacher of Alexander.' Now consider these sentences:

(6) Aristotle could have failed to be Aristotle

(7) Aristotle could have failed to be the teacher of Alexander

Presumably, Alexander could have had some teacher other than Aristotle (Plato, for example), so that (7) is true. On the other hand, Aristotle could not have failed to be himself, so that (6) is false. The problem for Frege and Russell is that, if 'Aristotle' contributes to the truth conditions of (6) and (7) something like the description 'The teacher of Alexander,' then (6) and (7) should express exactly the same proposition and cannot differ in truth value. But they do differ in truth value, and so the Frege-Russell treatment of 'Aristotle' fails—and it is easy to see that the same will be true for any other choice of sense for the name 'Aristotle,' and also for many other proper names.

But it is not clear that the same argument works against the Frege-Russell treatment of *empty* names. For consider:

(8) Santa Claus could have failed to be Santa Claus

(9) Santa Claus could have failed to be the red-suited, white bearded man who rides a flying sleigh

(8) and (9) are either both false or both truthvalueless, on account of the fact that the name 'Santa Claus' is empty. If this is right, this would show that Kripke's modal argument would not work against the descriptivist treatment of empty names. And it suggests that perhaps the Frege-Russell view of proper names is tenable, if restricted to empty names.

Given that the Frege-Russell treatment of empty names has so many advantages, and that it seems immune to Kripke's modal argument, why not going the descriptivist way? Why not say that the beliefs Johnny and Carlitos express by means of (2) and (3) are general propositions?

[6]For which, see Kripke (1980) and Donnellan (1970).

2.2.3 A Problem For Descriptivism About Empty Names

There nevertheless is a problem for the descriptivist treatment of empty names, a problem that is not entirely unrelated to Kripke's modal argument. The argument relies on the possibility that an agent accept (9) but reject (8). We have assumed that Johnny is a kid, and as the kid he is, he probably does not understand modal notions enough to understand either of (8) or (9). But there is no reason why someone could not grow up to understand modal notions before realizing that Santa Claus does not exist. Call such a person *Big John*. After studying modal notions, Big John could go through the following train of reasoning:

> Suppose one day Santa Claus gets injured —that is very likely, what with his work requiring him to walk on so many old roofs, and to go down so many old chimneys. Who could replace him? Maybe one of the three Kings of Orient could. They already have the practice, after delivering so many gifts to so many kids every January 6th. Of course, things would have to be arranged so that he looks like Santa Claus: It would not do to have someone deliver gifts on Christmas Eve without the white beard, or the red suit. If that ever happens, curious as it may be, it would turn out that the red-suited, white-bearded man who rides a flying sleigh would not be Santa Claus, but someone else.

After which it would be very natural for Big John to accept (9) as true, while rejecting (8).

If Big John accepts (9) but rejects (8), it must be because he believes that each of those sentences expresses different propositions. The problem is that descriptivism cannot distinguish those propositions. Suppose that we say that Big John associates one and only one sense with 'Santa Claus,' namely the description 'The red-suited, white bearded man who rides a flying sleigh.' That is itself very plausible, but then, descriptivism would have the consequence that the propositions that Big John associates with (8) and (9) are the same. We could try to avoid the argument by saying that Big John associates some other description with 'Santa Claus,' but this move would be futile, for then the argument could by substituting that other description for 'The red-suited, white bearded man who rides a flying sleigh,' and imagining a context similar to the one we have described.

This shows that general propositions will not do either as the content of the beliefs expressed by agents who assert sentences that contain empty names, and now we seem to be running out of candidates. Some philosophers have drawn the conclusion that the traditional view that belief is a relation to a proposition is falsified by cases like Johnny, Carlitos and Big John. On their view, cases like these ones show that we

should say instead that belief is a relation to a mode of presentation, or perhaps to a sentence of the language of thought[7]. But before we give up hope, there is one last chance of rescuing the traditional view.

2.3 Walton's Prop-Oriented Make-Believe

2.3.1 Commitment and Fiction

In the views so far examined, there are two features that I find valuable. On the one hand, there is the idea that, when a subject asserts a sentence that contains a proper name, she thereby expresses a belief whose content is a proposition that is about an object, rather than a general proposition. This seems to capture better the way competent speakers use sentences with proper names. On the other hand, there is the descriptivist suggestion to analyze the proposition expressed by sentences containing proper names in a way that does not commit us to the existence of a referent for the name. If we could somehow combine these two features into a single analysis, so that we can talk about propositions that are about objects that do not exist, without thereby becoming committed to the existence of the objects that those propositions are about, that would be progress. But how is this to be done?

The prefix 'it is fictionally true that' has several virtues that, at this point, recommend it to our attention. Consider, to begin with, that we can say sentences like the following:

(10) It is fictionally true that Superman can fly

(11) It is fictionally true that Santa Claus rides a flying sleigh

(12) It is fictionally true that Sherlock Holmes wore a deer-hunting cap

and say a truth each time, even if there is nothing that the names 'Superman,' 'Santa Claus,' and 'Sherlock Holmes' refer to. Thus, the prefix 'It is fictionally true that' works so that a sentence *S* prefixed by it can be true, even if *S* contains names that do not refer.

Another interesting feature is that the prefix 'It is fictionally true that' in effect *cancels* the ontological commitments of the complement sentence. This is so because everyone can agree, I think, I think, that

[7]See Braun (1993), for an example of this attitude.

we can assert (10–12) without committing ourselves to the existence of Superman, Santa Claus or Sherlock Holmes.

I think that the prefix 'It is fictionally true that' is what we need to solve our problems. By using this prefix, we can, for example, *talk* about propositions that are about fictional objects without thereby committing ourselves to the existence of those objects. For example, we can truly say the following:

(13) It is fictionally true that Santa Claus will bring videogames

And when we say (13), it seems for all the world as if we are saying something about the proposition that Santa Claus will bring videogames, without committing ourselves to the existence of Santa Claus.

If this idea is to be put to work, there are two directions in which it needs to be developed. In the first place, we need an explanation of exactly how it is that sentences of the form:

It is fictionally true that N is F

can ever get to be true, when the name N does not refer. There are two reasons for wanting this. One reason is that one would like the semantics of the prefix clarified, if it is going to be put to some substantive philosophical use. Another is that, in any case, we need to make sure that the truth sentences of this form does not commit us to the existence of objects that do not exist, and the way to make sure of this is by having an idea of what the semantics of the prefix looks like.

In the second place, it must be explained how the prefix 'It is fictionally true' can be used to say what we want to say, when we describe the belief state of people like Johnny. It is not obvious that it can; think for a moment what we would do with it. For example, suppose that we use it to describe Johnny's belief state in the following way:

(14) It is fictionally true that Johnny believes that Santa Claus will bring videogames

(14) is not satisfactory: Though (14) does not imply that Santa Claus exists, it does not imply that Johnny exists either! Surely this is not right, since we want to describe Johnny's belief state in a way that our audience understands that we are committed to the existence of Johnny and to his having a certain belief.

Another option would be to use the prefix to qualify *just* the proposition believed by Johnny:

(15) Johnny believes that it is fictionally true that Santa Claus will bring videogames

However, (15) does not imply that Johnny believes that Santa Claus is a real person. The problem is that Johnny does believe that Santa Claus is a real person, and we want to be able to say this in our description of Johnny.

Our task can be summarized in the following way. We have a prefix, 'It is fictionally true that,' which seems to allow us to say true things, even when its complement sentence contains names that do not refer. How exactly does this prefix work? And how can it be used to describe the belief state of people like Johnny, who are real people, and believe that Santa Claus is a real person?

To solve these two problems, I will assume an analysis of the prefix inspired in Kendall Walton's work, with a particular emphasis on the notion of *Prop-Oriented* make-believe. For the sake of convenience, I will break the explanation into three blocks[8].

2.3.2 Make-Believe

First, I will assume an analysis of the prefix 'It is fictionally true that' in terms of make-believe. The notion of make-believe that I have in mind is the ordinary notion of make-believe, on which make-believe is what children engage in in many of their games, and what grown-up people do when they read a novel, watch a theater play, or let their imagination fly free.

Make-believe is a species of a positive attitude, similar to belief in some respects. Like belief, make-believe can guide behavior: In some cases, our doing a certain action A can be explained by the claim that, at the time, we were making-believe that P, for some P. But there are also certain crucial differences: While belief tends towards the truth, make-believe does not so tend. While we would immediately abandon the belief that P upon learning that P is false, there is no reason why we should cease to make-believe that P, upon learning that P is false. These differences spring from the different *purposes* that belief and make-believe serve. While the proper purpose of belief is to guide behavior, the proper purpose of make-believe is not necessarily to guide behavior. Rather, make-believe serves a variety of purposes: Children engage in

[8]The following elaboration of Walton's views is based on Walton (1973) and Walton (1993).

make-believe because it is fun; grown-up people engage in make-believe to enjoy a novel, or a theater play, or perhaps even to figure out a way of testing some scientific hypothesis.

My proposal is to use the notion of make-believe, as here explicated, in the analysis of the prefix 'It is fictionally true that' in the following way:

> For any sentence *S*, assertions of ⌜It is fictionally true that *S*⌝ are true if, and only if, there is a relevant game of make-believe in which it is appropriate to make-believe that *S* is true

2.3.3 Principles of Generation

It is not true that, when we engage in a make-believe, everything goes. Walton has emphasized the point that, when we engage in a make-believe, there are certain rules that determine what is true and what is false within the make-believe. For example, if we make-believe that you are Holmes and I am Watson, then the parameters of our make-believe will be determined, to the most part, by the Sherlock Holmes novels. This means, for example, that I would not be playing the make-believe well if I were to solve the case (since I am making believe that I am Watson, and Watson never solves a case), and that you would not be playing the make-believe well, if you were to play the flute when you concentrate (since Holmes is known to play the violin, not the flute).

Walton has called the rules that determine what is true and what is false in a make-believe the *Principles of Generation* of the make-believe. Depending on how the make-believe is set-up, and our interest in playing it, the Principles of Generation may vary along different dimensions. One interesting dimension is the amount of creativity that they allow to the players. In the case of the make-believe in which you are Holmes and I am Watson, the rules do not allow for much freedom, for we are supposed to reenact some of the episodes of the Sherlock Holmes novels.

Other games of make-believe are not like that. Take, for example, Walton's classic example of the game of mud-pies[9]. In this game, children make-believe that certain globs of mud are pies. One of the principles governing the game could be formulated as follows[10]:

[9] First presented in Walton (1973), esp. pp. 287 and ff.
[10] The formulation is from Evans (1982b), p. 354.

MUD PIES: For all x, if x is a glob of mud, and is fashioned into pie-shape, then it will be true, within the make-believe, that x is a pie

This principle, by itself, does not determine how many pies there will be made-believe. That will be up to the participants, by creating globs of mud. Thus, this is an example of a game of make-believe in which the principles of generation, by themselves, do not settle what is true within the make-believe. Rather, that is settled by the principles, together with the conditions in which the game is played (some of which can be manipulated by the participants).

2.3.4 Prop-Oriented Make-Believe

Walton has drawn attention to a particular kind of make-believe that he calls *Prop-Oriented* make-believe. The feature that distinguishes this kind of make-believe is that it is set up so that certain features of the actual world play a role within the make-believe. To illustrate this, take again the case of the game of mud pies. In that game, the globs of mud, which are real objects, are used as *props* that play the role of pies within the make-believe.

The interesting thing about prop-oriented make-believe is that, if we engage in it, then we can communicate information about the real world by making assertions within the make-believe game. Suppose, for example, that we are playing the make-believe game, and that someone wants to know where Susie is. Susie is at the far corner of the playground, next to a glob of mud. If it is common knowledge between the participants in the conversation that we are playing the game of mud pies, then I can tell where Susie really is by saying:

(16) Susie is next to that pie in the far corner

Literally taken, my assertion is not true. But it can communicate to someone information about where Susie really is, because, given the way in which we have set up the make-believe, fictional truths about pies are correlated with real truths about globs of mud. Thus, my audience can hear (16), and infer from it that Susie is next to that glob of mud in the far corner.

How much information about the real world will be carried by assertions within the make-believe will depend on how the make-believe is set up. As the game of mud pies is set up, make-believe assertions

about pies carry information about real globs of mud. There may be other features of the make-believe that do not correspond to anything in the environment in which the game is played. For example, we can make-believe that certain pies are covered with chocolate sprinkles, and that others are covered with frosting, even if no actual physical property of the globs of mud is correlated with that. In general, to determine how much information is carried about the real world is carried by an assertion made within the make-believe, we will have to look at the Principles of Generation of the make-believe in question.

We now have all the elements in place to see how this account of the prefix 'It is fictionally true that' can help us with our problems.

2.4 Applications

2.4.1 Johnny and Carlitos

The idea is that, to describe the belief state of Johnny, we will engage in a prop-oriented make-believe inspired in Johnny's life. In particular, the principles of generation will be formulated so that there is a person called 'Johnny' within the make-believe, with more or less the same properties that the real Johnny has. This means, among other things, that if the real Johnny behaves in a certain way, then it will be true within the make-believe that the Johnny within the make-believe behaves in that way too.

For the most part, the make-believe will resemble the real world, though it will not be completely identical to it. The reason why we want to engage in make-believe is to be able to describe Johnny's belief in Santa Claus. Because of this, the principles of generation of the make-believe will stipulate that there is a person named 'Santa Claus' who has all the properties that Johnny attributes to him. This means, for example, that within the make-believe it is true that there is a red-suited, white-bearded man who rides a flying sleigh, who brings gifts for all well-behaved children on Christmas Eve, and whose name is 'Santa Claus.' Further, when, within the make-believe, Johnny goes to a department store and sees someone dressed in a red suit and wearing a white bear, it is true within the make-believe that the person he is seeing is Santa Claus.

Summarizing, the idea is that, to describe the belief state of people like Johnny, we must engage in a make-believe governed by the following principles of generation:

SANTA CLAUS:

(i) There is a person that is called 'Santa Claus'

(ii) The person called 'Santa Claus' has a property F if, and only if, the real person Johnny is willing to assert to $\ulcorner$Santa Claus is $G\urcorner$, for some predicate G that expresses F

(iii) For any sentence S, S is true within the make-believe if S is true in the actual world and is not inconsistent with any of the propositions that, according to principles (i) and (ii), are true within the make-believe

Let me now illustrate some of the virtues of this analysis. Perhaps it will be best to begin by showing that this framework implies that (17) is true:

(17) It is fictionally true that Johnny believes that Santa Claus will bring videogames

On the present account, (17) is true, even if the principles of generation of the make-believe, by themselves, do not imply that (17) is true. Rather, the truth of (17) is *generated* within the make-believe, so to speak. In particular, the principles of generation determine that, within the make-believe, Johnny will behave in a certain way, and will interact with certain objects. The interpretivist view of belief defended in chapter 1 will then determine a certain description of Johnny's belief state, within the make-believe. In all likelihood, the interpretivist framework will imply that, within the make-believe, Johnny believes that 'Santa Claus' refers to Santa Claus, that Santa Claus will bring videogames and so. The upshot is that, since it is stipulated that within the make-believe Johnny behaves as he does in the real world, and also that within the make-believe we can talk about all the objects we need to describe Johnny's belief state, the QRI would have all she needs to describe Johnny's belief state, within the make-believe.

In the second place, this proposal allows us to say that Johnny believes in Santa Claus without committing ourselves to the existence of Santa Claus. In particular, the truth of (17) does not imply that Santa Claus exists. Here the key is that the make-believe has been set up so that the truth of claims like 'Santa Claus exists' within the make-believe does not imply the truth of the proposition that Santa Claus exists, in the real world.

In the third place, the truth of (17) does imply that Johnny is a real person. Here it is crucial to realize that the make-believe, on which the truth conditions of (17) has been set up so that many of the actual facts about Johnny are also true within the make-believe. And so, in particular, the fictional truth that Johnny exists implies, given the way the make-believe has been set up, that Johnny is a real person.

In the fourth place, this proposal allows us to distinguish Johnny's beliefs from Carlitos' beliefs. Remember that Johnny has a belief which he expresses by saying:

(2) Santa Claus will bring videogames

And that Carlitos has a belief which he expresses by saying:

(3) Baltasar will bring videogames

To describe the beliefs of both, we will engage in a make-believe according to which there are two different people called 'Santa Claus' and 'Baltasar,' each of which has different properties. The principles of generation of such make-believe would be spelled out as follows:

> SANTA CLAUS AND BALTASAR:
> (i) There is a person called 'Santa Claus'
> (ii) There is a person called 'Baltasar'
> (iii) The person called 'Santa Claus' has a property F if, and only if, the real person Johnny is willing to assert to $\ulcorner$Santa Claus is $G\urcorner$, for some predicate G that expresses F
> (iv) The person called 'Baltasar' has a property F if, and only if, the real person Carlitos is willing to assert to $\ulcorner$Baltasar is $G\urcorner$, for some predicate G that expresses F
> (v) For any sentence S, S is true within the make-believe if S is true in the actual world and is not inconsistent with any of the propositions that, according to principles (i–v), are true within the make-believe

Given the way in which the make-believe is set up, it is clear that, within the make-believe, Santa Claus and Baltasar will have different properties, since Johnny attributes to Santa Claus properties which Carlitos does not attribute to Baltasar (for example, Carlitos attributes to Baltasar the property of having dark skin, while Johnny attributes to Santa Claus the property of having fair skin). We can then apply Leibniz's principle to derive the result that, within the make-believe, Santa Claus and Baltasar are different people; and the principle of CHARITY to determine which of those two different people are the ones which, within the make-believe, Johnny and Carlitos have beliefs about.

2.4.2 Hesperus and Phosphorus

My general proposal is that, whenever we are forced to describe the belief state of people who seem to believe in fictional entities, we will use Walton's notion of Prop-Oriented Make-Believe to do so. The recipe

to set up the appropriate make-believe has two steps. First, we assume a principle of generation to the effect that there is a fictional entity which, within the make-believe, has the properties which the subject attributes to it. Second, we fill out the rest of the details of the make-believe by stipulating that any proposition *P* that is true in the actual world would also be true in the make-believe, as long as that does not conflict with the make-believe truths determined by the first step[11].

The case of Charles, examined in chapter 1, §1.4.2, can be handled in this way. Remember that the problem was that, when we describe the belief of Charles, we run out of planets to characterize Charles' view of the world. For example, we did not have enough planets to characterize the two propositions that Charles takes himself to express when he asserts:

(18) Hesperus is nice

(19) Phosphorus is not nice

According to the present account, what we have to do is to engage in the appropriate kind of make-believe. A proposal about how the make-believe should be set up is this:

> HESPERUS AND PHOSPHORUS:
>
> (i) There is a planet called 'Phosphoria'
>
> (ii) The planet called 'Phosphoria' has a property *F* if, and only if, the real person Charles is willing to assert to ⌜Phosphorus is *G*⌝, for some predicate *G* that expresses *F*
>
> (iii) For any sentence *S*, *S* is true within the make-believe if *S* is true in the actual world and is not inconsistent with any of the propositions that, according to principles (i–iii), are true within the make-believe

In effect, these principles of generation determine a make-believe in which there is a real person called 'Charles,' and who is very similar, in most respects, to the real person Charles. Also, in that make-believe, there is a planet called 'Phosphoria,' which appears only in the morning; and another planet called 'Venus,' which appears only in the evening. All in all, it is true within the make-believe that there are ten different planets, and that Venus and Phosphoria are two of them.

[11] This recipe is reminiscent of a proposal by David Lewis about how to analyze the prefix 'It is fictionally true that.' According to Lewis, the truth of statements containing the prefix in terms of truth in a possible world in which the facts that are true are the facts described in the fiction, and the rest of the facts in the world are the facts that actually occur in the actual world. For details, see Lewis (1978), esp. pp. 268 and ff.

The idea is then that, when the QRI describes the belief state of Charles, she should engage in this sort of make-believe. She would then be able to describe the belief state of Charles by means of (20) and (21):

(20) It is fictionally true that Charles believes the proposition that Venus is nice

(21) It is fictionally true that Charles believes the proposition that Phosphoria is not nice

Given the way in which the make-believe has been set up, (20) and (21) have the following implications. In the first place, (20) implies that there is a real person called 'Charles,' that Charles has a belief that within the make-believe can be described by the proposition that Venus is nice, and that this proposition is a real one. On the other hand, (21) implies that there is a real person called 'Charles,' that Charles has a belief that within the make-believe can be described by the proposition that Phosphoria is not nice, and that this proposition is one that exists only within the make-believe.

2.4.3 The Scarcity Argument Revisited

At the end of the previous chapter we reviewed what I called the *Scarcity* argument against the traditional view that belief is a relation between a person and a proposition. The argument started with the premise that the traditional view does not have enough propositions to characterize the belief state of people like Charles: When Charles asserts 'Hesperus is nice' and 'Phosphorus is nice,' he expresses a pair of different beliefs; but it is not clear that there are enough propositions to characterize the difference between those beliefs. As we saw at one point, the QRI was stuck with only one kind of singular propositions, propositions about Venus, to characterize both of them. This is, of course, unacceptable. The Scarcity argument concluded that the traditional view does not have enough resources to describe the belief state of agents like Charles.

I grant the premise of the argument. On the present account, there is no Santa Claus, no Baltasar, no tenth planet 'Phosphoria,' and no fictional entities in general. Therefore, there are no singular propositions about Santa Claus, nor about Baltasar, nor about 'Phosphoria.' And it is true, there are not enough singular propositions to describe the belief state of people like Charles, or like Big John, for that matter.

What I reject is the inference from this premise to the claim that the traditional view of belief is powerless to describe the belief state of people like Charles. My point is that there need not be propositions, in order for us to use them when describing the belief state of other people. It is enough to engage in games of make-believe in which Santa Claus, Baltasar and Phosphoria exist. For as long as we engage in a make-believe of that sort, we will be able to talk about singular propositions about Santa Claus, Baltasar and Phosphoria. The tricky part in developing this proposal was to show how this notion of make-believe can be used to say what we want to say, when we describe the belief state of real people, like Johnny, who believe that Santa Claus is a real person; it was here that Walton's notion of Prop-Oriented Make-Believe was invaluable.

2.5 Externalism and Belief About Nothing

Externalism was proposed first by Hilary Putnam and Tyler Burge in a couple of immensely influential papers.[12] The doctrine arises from a series of well-known thought experiments, and one of them runs as follows[13]:

> TWIN WATER: Suppose Twin-Earth were a planet that is quite similar to ours, except for the fact that water does not exist there. Instead, the liquid that fills the lakes and oceans, flows in rivers, and falls from the sky in the form of rain, were instead a quite different stuff, whose chemical composition is a quite complicated formula that we will abbreviate as 'XYZ.' Suppose now that in Twin-Earth there is a physical duplicate of you: Someone who looks like you, acts in the way you do, grew up in the way you did, utters exactly the same words that you do, and is otherwise identical to you. Let us call that person your *doppelganger*. In these circumstances, suppose your doppelganger says:
>
> (22) There is water in the bathtub
>
> It seems clear that, when your doppelganger says (22), he says something about XYZ, and that he expresses a belief about XYZ.

One moral extracted from this thought experiment was purely negative. Let us say that *internalism* is the view that the content of your beliefs supervenes either on your physical constitution, or else on whatever properties supervene on your physical constitution. The thought experiment shows that internalism is false, for while you and your doppelganger are physically identical, you have beliefs about water, while your

[12]See Putnam (1975) and Burge (1979).

[13]The following is inspired in the thought experiment in Putnam (1975).

doppelganger has beliefs about XYZ. Therefore, it is not true that the content of your beliefs supervenes on your physical constitution, or on any set of properties supervening on those.

But the thought experiments do not support just this negative thesis. To be sure, they also suggest a positive thesis about the nature of thought, since the conclusion of the thought experiment is supposed to affect more beliefs than your beliefs about water, more people than yourself, and more situations than just the one described as TWIN WATER. It is natural to think that TWIN WATER, and other thought experiments like it, support something like the following thesis:

> STRONG EXTERNALISM: It is a necessary condition for an agent to have a belief about X that X exist in the agent's environment

In the thought experiment, whether you or your doppelganger have a belief about H_2O or XYZ seems to depend on what kinds of things exist in your environment. And wouldn't it then be natural to conclude that you can *only* have beliefs about objects and properties that exist in your environment?[14]

If STRONG EXTERNALISM were right, that would be bad news for the theory presented in the previous two sections. My theory allows that there is a sense in which we can say that people like Johnny believe in things that do not exist, as long as we say so while engaging in a make-believe. What we gain by engaging in make-believe is to gain access to propositions to which we would otherwise not have access to, and which are the ones necessary to describe the belief state of people like Johnny. But if STRONG EXTERNALISM is true, then there is no sense in which we can say that Johnny believes in things that do not exist. Indeed, if STRONG EXTERNALISM is true, then the presumption that this chapter is based upon, namely that there are people who have beliefs about things that do not exist, is simply false. The belief state of those people has to be described in some way, but not as beliefs about things that do not exist.

However, I think that this argument is too quick. In particular, I do not think that thought experiments like TWIN WATER really support STRONG EXTERNALISM. There are two things that call for attention in the

[14]It is not clear to me whether anyone has actually endorsed Strong Externalism. Paul Boghossian seems to me to come close to it when he says:

> ...[There is a commitment] to a relationist conception of content: the view that the content properties of mental states and events are determined by, or supervenient upon, their *relational* properties... (Boghossian (1989))

The claim that the content of a belief supervenes upon its relational properties straightforwardly suggests that we cannot have beliefs about Santa Claus: For, simply put, no one of us is related to Santa Claus, since there is nothing to be related to. However, compare with Boghossian (nd), especially the section titled 'The Empty Case.'

thought experiment. One is that, before determining what you and your doppelganger's beliefs are about, it is possible to say something about the kind of thing that must satisfy those beliefs. For example, both you and your doppelganger are inclined to assent to the sentence 'Water is the liquid that fills lakes, flows in rivers, and falls from clouds in droplets when it rains.' This, and other sentences like this, determine a certain *role* that the referent of 'water' has to play: It is the role of being that stuff which fills lakes, flows in rivers, and falls from clouds in droplets when it rains.

The other feature of the thought experiment is that, in the environment in which you and your doppelganger exist, there is exactly one kind of substance that plays the role of water. In your case, it is water, for water is the only liquid that fills lakes, flows in rivers, and falls from clouds in droplets when it rains; and in the case of your doppelganger, it is XYZ, for XYZ is the only liquid that fills lakes, flows in rivers, and falls from clouds in droplets when it rains.

Taking this into account, we can now argue that thought experiments like TWIN WATER do not support STRONG EXTERNALISM, but rather the truth of the following, weaker thesis:

> WEAK EXTERNALISM: If an agent has a series of beliefs according to which whatever satisfies a certain name N has certain properties $P_1, ..., P_n$ (for some $n > 0$), and if in the environment of the agent there is a unique object (or stuff) X that instantiates $P_1, ..., P_n$, then the beliefs the agent expresses by using the name N are beliefs about X

It is interesting to note that WEAK EXTERNALISM is compatible with the idea that there can be belief about objects that do not exist. In particular, whenever an agent has in her vocabulary a singular term such that there is no unique object in the subject's environment that satisfies the role associated to that name, WEAK EXTERNALISM does not imply anything about the beliefs that such an agent expresses by means of sentences containing the singular term in question.

Even if Twin Water does not support STRONG EXTERNALISM, the question now arises whether there could not be some version of the thought experiment in which the relevant conditions are not met —in which there is no unique object or property playing the relevant role. Would such a thought experiment support STRONG EXTERNALISM? Or would it tell against it? Let us take a look at it.

Twin Earth thought experiments in which there is no object or property playing the role of water are hard

to come by because it is hard to see how a doppelganger of us could exist in a place in which, by design, there is nothing that plays the water role (no water, no XYZ, nothing). Perhaps such examples are simply not possible; nevertheless, if we use another term, different from 'water,' one that refers to something that is less crucial to life, such examples are easier. Consider the following example:

> THE ILLUSION WORLD: Suppose the Illusion World is a planet exactly similar to ours, except that, in it, there is no gold, and there has never been gold. Some very clever aliens have come up with theories about the effect of gold on the human psyche which they would like to test, but unfortunately they have not come by enough gold in their interstellar travels to carry it out. So they resort to fooling the humans by a series of very clever illusions, both visual and tactile, destined to make humans believe that there is gold in their environment. In this world, there is no gold kept under guard in Fort Knox; instead, there is only an illusion created by the aliens. This would be amazingly complicated, but there is no reason to suppose that there could not be aliens smart enough to pull this off.

Does this thought experiment show that STRONG EXTERNALISM is false? On the face of it, there is a clear intuition that when your doppelganger says:

(23) There is gold in Fort Knox,

she is saying something false; and moreover, that the belief she is expressing is a belief about gold. (Notice that the situation of your doppelganger here is akin to those ancient physicists and chemists who believed in *phlogiston* or *caloric*, substances which turned out not to exist. It is very natural to describe their beliefs, and their corresponding assertions, as false too.) If we take this intuition at face value, that would seem to show the falsehood of STRONG EXTERNALISM, and would give support to our claim that one can have beliefs about objects that do not exist.

The friend of STRONG EXTERNALISM can reject this intuition, seemingly without inconsistency. He can point out that the thought experiment is *worded* in a tendentious way: It says that your twin is under the *illusion* that there is gold in Fort Knox; and because *having the illusion that P* implies *believing that P*, one could claim that the intuition that your doppelganger is expressing a belief about gold when she utters (23) is caused by the wording of the thought experiment, rather than by the substantive features of the case.

But even if the friend of STRONG EXTERNALISM can reject this intuition, notice that the way ahead of him is not easy: He has to explain which is the belief (if any) that your doppelganger expresses, when

she asserts (23). One option is to say that, when your doppelganger asserts (23), she expresses no belief whatsoever. However, this seems hard to believe, since your doppelganger will routinely assert sentences containing the word 'gold' to other Twin-Earthians, and manage, on occasion, to affect the behavior (and presumably, the beliefs) of other people. This strategy would have to make sense of the idea that one can change the beliefs of another person by asserting sentences that express no belief, which seems hard to do.

Another possibility would consist in saying that when your doppelganger asserts (23), he expresses a general belief, one that could perhaps be expressed by means of the following sentence:

(24) In Fort Knox there is a yellow metal that gets dissolved in aqua regia and has atomic number 79

The friend of STRONG EXTERNALISM would then have to provide some explanation of why (24) is *not* what people *in Earth* regularly express when they assert (23). It is not clear whether there is a satisfactory way in which this can be done.

None of these strategies seems plausible, and thus I conclude that the ILLUSION WORLD is a counterexample to STRONG EXTERNALISM. The argument against the claim that there can be people who believe in objects that do not exist is therefore defused.

Chapter 3

Assent and Dissent: A Problem for Salmon and Soames' Theory of Belief Attribution

3.1 Recognition Failure and Belief Attribution

You've surely heard this story before. Charles is confused about Hesperus and Phosphorus: he thinks that they are two different planets, while in reality they are one and the same, the planet Venus. Thus, on some nights, Charles assents to:

(1) Hesperus is visible,

while at the same time he dissents from:

(2) Phosphorus is visible

We can then report, very naturally, what Charles believes by means of the following belief attributions:

(3) Charles believes that Hesperus is visible

(4) Charles does not believe that Phosphorus is visible

To describe Charles' belief state by means of both (3) and (4) would seem the most natural thing in the world. Moreover, there is a very plausible thesis about the connection between what a speaker says and what he believes that has the consequence that (3) and (4) are both true:

> DISQUOTATION: Suppose English sentence ⌜S⌝ expresses proposition P with respect to context C; and suppose also that A is a competent speaker of English who, in C, is speaking sincerely. Then:
> (i) if A has sincerely assented to ⌜S⌝ in C, then ⌜A believes that S⌝ is true in C, and in any other context in which the embedded sentence would express C if asserted; and
> (ii) if A has sincerely dissented from ⌜S⌝ in C, then ⌜A believes that S⌝ is false in C, and in any other context in which the embedded sentence would express C if asserted

But there is a handful of very plausible theses about the semantics of belief attribution that are in tension with these straightforward observations. I am going to call these theses the THEORETICAL SUPPORT; they are:

> THEORETICAL SUPPORT:
> •The only semantic function of proper names is to contribute their referents to the proposition expressed
> •The verb 'believes' always expresses the same two-place relation between a person and a proposition
> •The denotation of an embedded clause ⌜that S⌝ in a context C is the proposition that would be expressed by an assertion of S in C

When these theses are conjoined with the factual assumption that both 'Hesperus' and 'Phosphorus' denote Venus, the theoretical support has the consequence that (4) expresses a proposition which is the negation of the one expressed by (3). For, on this view, both (1) and (2) express the proposition that Venus is visible, which is the proposition denoted by the *that*-clauses of (3) and (4). Because 'believes' always expresses the same two-place relation, (3) says that Charles is in a certain relation to the proposition that Venus is visible, and (4) says that Charles is *not* in *that* relation to *that* proposition.

And now we are in trouble. On the one hand, DISQUOTATION implies that both (3) and (4) are true. On the other hand, the THEORETICAL SUPPORT has the consequence that the proposition expressed by (4) is the negation of the proposition expressed by (3). But a proposition and its negation cannot both be true; that would be a contradiction.

It is worth getting clear on what we would like to require of a satisfactory solution to our problem. Semantic theories usually try to strike a compromise between theoretical pressures, on the one hand, and empirical pressures, on the other. In the case of our problem, the compromise is between respecting our intuitions and the DISQUOTATION principle, on the one hand, and holding on to the THEORETICAL SUPPORT. In advance of philosophical enquiry, it is simply not possible to say that one way of striking the compromise is better than the other, without begging substantive questions. But at the very least, I think we can agree that, whatever strategy one chooses to solve our problem, the solution must not be arbitrary. This is so because, as I have been trying to emphasize, the theses and the intuitions that give rise to our problem seem very plausible and very reliable, respectively. It is no solution to our problem to simply *say* that a certain thesis, or a certain set of intuitions, must be abandoned; a satisfactory solution to our problem must also give a reasonable explanation of *why* the thesis or intuitions that are put in question must be abandoned.

I think that this requirement of *non-arbitrariness* does not beg any important questions, and yet I am going to argue that, if we accept it, then we would have good reason to reject a popular strategy to solve our problem, the strategy defended by Scott Soames and Nathan Salmon, among others[1]. Salmon and Soames choose to question some of our intuitions, while holding on to what I called the THEORETICAL SUPPORT. Because they endorse the THEORETICAL SUPPORT, they are committed to the claim that (4) is the negation of (3), and thus to the claim that one of them is true, and the other is false. Well, which one is true? Salmon and Soames' view is that (3) is true but (4) false, and thus that our intuitions about (3) are right, while our intuitions about (4) are mistaken. But, I am going to argue, they have not provided us with any reason to prefer this strategy to the opposite one of making (4) true and (3) false. In particular, I will argue that, when properly understood, Salmon and Soames' theory shows at most that, *if* our intuitions about (3) are reliable, then our intuitions about (4) need not be respected. The problem is that they have nowhere shown that our intuitions about (3) are reliable. My conclusion will be that their theory asks us to reject some of our intuitions, but that the choice of the intuitions to be rejected is arbitrary. Thus Salmon

[1] For the classic statements of the view, see Salmon 1986, Soames 1987a, Soames 1987b. For some recent discussions of the theory, which are nevertheless subject to the same difficulty, see Braun 1998 and Saul 1998.

and Soames' theory flaunts the non-arbitrariness requirement.

3.2 Salmon and Soames' Theory

Let me begin by presenting Salmon and Soames' theory. They accept the theses in the THEORETICAL SUPPORT, and make two additions to it. The first addition is meant as a replacement for DISQUOTATION[2]:

> ASSENT: Suppose English sentence ⌜S⌝ expresses proposition P with respect to context C; and suppose also that A is a competent speaker of English who, in C, is speaking sincerely. Then:
> (i) if A has sincerely assented to ⌜S⌝ in C, then ⌜A believes that S⌝ is true in C, and in any other context in which the embedded sentence would express C if asserted
> (ii) Part (ii) of DISQUOTATION is false

The assumption of ASSENT instead of DISQUOTATION blocks the derivation of a contradiction. In particular, ASSENT implies that (3) is true and that (4) is false. Since Charles accepts (1) as true, ASSENT implies that (3) is true, and also that (4) is false, since (4) is the negation of (3). But notice that, even if Charles dissents from (2), ASSENT does *not* also entail the truth of (4), since ASSENT lacks the half of DISQUOTATION that would yield that result.

ASSENT raises a substantive question, however. Earlier we said that the claim that both (3) and (4) are true was very natural; but Salmon and Soames propose that (4) is indeed false. The problem is, if (4) is really false, how come that we have the intuition that (4) is really true? To explain away this intuition, Salmon and Soames propose to appeal to conversational implicatures:

> IMPLICATURE: Our intuitions about the truth value of some belief attributions are explained by the fact that those attributions convey misleading conversational implicatures

Let me illustrate this idea. On Salmon and Soames' view, (4) is really false, but if it were asserted, it would convey a conversational implicature to the effect that Charles does not accept as true the sentence 'Phosphorus is visible.' And this implicature is true. The fact that the implicature is true helps explain why we have the intuition that (4) is true: we simply confuse what is implicated by an assertion of (4) for what

[2] See Soames 1987a, 217–218; and also Salmon 1986, 132.

is semantically encoded by (4). Our intuitions do not reliably indicate the truth value of the proposition expressed by (4), but rather the truth value of the conversational implicature associated with it.

IMPLICATURE has generated a wealth of discussion. Topics that have been addressed in the literature include: the sort of implicatures that IMPLICATURE requires; the content of such implicatures; the felicity of having a theory of belief attribution that tries to locate the source of our intuitions in the pragmatics, rather than in the semantics[3]. These are legitimate questions, though I mention them only to put them aside. In this paper, my main concern is going to be with the other part of their theory, the thesis which I have called ASSENT. Let me now raise a question about it.

3.3 The Problem: Assent and Dissent

On Salmon and Soames' theory the contradiction is avoided by making (3) true and (4) false. But why make (4), rather than (3), the false one? Why not the opposite tack?[4]

Of course, Salmon and Soames justify their assignment of truth values to (3) and (4) by virtue of their endorsement of ASSENT, which allows them to dodge the contradiction. But the truth is that, by itself, this is not a convincing argument for that assignment, for there is an alternative weakening of DISQUOTATION on which the contradiction is avoided, and in which (3) and (4) receive an alternative assignment of truth values. The weakening consists in saying that it is acts of *dissent*, rather than acts of *assent*, the ones that are indicative of belief:

> DISSENT: Suppose English sentence ⌜S⌝ expresses proposition P with respect to context C; and suppose also that A is a competent speaker of English who, in C, is speaking sincerely. Then:
> (i) if A has sincerely dissented from ⌜S⌝ in C, then ⌜A believes that S⌝ is false in C, and in any other context in which the embedded sentence would express C if asserted
> (ii) Part (i) of DISQUOTATION is false

Given that Charles sincerely dissents from (2), DISSENT implies that (4)—'Charles does not believe that Phosphorus was visible'—expresses a truth. At the same time, by the truth-functionality of negation, it fol-

[3] Good discussions of these issues are to be found in Saul 1998 and Braun 1998.
[4] As far as I can tell, this question was first raised by G. W. Fitch. See Fitch 1993, 474.

lows that (3)—'Charles believes that Hesperus is visible'—expressed a falsehood; but notice, in particular, that since we are not assuming the full DISQUOTATION, it does not also follow that (3) expresses a truth. Thus, the conjunction of DISSENT with the THEORETICAL SUPPORT also avoids falling into a contradiction.

Salmon and Soames argue that our intuitions about (3) are correct, and those about (4) are not. But if this proposal is to be taken seriously, then we need some reason why ASSENT should be preferred to DISSENT. The observation that ASSENT can be used to avoid the contradiction is not enough, for DISSENT can also be so used. And is there any other reason?

Here is a quick try, appealing to conversational implicatures. One salient feature of ASSENT and DISSENT is that both have counterintuitive consequences: According to ASSENT, (4) is false, which does not sound right. On the other hand, according to DISSENT, (3) is false, which does not sound right either. Now, according to Salmon and Soames, we can explain away the counterintuitive consequences of ASSENT by claiming that ordinary speakers are inclined to judge (4) as true in spite of its being false because they confuse the proposition expressed by that sentence with what it implicates. Since what it implicates is a true proposition, it is somewhat natural that ordinary speakers tend to judge (4) as true. The hope now is that, since the counterintuitive consequences of ASSENT can be explained away, maybe this constitutes a reason for preferring ASSENT to DISSENT.

But the truth is that the same kind of defense can be given for DISSENT. DISSENT has the unhappy consequence that (3) is false, while we think it is true. The analogous strategy in this case would consist in arguing that competent speakers of English perceive (3) as true, in spite of its being false, because assertions of (3) typically carries a true conversational implicature to the effect that Charles accepts as true the sentence 'Hesperus is visible.' This implicature is true, and if we assume that sometimes speakers confuse what is said for what it is implicated (something that neither Salmon nor Soames can consistently deny), the friend of DISSENT will have a way of explaining away our intuitions about (3). Thus, both theses have counterintuitive consequences, but both are compatible with equally good explanations for why those consequences are not counterintuitive. Therefore, the issue between them cannot be resolved in this way.

Earlier I emphasized that any solution to our problem that proposes to reject some of our intuition should provide a reasonable explanation of *why* those intuitions are to be rejected. The preceding suggests

that Salmon and Soames' appeal to conversational implicatures cannot constitute such an explanation. At most, their appeal to conversational implicatures can be used to show that *if* our intuitions concerning (3) are reliable, then our intuitions concerning (4) need not be respected. But, by itself, the appeal to conversational implicatures does not explain, or even show, that our intuitions concerning (3) are reliable. What we want is a reason to choose between ASSENT and DISSENT, and the appeal to conversational implicatures, by itself, does not give us such a reason.

Our question is, is there any good reason to prefer ASSENT over DISSENT? Salmon and Soames have provided two different answers to this question, and I will discuss them in the next two sections.

3.4 Soames' Argument

Soames has argued that the following argument from Mark Richard manages to show that, as he puts it, dissent is not a "reliable" guide to belief [5]:

> Consider *A*—a man stipulated to be intelligent, rational, a competent speaker of English, etc.—who both sees a woman, across the street, in a phone booth, and is speaking to a woman through the phone. He does not realize that the woman to whom he is speaking—*B*, to give her a name—is the woman he sees. He perceives her to be in some danger—a runaway stemaroller, say, is bearing upon her phone booth. *A* waves at the woman; he says nothing into the phone.... If *A* stopped and quizzed himself concerning what he believes, he might well sincerely utter:
>
> (5) I believe that she is in danger
>
> but not:
>
> (6) I believe that you are in danger
>
> Many people, I think, suppose that... [these sentences] clearly diverge in truth value, (5) being true and (6) being false.... But [this] view... is, I believe, demonstrably false. In order to simplify the statement of the argument which shows that the truth of (6) follows from the truth of (5), allow me to assume that *A* is the unique man watching *B*. Then we may argue as follows:
>
> Suppose that (5) is true, relative to *A*'s context. Then *B* can truly say that the man watching her—*A*, of course— believes that she is in danger. Thus, if *B* were to utter
>
> (7) The man watching me believes that I am in danger
>
> (even through the telephone) she would speak truly. But if *B*'s utterance of (7) through the telephone, heard by *A*, would be true, then *A* would speak truly, were he to utter, through the phone
>
> (8) The man watching you believes that you are in danger

[5] See Soames 1987a, 217–218.

> Thus, (8) is true, taken relative to *A*'s context.
> But, of course,
>
> (9) I am the man watching you
>
> is true, relative to *A*'s context. But (6) is deducible from (8) and (9). Hence, (6) is true, relative to *A*'s context. (Richard 1983, 184–86; examples have been renumbered)

As I understand it, the gist of this argument is to show that DISSENT has false consequences. Richard begins with the assumption that (5) and (8) are true, and proceeds, via a deductively valid argument, to derive the conclusion that (6) is true. Because *A* has dissented from 'You are in danger,' when talking on the phone, DISSENT implies that (6) is false. Thus, according to this argument, DISSENT is false, because it implies the falsehood that (6) is false.

Nevertheless, it is hard to see this argument as providing any reason to prefer ASSENT over DISSENT. To get their argument going, Richard and Soames need the assumptions that (5) and (8) are true (Richard even *says so* explicitly in the passage); but how is this assumption to be justified? Presumably, by appeal to the principle of ASSENT itself. Thus the argument succeeds in proving that DISSENT is false only on the assumption that ASSENT is true. But what we are trying to determine is precisely whether there is any reason to prefer ASSENT over DISSENT, so, in this respect, the argument completely begs the question.

The friend of DISSENT could even counter Soames' argument with a *reductio* of his own. Such *reductio* would begin from the claim that (6) is false, and then proceed, via (9), to derive the claim that (5) and (8) are really false. This alternative *reductio* would be as unconvincing as the one outlined above, for it too assumes the truth of the doctrine that is at stake. This suggests that, by itself, Richard's (and Soames') argument does not provide any substantive reason to prefer ASSENT over DISSENT.

3.4.1 A Principle About Perception and Evidence

An alternative way of looking at Richard's argument draws our attention to the different *justifications* for (5) and (6)[6]. (5) can be justified on the strength of the visual evidence available to *A*, which presents *B* in a phone booth with a steamroller bearing upon her. Because it seems that the following principle is extremely plausible:

[6] I am grateful to Alex Byrne for suggesting this reply on Soames' behalf, and for a conversation in which the following argument emerged.

VISUAL EVIDENCE AND BELIEF (*VEB*): If F is a property that can be visually recognized, X is an agent with normal vision, and in good atmospheric conditions, then:
(i) If it looks to X as if a certain object O has F, then $\ulcorner X$ believes that O is $F\urcorner$ is true,
(ii) If it looks to X as if a certain object O does not have F, then $\ulcorner X$ believes that O is $F\urcorner$ is false

it would seem that we have the best evidence that we might want for the truth of (5)—O being the woman in the booth, and F the property of being in danger. On the other hand, the justification for (6) does not depend on the *visual* evidence at the disposal of A. The two things that, in all likelihood, would incline one to assent to (6) are the fact that A has dissented from 'You are in danger,' and the fact that, through the phone, B has indicated that she is not in danger. Whether dissent is an adequate guide to belief is what is under discussion, but in any case it seems that the testimony from B is not as good evidence to A that B is in danger. At the very least, it seems clear that the principle:

TESTIMONY AND BELIEF (*TB*): If X is told by a source that a certain object O does not have property F, then $\ulcorner X$ believes that O is $F\urcorner$ is false,

is much less plausible than VEB.

We can now count the costs and benefits of endorsing ASSENT or DISSENT. By endorsing ASSENT, we have to renounce to TB, which in any case is not a very plausible principle. On the other hand, by endorsing DISSENT, we would have to renounce to VEB, which *is* a very plausible principle. On these grounds, one could argue that Richard's example does offer some grounds to prefer ASSENT over DISSENT, by specifying one situation which makes clear that the costs of endorsing DISSENT are much higher than the costs of endorsing ASSENT.

Nevertheless, I do not think that this reply succeeds. Initially, the reply seems succesful because, in this particular case, the visual evidence at the disposal of the agent tells a story that accords with ASSENT. But that is a very contingent feature of the example. There are alternative versions of this same story on which the visual evidence at the disposal of the agent tells a story that accords better with DISSENT. Consider, for example, the following variation on Richard's example:

Suppose A both sees a woman B, across the street, in a phone booth, and is speaking to a woman through the phone. He does not realize that the woman to whom he is speaking is

the woman he sees. Visually, he does not perceive her to be in danger: Her demeanor is quite natural, and the street looks clear. But there is a steamroller bearing down on her, a steamroller that *A* cannot see because there are some cars blocking the view.

On the other hand, *B* has perceived that there is a steamroller bearing down on her, and is calmly deciding what to do. He tells *A* on the phone: 'A runaway steamroller is bearing down on me; what shall I do?'

If *A* stopped and quizzed himself, he might well sincerely assent to:

(10) You are in danger

in connection with his phone conversation, and dissent from:

(11) She is in danger

while demonstrating the woman on the phone booth.

Now consider the following two attributions, said by *A*:

(12) I believe that you are in danger (said on the phone)

(13) I believe that she is in danger (accompanied by a demonstration of the phone booth)

Intuitively, (12) is true and (13) is false. Salmon and Soames, by virtue of their endorsement of ASSENT, are committed to the claims that (12) is true—since *A* has assented to (10)—and that (13) is true too, since, on their view, it expresses the same proposition as (12). The interesting point is that, this time, endorsement of ASSENT implies rejection of VEB. In this version of the example, it looks to *A* as if *B* is not in danger, and moreover *A* has normal vision, and the atmospheric conditions are right: Thus VEB implies that (13) is false. On the other hand, endorsement of DISSENT with respect to this example would be quite compatible with VEB. Thus, we see that the argument that endorsement of DISSENT was more costly than endorsement of ASSENT was really the product of choosing the adequate example. Change the example, and it would look as if endorsing ASSENT will be equally costly.

One possible strategy that the friend of ASSENT might pursue further would consist in rejecting part (ii) of the principle VEB, while retaining part (i). I cannot imagine what such an argument would look like, especially since any reason to doubt part (ii) of VEB would also seem a reason to doubt part (i). I conclude that Soames' argument does not offer us any reason to prefer ASSENT over DISSENT.

3.5 Salmon's Argument

3.5.1 Salmon's Semantics

Salmon has presented an argument in favor of ASSENT and against DISSENT that is somewhat different from Soames' argument. To understand Salmon's argument, we first need to take a look at Salmon's proposal about the nature of the belief relation and the denotation of 'believes;' it has three main elements[7].

The first element is the thesis that the acquaintance relation between a person and a proposition is *mediated* by something which he calls a *mode of presentation* of a proposition. It is a good question what these modes of presentation could be, but he nevertheless identifies certain examples of them: A sentence, relative to a context, is a mode of presentation of the proposition it expresses in that context. For the purposes of the present discussion, I will assume that there is nothing problematic about modes of presentation; the important thing to remember is that, on Salmon's view, our acquaintance with propositions is mediated by these modes of presentation.

The second element of Salmon's proposal builds on the first. On Salmon's view, belief is a sort of inward assent to a proposition. But, because our acquaintace with propositions is *mediated* by modes of presentation, there cannot be anything like inward assent to a *bare* proposition; rather, one is disposed to give inward assent to a proposition *under a mode of presentation* of that proposition. This suggests that belief is, really, a three place relation between individuals, propositions, and modes of presentation. Salmon calls this relation *BEL*. On Salmon's view, BEL holds between an individual X, a proposition P, and a mode of presentation m if, and only if, X has given assent to m and m is a mode of presentation of P.

The third element in Salmon's proposal is the view that the verb 'believes' denotes a two-place relation between individuals and propositions. In particular, that the verb 'believes' expresses the existential generalization, over modes of presentation, of the BEL relation. Thus, on his view, the following schema gives the truth conditions of belief attributions:

SALMON'S SEMANTICS: An assertion of $\ulcorner X$ believes that $S \urcorner$ in a context C is true

$$\leftrightarrow \exists m \text{ (Grasps } (X, P, m) \text{ \& BEL } (X, P, m))$$

[7]For Salmon's presentation of his own view, see chapter 8 of Salmon 1986, 103–118.

Now we are in a position to understand Salmon's argument, in which SALMON'S SEMANTICS plays a key role. If SALMON'S SEMANTICS is right, then it follows that the truth conditions of *negative* belief attributions are the following:

An assertion of ⌜X does not believe that S⌝ in a context C is true
$\leftrightarrow$ $\neg\exists\, m$ (Grasps (X, P, m) & BEL (X, P, m))

If this is right, then we can show that (3) is true and (4) is false: Because there is a mode of presentation under which Charles gives inward assent to the proposition that Venus is visible (namely, the sentence 'Hesperus is visible') that fact alone is enough, given SALMON'S SEMANTICS, for the truth of (3) *and* the falsehood of (4). In particular, the fact that Charles inwardly dissents from the proposition that Venus is visible under the mode of presentation 'Phosphorus is visible' is not enough for the truth of (4), since the truth of (4) requires—given SALMON'S SEMANTICS—that Charles had not given assent to the proposition that Venus is visible under *any* mode of presentation. Since DISSENT implies that (4) is true, this is, in effect, an argument against DISSENT, and in favor of ASSENT[8]. How convincing is this argument?

3.5.2 A Reply To Salmon

Salmon's proposal is quite complex, and as with all complex proposals, there are many elements that one could question. For example, Salmon assumes an account of belief as a three-place relation which I am not sure is right, or well-motivated. But I am not going to question it here; I think that it is possible to show that Salmon has still not provided a good reason to prefer ASSENT over DISSENT, *even* if we grant him his analysis of belief in terms of BEL.

A crucial element of Salmon's argument against DISSENT is the assumption of what I called SALMON'S SEMANTICS. Without it, it would not be possible to derive the claim that (3) is true and (4) is false, which is instrumental in showing that DISSENT is false. The problem is that Salmon has not said much to support this particular semantic proposal.

What is peculiar about Salmon's semantic proposal for 'believes' is that it gives more weight to assent

[8] Using Kripke's puzzle, Salmon suggests essentially this same argument. See Salmon 1986, 132.

than dissent. As we have seen, Charles assents to (1) and dissents from (2); SALMON'S SEMANTICS then guarantees that his assent will weigh more than his dissent, so (3) will be true and (4) false. But what is important to realize is that this feature is not a consequence of the analysis of belief in terms of BEL; rather, it is a quite arbitrary feature that has been built into the semantics. Indeed, it would be possible to use BEL to give an alternative semantics for 'believes' on which dissent has more weight than assent. To see this, we will proceed in two stages. In the first place, we shall define the following relation, which should correspond to our intuitive notion of *disbelief*[9]:

DISBELIEF:
$\forall m$ ((Grasps (X, P, m) & $\neg$BEL (X, P, m)) $\rightarrow$ DISBEL (X, P, m))
$\forall m$ ((Grasps (X, P, m) & BEL (X, P, m)) $\rightarrow \neg$ DISBEL (X, P, m))

Salmon can hardly deny the meaningfulness of this relation DISBEL, since it is defined from notions that he himself accepts. The second stage now consists in using this relation DISBEL to give an alternative semantics for negative belief attributions:

ALTERNATIVE SEMANTICS: An assertion of ⌜X does not believe that S⌝ in a context C is true
$\leftrightarrow$ $\exists m$ (DISBEL (X, P, m))

This semantics gives more weight to dissent than to assent. To see this, notice first that the ALTERNATIVE SEMANTICS implies that the truth conditions of *positive* belief attributions are as follows:

An assertion of ⌜X believes that S⌝ in a context C is true
$\leftrightarrow$ $\neg\exists m$ (DISBEL (X, P, m))

It is now easy to see that, by assuming this ALTERNATIVE SEMANTICS, it turns out that the fact that there is a mode of presentation under which Charles dissents from the proposition that Venus is visible would be enough for the truth of (4), and the falsehood of (3). Following the guidelines of Salmon's own argument, this observation can easily be turned into an argument in favor of DISSENT, and against ASSENT, an argument which will be as convincing as the previous one in favor of ASSENT, using SALMON'S SEMANTICS—and this means that neither is at all convincing.

[9] Note well that, though close to it, this notion is different from Salmon's notion of *withheld belief*.

To decide the issue between ASSENT and DISSENT, what we need is some reason to prefer SALMON'S SEMANTICS over its alternative. And here I think that we are back at square one, since there is no reason that Salmon, or a friend of his view, could offer in favor of SALMON'S SEMANTICS, and against the alternative. Salmon thought that ASSENT is really a *corollary* of his analysis of belief in terms of BEL:

> ...[A]t least some version of [ASSENT] is unobjectionable... Kripke remarks that, 'taken in its obvious intent, after all, the principle appears to be a self-evident truth.' What makes the principle self-evident is that it is a corollary of the traditional conception of belief as inward assent to a proposition. Sincere, reflective outward assent (qua speech act) to a fully understood sentence is an overt manifestation of sincere, reflective, inward assent (qua cognitive disposition of attitude) to a fully grasped proposition... (Salmon 1986, 129–130)

But it is hard to see why this should be so. One conclusion of the preceding discussion is that, by itself, the conception of belief as inward assent is not incompatible with the idea that there is such a thing as inward *dissent*, and that one can give a semantics for negative belief attributions in terms of this relation of inward dissent. Thus, by itself, the analysis of belief in terms of BEL does not give support to SALMON'S SEMANTICS, or to the principle of ASSENT that is at stake.

Ultimately, Salmon's proposal fails to answer the question we started out with: When evaluating the truth value of belief attributions, why should we trust assent more than dissent? Just offering a semantics for 'believes' which gives more weight to assent is not enough; one would also like to know *why* a semantics that gives more weight to dissent would not be satisfactory. Unfortunately, Salmon has not provided any explanation of why it should be so.

3.5.3 Yagisawa's Analogy

In a recent paper, Takashi Yagisawa has acknowledged the problem that occupies us, but has argued that Salmon is in a position to solve it[10]. Let us look at his argument. Yagisawa begins by explaining a situation in which there are two characters, Jane and Chuck, in the Hall of Mirrors. The argument then proceeds as follows:

[10] See Yagisawa 1997, especially pp. 357–359.

> ...Jane and Chuck are now in the Hall of Mirrors, so positioned relative to each other and to the mirrors that Jane sees one reflection of Chuck to her north and another reflection of Chuck to her south... Jane does not realize that she is looking at one and the same person...
>
> Jane sees Chuck in the north mirror and waves at him in that direction. This is sufficient for her greeting Chuck. Jane sees Chuck on the south mirror but does not wave at him in that direction. This is not sufficient for her not greeting Chuck. All in all, Jane does greet Chuck. Analogously, when Jane grasps a proposition *P* by means of sentence *S* and BEL (Jane, *P*, *S*)), it is sufficient for her believing *P*, whereas when Jane grasps *P* by means of another sentence *S'* but not-BEL (Jane, *P*, *S'*), it is not sufficient for her not believing *P*... (Yagisawa 1997, 358. The quotation has been slightly edited, substituting Yagisawa's references to particular sentences and propositions by schematic letters, for the sake of making the proposal general.)

If we could take Yagisawa's analogy between greeting and believing at face value, then we would have a reason to prefer SALMON'S SEMANTICS over the alternative, and ASSENT over DISSENT. But can the analogy be taken at face value?

I think that the issue one should concentrate on is whether there is some substantive reason to construe the semantics of positive belief attributions on the model of greeting. For example, we could have chosen instead to construe the semantics of *negative* belief attributions on the very same model. Just as it is sufficient for Jane to greet Chuck to wave at him via the north mirror, we could have said that it is sufficient for the truth of (4) that Charles dissents from (2). On this view, just as the fact that Jane fails to greet Chuck via the south mirror is not sufficient for her not greeting Chuck, the fact that Charles assents to (1) is not sufficient for the falsity of (4), or for the truth of (3). This construal of the semantics of negative belief attributions in effect implies that the semantics of positive belief attributions is not analogous to greeting.

What Salmon needs is an argument to the effect that the semantics of positive belief attributions are best construed as an analogy to greeting, and that the semantics of negative belief attributions cannot be so construed. Yagisawa has not provided any argument to that effect; and the truth is that it is not clear what such an argument would look like.

3.6 Concluding Remarks

I began by emphasizing that any solution to our problem about belief attribution that rejects some of our intuitions should give a reasonable explanation of why those intuitions are mistaken or unreliable. In particular, I argued that any attempt to develop this strategy will face the dilemma of choosing between the

following two options:

OPTION 1: (3) is true and (4) is false

OPTION 2: (4) is true and (3) is false,

and that it was the task of the friend of this strategy to choose one of these options, and explain to us why the other is wrong.

Salmon and Soames have chosen to pursue OPTION 1. In this paper I have argued that they have offered no good reason to pursue OPTION 1, rather than OPTION 2. At most, Salmon and Soames are in a good position to show that, *if* our intuitions concerning (3) are reliable, then our intuitions concerning (4) need not be. But they have nowhere shown that our intuitions concerning (3) *are* reliable.

My conclusion is this. Salmon and Soames urge us to reject our intuitions concerning (4), but they have failed to give us any reason to do that, rather than rejecting our intuitions concerning (3). Their proposal therefore seems to depend on a proposal about which intuitions should be rejected, and which accepted, which is quite arbitrary. As I have explained, it is *okay* for a semantic theory to reject some of our semantic intuitions, only if it provides a reasonable explanation of why those intuitions are mistaken or unreliable. Neither Salmon nor Soames do this; therefore, their theory should be rejected.

If the preceding is right, this is good news for all those who approach our problem about belief attribution by trying to formulate a semantics that respects our intuitions concerning the truth value of (3) and (4). These philosophers face the task of arguing against one or another of the theses in the THEORETICAL SUPPORT. I do not think that this task is easy, but I think it can be done. In any case, friends of this strategy do not have the problem of having to assign truth values to (3) and (4) arbitrarily, since the intuitions of competent speakers provide them with all the justification they will ever need to do that.

Chapter 4

The Subjective Attitude in Belief Attribution

4.1 Belief Attribution and the Subjective View

The familiar problem about substitution of coreferential names in belief attribution contexts arises in the following way. For theoretical reasons, one expects that substitution of coreferential names in a sentence will not affect the proposition expressed[1]. And yet it does. For example, everyone acquainted with the Superman stories knows that (1) is true, while (2) is false:

(1) Lois believes that Superman can fly
(2) Lois believes that Clark Kent can fly

But if (1) is true and (2) is false, then (1) and (2) must express different propositions. The challenge is to explain how this *is* possible, given that the only difference between (1) and (2) is that they have different but coreferential proper names.

There is an attractive strategy to solve this problem that goes as follows. It begins by drawing our attention to the fact that Lois is inclined to assent to (3), and to dissent from (4):

[1]Part of those reasons are the ones presented in Kripke (1980) and Donnellan (1970).

(3) Superman can fly

(4) Clark Kent can fly

If Lois assents to (3) and dissents from (4), it must be because she believes that they express different propositions. This is interesting, because (3) and (4) appear in the *that*-clauses of (1) and (2), respectively. We could now say that the *that*-clauses in (1) and (2) contribute to the proposition expressed *the proposition which Lois takes the embedded sentences to express.* On this view, (1) says that Lois is in the belief relation to a certain proposition, and (2) says that Lois is in the belief relation to a different proposition. Because these are different propositions, we see how (1) and (2) can differ in truth value.

Of course, this is just an instance of piecemeal theorizing; to take it seriously, we need to develop this suggestion in several directions. In the first place, we have to justify the claim that Lois does associate different objects of belief with (3) and (4); this is what I have tried to do in chapter 1.

In the second place, it has to be explained which are the two different objects of belief which Lois associates with (3) and (4), and this is a very controversial question. Many philosophers are convinced that cases like Lois' show that belief cannot be a relation between a person and a proposition, and argue that belief should be conceived instead as a relation between a person and something that they call a *mode of presentation*[2]. However, other philosophers claim that the apparatus of propositions is, after all, flexible enough to accommodate the case of Lois. In their view, Lois takes (3) and (4) to express propositions that are about different people. Because at most one of of those propositions can be about the real Superman, the other must be about a merely fictional, or merely possible, individual[3].

For our purposes, it will not matter who is right on the question of the nature of the object of belief. What matters is that, whoever wins that debate, there is a description of Lois' belief state on which she associates a certain object of belief with (3), and a different object of belief with (4). *This* assumption will be very important in what follows. At the same time, I will remain neutral on the question of the object of belief. For the sake of convenience, I will sometimes talk as if belief really were a relation to propositions, and other times talk as if belief were a relation to modes of presentation; but my way of talking should not

[2]Perhaps among other things. As we will see in a moment, Stephen Schiffer and Mark Crimmins endorse a *triadic* view of belief, on which belief is a *three*-place relation between individuals, propositions and modes of presentation.

[3]An example of this view is offered in Stalnaker (1987), Stalnaker (1986a), and Stalnaker (1986b), and of course, also the view defended in chapters 1 and 2 of this dissertation.

be interpreted as an expression of allegiance to one side of the debate, but rather as a device of convenience. The friend of one or another side in this debate will see that the argument in this section can be reconstructed in terms of her favorite object of belief.

Finally, the third direction in which our idea needs to be developed, and the one that will occupy us in this chapter, is this. To explain how (1) and (2) can differ in truth value, it is not enough to show that Lois associates with (3) and (4) different propositions. One also has to explain how those propositions get to be part of the truth conditions of (1) and (2). This project is by no means trivial. For example, one could let oneself be carried away by our success with (1) and (2), and propose that, in general, a belief attribution:

X believes that *S*

expresses the proposition that *X* is in the belief relation to the proposition that *X* takes (the utterance of) *S* to express. This theory would fail, because it presupposes that the subject of a true belief attribution has an opinion about the proposition expressed by the utterance of *S*; which sometimes is not true. For example, if María does not speak English, but otherwise has all the normal beliefs about Bill Clinton, it seems that we can truly say, in English, that María believes that Clinton is the US President. This simple proposal would prevent our attribution from being true, simply on the grounds that María does not know English, and this does not seem right.

Perhaps one could claim that *that*-clauses have a subjective interpretation only when it is known that the subject of the attribution is mistaken about some identity; otherwise, they denote the proposition expressed by their embedded sentences. The problem is that there are examples of attributions to subjects who are confused about some identity, and in which the subject does not have any opinion about the proposition expressed by the corresponding embedded sentences. In this chapter and the next I will pay special attention to two such cases. The first one concerns attributions containing demonstratives whose utterance the subject has not witnessed[4]:

THE STEAMROLLER: Suppose Alfred both sees a woman across the street in a phone booth, and is speaking to a woman through the phone. He does not realize that the woman to whom

[4]Inspired in an example in Richard (1983)

he is speaking—Betty—is the woman he sees. Through the window, he sees that the woman in the booth is in danger—a runaway steamroller is bearing upon the phone booth. Alfred waves at the woman; he says nothing into the phone.

Suppose that we are watching the scene from a van parked in the street. We have tapped Alfred's phone, so we are able to listen to Alfred's conversation with Betty. We know that the woman talking on the phone and the woman on the phone booth are the same, and also that Alfred has not realized this. Suppose now that at t_1 we say:

(5) Alfred believes that she is in danger

while pointing at the woman in the street, and that at t_2 we say (5) again, this time pointing towards the phone receiver in our ears.

Intuitively, our utterance at t_1 is true, and our utterance at t_2 is false. The problem is that, since Alfred has not witnessed our utterances of 'she is in danger,' he has no opinion about which is the proposition expressed by them. Therefore, the strategy used in the case of Lois cannot be applied to this case.

A similar problem arises in cases of attributions made in one language to a subject who does not speak that language:

CARLOS: Carlos is a monolingual speaker of Spanish who is also an amateur astronomer. He has studied, with his telescope, most of the planets in the Solar System, and knows many things about them. Indeed, he has studied all of them but Venus, which he does not even know exists.

One evening, he discovers a planet which is visible right after sunset, and he immediately decides that the planet is different from all the other planets he has studied so far. He names the planet in question—in Spanish, of course—'Héspero.'

Later, he also discovers a planet that is visible right before daybreak. Again he decides that that planet is different from all the planets he has studied before, the one he calls 'Héspero' included. He names this planet 'Fósforo.'

In his view, he has discovered two new planets; because of this, he goes around saying that there are *ten* different planets—and says different things about each of them.

Suppose that we are well acquainted with Carlos, and with his belief that 'Héspero' and 'Fósforo' refer to two different planets. We know that Carlos is mistaken, but we decide not to tell him.

Suppose that Carlos has asserted the Spanish sentence 'Héspero es brillante pero Fósforo no,' whose literal translation into English is 'Hesperus is brigh but Phosphorus is not.' Then suppose we say:

(6) Carlos believes that Hesperus is bright

(7) Carlos believes that Phosphorus is bright

Intuitively, (6) is true and (7) is false. The problem is that, because Carlos does not speak English, he does not have any opinion about the proposition expressed by the English sentences 'Hesperus is bright' and 'Phosphorus is bright.' Therefore, the strategy we used in the case of Lois cannot be applied to this case either.

I do not think that examples like these two are new to the literature, but it nevertheless seems fair to say that they have not received the attention they deserve. Most proposals about belief attribution manage to distinguish the truth conditions of (1) and (2) by relying on Lois' beliefs to do so. But this strategy does not work with our utterances of (5), or with (6–7), whose subjects simply do not have any opinion about the proposition expressed by (the utterances of) the embedded sentences. Stephen Schiffer, Mark Crimmins and Robert Stalnaker have defended alternative semantics for belief attribution that go beyond the simple *dual* theory discussed. Nevertheless, I am going to argue that neither of those theories provides an explanation of how our utterances of (5), and (6) and (7), can express different propositions.

4.2 Schiffer's Hidden Indexical

4.2.1 Schiffer's Theory

Stephen Schiffer has presented a theory of belief attribution that has at least the potential to handle the cases of THE STEAMROLLER and CARLOS[5]. On Schiffer's view, the verb 'believes' expresses a three-place relation between an individual, a proposition, and something that he calls a *mode of presentation*. Of these three relata, the more important one, and the one that will distinguish the truth conditions of pairs like (1) and (2), are the modes of presentation. On Schiffer's view, modes of presentation mediate a subject's epistemic access to a proposition, and they are used to explain how a subject can assent to and dissent from sentences that express the same proposition. When, for example, Lois assents to 'Superman can fly' and dissents from 'Clark Kent can fly,' she assents and dissents from the same proposition, but does so because she associates different modes of presentation with each sentence.

Schiffer argues that modes of presentation of propositions are structured entities. For example, a mode of presentation of the proposition that Superman can fly will be made up of at least two entities, a mode of presentation of Superman, and a mode of presentation of the property of being able to fly. Schiffer leaves open the question of what modes of presentation are, and for our purposes, it is not too important to decide what they are. For the sake of having a concrete proposal, I will assume that, in particular, modes of

[5]The following presentation of Schiffer's view is based on Schiffer (1992) and Schiffer (1995)

presentation of an individual object are something like *file cards* in which one writes the information that one has about an individual. On this view, the mistake of agents like Lois consists in having two different file cards about Superman, one that describes him as the superhero defending Metropolis —call it '$m_{superhero}$,' and another that describes him as the hapless, bespectacled reporter who works for the *Daily Planet* —call it '$m_{reporter}$.'

One important question that Schiffer has to answer is how modes of presentation come to be part of the truth conditions of belief attributions. On Schiffer's view, there is no overt constituent of belief attributions that has modes of presentation as its semantic value; rather, a mode of presentation comes to be part of the truth conditions of a belief attribution by being the denotation of a tacit *hidden indexical* that is part of the logical form of belief attributions, but does not appear in its surface structure. On Schiffer's view, for example, the hidden indexical that accompanies our utterance of (1) denotes $m_{superhero}$, while the hidden indexical that accompanies our utterance of (2) denotes $m_{reporter}$. If we let 'm_{flying}' name the mode of presentation of the property of being able to fly, then we can represent the propositions expressed by (1) and (2) as (8) and (9), respectively:

(8) *Believes* (Lois, *that Superman can fly*, $< m_{superhero}, m_{flying} >$)

(9) *Believes* (Lois, *that Superman can fly*, $< m_{reporter}, m_{flying} >$)

Schiffer's theory would handle the case of THE STEAMROLLER in a similar way. Schiffer can say, in the first place, that Alfred has two different modes of presentation that describe Betty in different ways: One that describes her as the woman standing in the booth, but not as the woman talking on the phone; and another that describes her as the woman talking on the phone, but not as the woman standing in the booth. Let us call these modes of presentation 'm_{booth}' and 'm_{phone},' respectively, and let us say that Alfred has a mode of presentation of the property of being in danger that we will call m_{danger}. Then Schiffer can argue that, in our utterance of (5) at t_1, the hidden indexical denotes a mode of presentation made up of m_{booth} and of m_{danger}. As a result, the proposition we express at t_1 is:

(10) *Believes* (Alfred, *that Betty is in danger*, $< m_{booth}, m_{danger} >$)

At the same time, Schiffer could argue that the hidden indexical that accompanies our utterance of (5) at t_2 denotes a mode of presentation made up of m_{phone} and of m_{danger}, so that the proposition expressed at t_2 is:

(11) *Believes* (Alfred, *that Betty is in danger*, $< m_{phone}, m_{danger} >$)

It is clear that (10) and (11) are different propositions, because the third relatum of the verb 'believes' is different in each case. It should also be clear how to extend this treatment to the case of CARLOS. The question now is, is this a satisfactory solution to our problems?

4.2.2 Figuring Out the Hidden Indexical

I do not think that merely producing (10) and (11) is enough to explain how our utterances of (5) could express different propositions each time. That Schiffer can produce (10) and (11) is a testimony, perhaps, to the underlying philosophy of mind to which he is committed. However, as semanticists, we cannot rest content with the claim that (10) and (11) are the propositions expressed by our utterances of (5) at t_1 and t_2; we also need a systematic explanation of how the semantics of (5), together with the circumstances in which (5) is uttered, implies that (10) and (11) are the propositions expressed by our utterances at t_1 and t_2. In particular, we need an explanation of how the modes of presentation m_{booth} and m_{phone} come to be part of (10) and (11), respectively. On Schiffer's view, those modes of presentation are contributed by a hidden indexical, and what we therefore need is an explanation of how the hidden indexical accompanying our utterances at t_1 and t_2 manages to denote those modes of presentation. And it is here that Schiffer's theory has a serious problem.

Let us begin by recalling that indexical expressions are characterized by two features:

> INDEXICALS:
>
> (i) An indexical expression has an associated semantic rule that determines the denotation of the indexical, relative to a context of utterance
>
> (ii) For any indexical *I*, there are different contexts C and C′ such that an utterance of *I* in C has a different denotation from an utterance of *I* in C′

An example of an indexical expression, according to this definition, is the word 'today.' The semantical rule that determines the denotation of the word 'today' says that an utterance of the indexical will denote the day in which the indexical is uttered.

Now, if we are to take seriously Schiffer's suggestion that there is a tacit hidden indexical in the logical form of belief attribution sentences, and that this indexical plays a crucial role in the resolution of the puzzles, we need to have an idea of what is the semantic rule that determines the reference of the indexical, relative to a context of utterance. The problem is that Schiffer has not said much about what the semantic rule that determines the denotation of the hidden indexical might look like, and this is a serious problem because, without such a rule, we do not have an explanation of how (10) and (11) get to be the propositions expressed by our utterances of (5) at t_1 and t_2.

In defense of Schiffer, one is tempted to draw a parallel between the hidden indexical, and other indexicals about whose semantic rule we cannot say much. A relevant item of comparison is the indexical 'he.' It seems that the most we can say about the meaning of 'he' is that it denotes, in a context, the male individual that is *most salient* in that context. Another relevant item of comparison are other hidden indexicals that Schiffer takes to be present in the logical form of certain incomplete sentences, such as 'It's raining.' Schiffer argues that, when we utter 'It's raining,' we mean (at least sometimes) that it is raining at some (more-or-less) determinate place. Suppose, for example, that we are talking on the phone, that you are in Chicago while I am in Boston, and that you ask me about the weather in Boston. If I then say 'It's raining,' it seems clear that I mean it's raining *in Boston*. Schiffer's idea is that *Boston* gets to be part of the proposition I would express because it is contributed by a hidden-indexical that accompanies my utterance. As in the case of 'he,' it seems difficult to say what the semantic rule of this indexical could be, except that it is an indexical that denotes the place that is *most salient* in that context.

Based on examples like this, one could argue that the most we can say about the semantic rule that determines the denotation of the hidden indexical in belief attributions is something like the following:

> HIDDEN INDEXICALS: The denotation of a hidden indexical accompanying a belief attribution in a context C is the mode of presentation that is *most salient* in C

However, even if this were right, this still would not help us to understand how (10) and (11) could be the propositions expressed by our utterances at t_1 and t_2, since we have not been explained what makes a mode of presentation *salient* in a given context.

Note that, in this respect, there is a stark contrast between the hidden indexical that accompanies belief attributions and the (putative) hidden indexical that accompanies 'It's raining' and the indexical 'he.' The latter two denote places and male people, and we have a reasonable good understanding of how places and male people can become relevant in a context, we do not have a comparable understanding of how modes of presentation can become relevant in a context. Indeed, the reason why we cannot give an informative rule for the denotation of 'he,' or for the denotation of the indexical 'It is raining,' is probably because we are aware of *many mechanisms* that could raise a male person or a place to saliency. The contrast with the case of the hidden indexical in cases of belief attributions is very marked, since in this case we have not been explained even a single mechanism that could raise modes of presentation to saliency.

Of course, one could try to supplement Schiffer's theory with an account of what makes a mode of presentation salient in a context. One could try to do this, for example, by assimilating the hidden indexical to an ambiguous expression. One way in which cooperative speakers figure out the meaning of ambiguous expressions is by figuring out which interpretation of the expression in question makes the utterance meaningful and true. One could then formulate a criterion of saliency for modes of presentation based on this idea, perhaps along the following lines:

> SALIENCY FOR MODES OF PRESENTATION: Let C be a context in which a belief attribution B is uttered. A mode of presentation M is *salient* in C if B would express a truth on the assumption that the hidden indexical that accompanies B denotes M

Nevertheless, this account of saliency would be very problematic. On the one hand, it would assign the wrong truth conditions to some utterances of belief attributions. Consider, for example, the following sentence:

(12) Lois believes that Clark Kent can fly

Intuitively, (12) is false. But on the current proposal about what makes a mode of presentation salient in a context, it is true. On Schiffer's view, (12) predicates a relation between Lois, the proposition that Superman

can fly, and a mode of presentation of this proposition made up of modes of presentation of Superman and of the property of flying. The question is, which mode of presentation of Superman is denoted by the hidden indexical? Well, it seems clear that (12) would be true if the mode of presentation denoted were one that describes Superman as the superhero defending Metropolis, and false if it denoted a mode of presentation that describes Superman as the hapless, bespectacled reporter who works for the *Daily Planet*. Therefore, by the definition of saliency just given, the mode of presentation that is salient in this context is the one that describes Superman as the the superhero defending Metropolis. By the semantical rule for the hidden indexical, this mode of presentation is also the one denoted by the hidden indexical accompanying (12). If we call this mode of presentation '$m_{superhero}$,' and we stipulate that Lois thinks of the property of being able to fly by means of the mode of presentation 'm_{flying},' then we can represent this proposition as follows:

(13) Believes (Lois, *that Superman can fly*, $< m_{superhero}, m_{flying} >$)

The problem, of course, is that (13) is true, since Lois does believe the proposition that Superman can fly, under this mode of presentation. Thus, this proposal about what makes a mode of presentation salient in a context cannot account for the falsehood of (12).

Against this argument, one could object that perhaps we do not *need* a full-fledged theory of what makes a mode of presentation salient in a context. Take, for example, our utterance of (5) at t_1. We want it to be the case that the hidden indexical that accompanies this utterance denote the mode of presentation m_{booth}, which describes Betty as the person who is standing on the booth. But the demonstration which accompanies our utterance at t_1 demonstrates *the booth*. Couldn't one claim that our demonstration of the booth raises to saliency precisely m_{booth}? If this is right, then the hidden indexical that accompanied our utterance of (5) would denote m_{booth}, which is what we want. (And it is easy to see how to formulate a similar story for our utterance of (5) at t_2.)

There are two problems for this suggestion. First, *granting* that our demonstration of the booth raises m_{booth} to saliency, it is not clear that this suggestion would manage to solve our general problem. What we need is a general explanation of how modes of presentation are raised to saliency, and this proposal about

the case of THE STEAMROLLER, by itself, does not suggest one. The failure becomes more evident if we recall for a moment the case of CARLOS. In that case, I argued that (6) and (7):

(6) Carlos believes that Hesperus is bright

(7) Carlos believes that Phosphorus is bright,

express different propositions. Schiffer's explanation would be that the propositions expressed by (6) and (7) contain different modes of presentation of Venus. But in this case, there are no demonstrations accompanying our utterance, and therefore what raises modes of presentation to saliency would have to be something other than demonstration. But what could it be? Those of us who endorse the *Direct Reference* view of proper names think that proper names do not have descriptive senses associated with them, and it is hard to see what else, besides the names themselves, could do this job.

The second problem is that it is not entirely clear why our demonstration of the booth at t_1 should manage to raise to saliency just the mode of presentation m_{booth}. On the face of it, our demonstration at t_1 is addressed towards the booth, and is expected to raise to saliency all the objects lying in that general direction, and, presumably, also the modes of presentation related to those objects (otherwise, how would m_{booth} ever be raised to saliency by our demonstration of the booth?). Now, *Betty*, who is inside the booth, is one of the objects that is surely raised to saliency by the demonstration. As I argued, Alfred has a mode of presentation m_{phone}, which describes Betty as the person speaking on the phone. The question is, why shouldn't our demonstration also raise to saliency the notion m_{phone}, which is a notion of Betty? What we need to answer this question is an account of how a demonstration of an object O can raise to saliency some, but not all, of the modes of presentation associated with O; however, it is not clear what such an account would look like, and certainly Schiffer has not given any indication of this.

4.2.3 Further Developments

It is worth pointing out that, for independent reasons, Schiffer has departed in some extent from the theory presented here; however, none of the departures can handle our present difficulties. In the first place, Schiffer has argued that the propositions expressed by belief attributions do not contain modes of presentation,

but rather are quantifications over them. On this view, the logical form of a belief attribution:

A believes that *P*

is something like the following:

$$\exists m\ (\phi m\ \&\ \text{Believes}\ (A, \textit{that}\ P, m)),$$

where ϕ is a contextually determined *property* of modes of presentation. On this version of the view, the hidden indexical that accompanies belief attributions does not contribute a mode of presentation to the proposition expressed, but rather a type of mode of presentation (symbolized as 'ϕ')[6].

This version of Schiffer's view suffers the same problems as the one above: To see how this view could solve our problems, it is still crucial to understand how the hidden indexical can contribute the appropriate types of modes of presentation to the proposition expressed, and this, Schiffer has not explained.

In a more recent paper, Schiffer has departed from the theory presented above in yet another way. For independent reasons, Schiffer has apparently despaired of the prospects of finding a semantic rule that determines, relative to a context, the denotation of the hidden indexical accompanying belief attributions. On Schiffer's view, the truth conditions of belief attributions are vague, in that it is indeterminate which is the mode of presentation denoted by the hidden indexical, relative to a context. Schiffer then gives an account of the truth conditions of belief attributions based on the idea of an *admissible precisification*: A belief attribution is true if it is true under all admissible precisifications; false in case it is false under all admissible precisifications; and neither true nor false otherwise[7].

It is not clear that this embellishment of Schiffer's theory can solve our problems. In particular, to show that (1) is true and (2) is false:

(1) Lois believes that Superman can fly

(2) Lois believes that Clark Kent can fly,

[6] See for example Schiffer (1992), p. 503.
[7] See Schiffer (1995), esp. pp. 110–114.

Schiffer has to give us a notion of admissible precisification on which, for example, it turns out that there is no admissible precisification of (1) on which the hidden indexical denotes the mode of presentation $m_{reporter}$, and on which there is no admissible precisification of (2) on which the hidden indexical denotes the mode of presentation $m_{superhero}$. But Schiffer has not indicated what makes a precisification of the meaning of a belief attribution admissible. In all likelihood, this will be a context-sensitive notion, in which case there will be reason to believe that all the problems mentioned in the previous section would reappear here.

In complete fairness, it has to be acknowledged that, though Schiffer actively presents this theory as having some virtues, he does not completely endorse it. Rather, he conceives of it as the best semantic theory for belief attribution, relative to certain assumptions; though he expresses doubts that it can be true[8].

From my point of view, there is no reason why Schiffer's theory could not turn out to be true. The problem I have been stressing is that the theory relies on a hidden indexical whose semantic properties have never been made clear, and on a notion of saliency for modes of presentation that has not been clarified. Perhaps there is a way of carrying out these two tasks, and Mark Crimmins' theory, discussed in the next section, might teach us how to carry them out.

4.3 Crimmins' Providing Conditions

Mark Crimmins has presented a semantics for belief attributions that, in many ways, complements Schiffer's theory[9]. Like Schiffer, Crimmins assumes that belief is a relation between a person, a proposition, and a mode of presentation, and like Schiffer, he leaves open the question of what modes of presentation are[10]. Like Schiffer, Crimmins also says that there is no overt component of belief attributions that contributes modes of presentation to the proposition expressed. The important difference with Schiffer's theory is that Crimmins is not committed to the presence of a hidden-indexical in the logical form of belief attributions.

[8] See especially Schiffer (1992)

[9] The theory is presented in Crimmins (1992). Also relevant is Crimmins and Perry (1989), which presents a simplified version of the same theory

[10] However, it is worth pointing out that, at the same time, Crimmins strongly endorses the hypothesis that modes of presentation are sentences in a language of thought. See Crimmins (1992), chapter 4

Rather, he says that modes of presentation are contributed to the truth conditions of belief attribution by certain pragmatic constraints that he calls *providing conditions*. Crimmins' idea is that the providing conditions associated with a belief attribution determine which is the mode of presentation that is part of the truth conditions of the attribution. Also, he argues that the providing conditions associated with an attribution may change from context to context. It is thus crucial for Crimmins to tell us more about what providing conditions are like, and about which providing conditions obtain in which contexts. Crimmins identifies several kinds of providing conditions; however, I will argue that neither of those can handle the cases of THE STEAMROLLER and CARLOS.

4.3.1 Normality

Crimmins argues that there are such things as *normal* modes of presentation. Using our metaphor of the file cards, we can define a normal mode of presentation of an object x as a mode of presentation that gives a normal description of x. Normal modes of presentation play an important in Crimmins' first example of providing condition. Crimmins argues that, when we report the beliefs of a subject whom we think normal, then using a particular name in the *that*-clause of the attribution will typically raise to saliency the *normal* mode of presentation associated with the referent of that name[11]. These providing conditions allow Crimmins to explain, for example, the case of María: Because we know that María has the normal beliefs about Clinton, using the name 'Clinton' in the *that*-clause of an attribution to her will typically indicate that the proposition expressed by the attribution contains María's normal mode of presentation of Clinton. If we use '$m_{Clinton}$' to name the normal mode of presentation of Clinton, '$m_{president}$' to denote the normal mode of presentation of the property of being the US President, we can now represent the proposition expressed by 'María believes that Clinton is the US President' like this:

$$\textit{Believes}\ (\text{María}, \textit{that Bill Clinton is the US President}, < m_{Clinton}, m_{president} >)$$

Because María thinks of Clinton by means of the normal mode of presentation of Clinton, this proposition is true, which is the right result.

[11] See Crimmins (1992), pp. 158–161.

However, this providing condition cannot assign the right truth conditions to (5–7). Presumably, there is just one normal mode of presentation of Betty. If the propositions expressed by our utterances of (5) were about that normal mode of presentation, we would have to conclude that our two utterances of (5) are about the same mode of presentation, and that therefore they express the same proposition. That would not be right. (And it is easy to see that the same problem would arise in the case of CARLOS.)

It would be possible to claim that there can be more than one normal mode of presentation of an object[12]. If one could make the case that, for example, there is a normal mode of presentation of Venus that is associated with the name 'Hesperus,' and another, different normal mode of presentation of Venus that is associated with the name 'Phosphorus,' then perhaps that could be used to distinguish the truth conditions of (6) and (7). However, it is not clear how this would work out. In the case of CARLOS, as we explained, it is *we* who assert (6) and (7), and *we* know that Hesperus *is* Phosphorus, and also that the referent of 'Hesperus' has exactly the same properties as the referent of 'Phosphorus.' Presumably, any normal mode of presentation of Venus, in that context, would have to reflect our knowledge that Hesperus is Phosphorus; but then it is not clear how we could distinguish between the modes of presentation associated with 'Hesperus' and 'Phosphorus.'

4.3.2 *De Dicto*

The second providing condition operates in cases of what Crimmins calls *de dicto* belief reports, which he defines as reports in which we have reason to believe that the utterer is reporting a belief by using the very same words that the subject would use to do so. In those cases, Crimmins argues that the proposition expressed by the report will contain the mode of presentation which, in the mind of the subject, is associated with (the utterance of) *S*[13]. This providing condition would succeed in assigning different propositions to (1) and (2). If we adopt our conventions from the discussion of Schiffer, we can represent the proposition expressed by (1) as:

Believes (Lois, *that Superman can fly*, $< m_{superhero}, m_{flying} >$),

[12] I am grateful to Stephen Yablo for this suggestion

[13] See Crimmins (1992), pp. 165–68.

because Lois associates the name 'Superman' with her mode of presentation $m_{superhero}$. Also, the proposition expressed by (2) is:

$$\textit{Believes}\ (\text{Lois, } \textit{that Superman can fly}, < m_{reporter},\ m_{flying} >),$$

because Lois associates the name 'Clark Kent' with her mode of presentation $m_{reporter}$. These two propositions are different, because the third relatum in each case is different. Nevertheless, this providing condition cannot get the right results in the cases of THE STEAMROLLER and of CARLOS. In THE STEAMROLLER, Alfred has not witnessed our utterances of 'she is danger,' and therefore does not associate any mode of presentation with them. As for the case of Carlos, remember that he does not speak English, and therefore does not associate any mode of presentation with any English sentence.

4.3.3 Self-Attributions

The third providing condition operates in connection with self-attributions of belief (that is, attributions of the form: I believe that *P*). Crimmins argues that, in this case, the mode of presentation that forms part of the truth conditions is the one which the *utterer* of the attribution himself attaches to (the utterance of) the embedded sentence[14]. It is interesting to note that Crimmins invokes this providing condition to solve a version of THE STEAMROLLER in which *Alfred* asserts the following[15]:

(14) I believe that she is in danger [pointing at phone booth]

(15) I believe that she is in danger [pointing at the phone receiver]

Crimmins' rule for self-attribution would distinguish the propositions expressed by (14) and (15), since Alfred does associate different mode of presentations with his two utterances of 'she.' But the rule cannot handle our version of THE STEAMROLLER, nor the case of CARLOS, which involve attributions to a third person.

[14]See Crimmins (1992), pp. 163–165.

[15]This is the version of the puzzle presented by Richard in his original paper. See Crimmins (1992), p. 164.

4.3.4 Saliency and Relevance

The fourth and most intriguing of the providing conditions Crimmins discusses operates in connection with belief attributions that are motivated by some particular action of the subject of the attribution[16]. Crimmins argues that we can gain epistemic access to a person's mode of presentations simply by learning about which actions that person performs, and/or what perceptions the person is having. If, for example, a person throws a rock in the air, and he appears to have done it on purpose, then we can assume that he has a mode of presentation of the rock. Crimmins then offers the following account of how such modes of presentation get to be part of the truth conditions of belief attributions (keep in mind that for Crimmins *notions* are the modes of presentation of individual objects):

> A notion linked to action and/or perception is provided as an unarticulated constituent of a belief report only when it is both salient in the circumstances and relevant to the report. An example where a notion... is salient but not relevant... is the following. Fred is moving all the books in his office... He is doing this quickly, without noticing the titles of the books he moves. Watching him, we see that he is moving his copy of *Oliver Twist*. Susan remarks, "Fred believes that that book once belonged to Dickens." Here, it is clear that Fred has a perceptual notion of the book, which is associated with the ideas of *being in hand, being a book*, and so on. But it cannot be *this* notion that Susan is talking about, since it is obvious that Fred does not know which book he has in hand as he moves it. Instead, it is clear in the circumstances that Susan is talking about Fred's stable notion of the book, which is no doubt associated with ideas of *being a copy of* Oliver Twist, *being a first edition*, and, if Susan is right, *being owned at one time by Dickens*... (Crimmins 1992, p. 163.)

According to this passage, a mode of presentation is part of the truth conditions of a belief attribution if and only if it is both *salient* and *relevant* in the context in which the attribution is uttered. As I said in the case of Schiffer, appealing to notions like *saliency* or *relevance* in the semantics of belief attributions does not automatically solve our problems; one has to explain what it is that makes a mode of presentation salient and relevant in a context. The passage suggests the following account of what the notions of *saliency* and *relevance* come to, when applied to modes of presentation.

First, saliency. Crimmins' idea seems to be that there are several mechanisms that raise a mode of presentation to saliency. A mode of presentation is typically associated with a proper name, so it can be raised to saliency in a context by using, in that context, the proper name to which it is associated. Also, a

[16]See Crimmins (1992), pp. 161–163.

mode of presentation refers to an object, so it can be raised to saliency in a context if the object that the mode of presentation refers to is itself salient in the context. The latter is, I think, the mechanism that explains why both of Fred's modes of presentation of the book come to be salient in the context Crimmins describes: The book is salient in the context described, and both modes of presentation refer to that book.

Second, relevance. Many contexts in which a belief attribution is uttered are contexts in which we are trying to explain an action or a perception of a subject. It will generally be the case that, in the context in which a belief attribution is uttered, speaker and audience will have a pretty good idea of which modes of presentation of the subject played a causal role in the action or perception being discussed, and which did not. As Crimmins explains, in the case of Fred speaker and audience have a pretty good idea that Fred's stable mode of presentation of the book, rather than his visual mode of presentation, is involved in his belief that the book once belonged to Dickens.

I think that we can grant Crimmins that he has explained some interesting notions of saliency and relevance, as applied to modes of presentation. The problem is that, for a mode of presentation to be part of the truth conditions of a belief attribution, it is not sufficient that the mode of presentation be salient and relevant, in Crimmins' sense. To see this, consider the following example. Suppose that we learn that Lois has rejected an invitation from Superman, under his reporter identity, but has accepted another invitation from him, this time under his superhero identity. In this context, we know that Lois has two modes of presentation of Superman, $m_{superhero}$ and $m_{reporter}$, and presumably both of these modes of presentation are relevant to explain why Lois acted in the way she did. Suppose now that in this context we say:

(16) Lois believes that Superman is attractive,

It seems clear that (16) would be true, and to honor this intuition, we need to say that the mode of presentation that is involved in the truth conditions of (16) is $m_{superhero}$, and not $m_{reporter}$. The problem is that both modes of presentation are both salient and relevant in the context in which (16) is asserted: Both are salient, because both are modes of presentation of Superman, and Superman is salient in the context; and both are relevant, because both play a role in the causal explanation of Lois' action.

One could protest that in this example the providing condition that is relevant to determine which mode

of presentation is part of the proposition expressed should be the one about *de dicto* attributions, discussed above; but this would not solve all the problems. To begin with, this in effect raises the issue of determining what happens in contexts in which more than one of the pragmatic providing conditions described by Crimmins can become operative. It is not clear what the answer to this question should be, and Crimmins does not offer any illumination on this matter.

On the other hand, notice that we can modify the example so that the rule to interpret *de dicto* attributions does not apply. Suppose that, in the circumstances described above, we assert (17):

(17) Lois believes that he is attractive,

and that we accompany our utterance by a demonstration of Superman, *at a moment when he is wearing his red cape and his blue tights*. Suppose also that Lois has not witnessed our utterance, and that therefore she does not associate any particular mode of presentation with it. It is clear that our utterance is true —and indeed that it should express the same as (16)—, but in this case the rule to interpret *de dicto* belief attributions cannot be operative, since Lois does not have any opinion about which belief is expressed by our utterance of the embedded sentence. In this case, if Crimmins' theory is to assign some truth conditions to our utterance of (17), it would be by means of the rule about saliency and relevance. The problem is that, like in the case of (16), both modes of presentation $m_{superhero}$ and $m_{reporter}$ are salient and relevant in the context in which (17) is asserted. Examples like these suggest that it is not enough for a mode of presentation to be part of the truth conditions of an attribution that the mode of presentation be both salient and relevant, in Crimmins' sense.

4.3.5 Translation

It would be possible to supplement Crimmins' account by supplying additional providing conditions. Indeed, the case of attributions made to subjects who do not speak the language of the attribution suggests one more: We could say that this kind of attributions raise to saliency the mode of presentation which the subject associates with an appropriate translation of the embedded sentence, into the language which the subject speaks. This could perhaps help Crimmins' theory solve the case of CARLOS. In that case, re-

member, Carlos is a monolingual speaker of Spanish who believes that morning appearances of Venus and evening appearances of Venus are appearances of different planets. If in those circumstances we say:

(6) Carlos believes that Hesperus is bright

(7) Carlos believes that Phosphorus is bright

we would say a truth, even if Carlos does not associate any mode of presentation with the English sentences 'Hesperus is bright' and 'Phosphorus is bright.'

Now, I do not see any reason why a friend of DIRECT REFERENCE could not say that the proper translation of the proper translation of the Spanish 'Héspero' into English is 'Hesperus,' and not 'Phosphorus;' and that the proper translation of the Spanish 'Fósforo' into English is 'Phosphorus,' and not 'Héspero.' If we now note that Carlos associates different modes of presentation with the Spanish sentences 'Héspero es brillante' and 'Fósforo es brillante,' we see how this suggestion could help us distinguish the truth conditions of (6) and (7).

This suggestion is promising, but very limited, in that there are many other cases which it cannot solve. To begin with, we could imagine a version of the case of CARLOS on which everything is as explained, except for the fact that Carlos does not have any proper names to talk about Venus, in any of its guises. In those circumstances, it would seem that we could still assert (6) and (7) and say a truth, but the problem is that, in this case, Carlos would not associate any mode of presentation with the sentences embedded in (6) and (7).

In the second place, it is clear that this suggestion would not help us with the case of THE STEAMROLLER. In this case, we are assuming that Alfred is a competent speaker of English, the language of the attribution. Therefore appeal to the notion of proper translation will not solve this problem.

4.3.6 Subject-Oriented Counterfactuals

Another possible providing condition is motivated by the thought that, in some occasions, the words used in the *that*-clause of a belief attribution represent our attempt to get at how the subject of the attribution *would* put her belief, if certain circumstances obtained.

One way to put this suggestion to work is this. In the case of THE STEAMROLLER, though Alfred does not have any opinion about the proposition expressed by our utterances of 'she is in danger' at t_1 and t_2, it is clear which opinion *he would have*, were he to learn that we made those utterances. The utterance at t_1 is accompanied by a demonstration of the phone booth, and the utterance at t_2 is accompanied by a demonstration of the phone; therefore, it seems clear that, were he to witness those utterances, he would believe that they express different propositions. This suggests that, in this context, there is operative a providing condition that raises to saliency the mode of presentation which the subject would associate with the embedded sentence, were he to witness the utterance.

Still, this proposal cannot be enough to deal with our general problem. To begin with, it is open to counterexamples in which the beliefs of the subject of the attribution change significantly upon witnessing the utterance in question. For example, consider the following version of THE STEAMROLLER. Remember that, in that story, we are watching the situation from a van parked in the street. From the van, we can see the phone booth through the windows, and we can also listen to the Alfred's conversation with Betty through a loudspeaker. Suppose now that we utter (5):

(5) Alfred believes that she is in danger

And suppose also that we do not accompany our utterance by any demonstration. Rather, at that moment there appears a huge banner hanging over the phone booth, with a big arrow pointing at it. Intuitively, our assertion is true. It is true that we not produced any demonstration accompanying the utterance of (5), but the banner that is hanging over the booth is a good substitute for it.

Now suppose that the banner that contains the arrow has, in addition to it, the inscription 'She is the woman talking to you on the phone, you fool.' That the banner has this inscription should not change the truth value of our utterance; so that it should still be true. Yet it also seems clear that, were Alfred to witness our utterance, and the banner with it, he would realize that the person standing in the phone booth *is* the person talking to him on the phone. And if this were to happen, then Alfred would come to associate with our utterance of 'she' a mode of presentation that describes Betty as the person both standing in the booth and speaking on the phone. Let us call this mode of presentation 'm_{betty}.' According to this proposal, the

proposition expressed by this utterance of (5) would be:

(18) Believes (Alfred, *that Betty is in danger*, $< m_{betty}, m_{danger} >$)

The problem is that (18) is false, because Alfred does not think of Betty by means of the mode of presentation m_{betty}. Thus this proposal would imply that our utterance of (5) in these circumstances would be false, when it seems clearly true.

Even if this problem could be finessed, it would still not be clear how the proposed account would handle the case of CARLOS. In that case, whether or not Carlos witnesses our utterance would not make any difference, since he does not know English, and therefore would not know what those utterances mean even if he witnessed them. Perhaps, in connection with the case of Carlos, one could suggest a providing condition that raises to saliency the mode of presentation which the subject of the attribution *would* associate with the embedded sentence, if he knew the language of the attribution.

The problem with this suggestion is that, on the face of it, if Carlos were to learn English he might find out that 'Hesperus' and 'Phosphorus' refer to Venus. (For example, if the textbook he uses exemplified synonymous proper names by using 'Hesperus' and 'Phosphorus.') Therefore, it is not so clear either that the proposed providing condition might manage to attribute the correct truth conditions to (6) and (7).

My conclusion is that Crimmins' proposal is incomplete, because none of the providing conditions which he offers can handle the problems raised by examples like THE STEAMROLLER and CARLOS. In his book, Crimmins says that "it is not at all a worry that the agent can be unfamiliar with the words the speaker uses" (Crimmins 1992, p. 204); but this does not appear to be true. Crimmins' theory can handle some simple cases of belief attributions whose agents are not familiar with the words the speaker uses, like the case of María; but it cannot handle cases in which, in addition to being unfamiliar with the words the speaker uses, the agent is also confused about the identity of the objects mentioned in the report.

4.4 Stalnaker's Contexts

Stalnaker subscribes to the view that the objects of belief are propositions, understood as sets of possible worlds[17]. Many have thought that this view is incapable of handling the cases of recognition failure, but Stalnaker has shown that it is actually quite resilient. The problem that many people see with it is that, for example, the sentences (3) and (4), repeated here:

(3) Superman can fly

(4) Clark Kent can fly

express the same proposition, the proposition that Superman can fly. For this reason, the argument goes, this view cannot distinguish the truth conditions of (1) and (2):

(1) Lois believes that Superman can fly

(2) Lois believes that Clark Kent can fly

since the *that*-clauses of (1) and (2) would contribute the same proposition to the truth conditions of (1) and (2). But this argument presupposes that a *that*-clauses denotes the proposition that would be expressed by its embedded sentence, if unembedded; and this is an assumption that Stalnaker rejects. On Stalnaker's view, there is more flexibility in the denotation of *that*-clauses than this argument makes it out to be.

Stalnaker's theory begins with a proposal about how the belief state of people suffering from recognition failure ought to be described:

> Consider the bilingual Pierre who sincerely assents, in French, to the statement "Londres est jolie" while dissenting, in English, from "London is pretty." It seems that Pierre has a false but coherent conception of the world. In the possible worlds that are the way Pierre thinks the world is, there are two distinct cities, one that is pretty and is called (in French) "Londres," and one that is not so pretty and is called (in English) "London." (Stalnaker (1986b), p. 126)

This description suggests that Pierre associates different propositions with the sentences "London is pretty" and "Londres est jolie," propositions that are about different cities. It is of course a good question what these

[17] A defense of this view is given in Stalnaker (1984). My presentation of Stalnaker's view about belief attribution and failures of substitutivity is based on Stalnaker (1986a), Stalnaker (1986b) and Stalnaker (1987).

cities are. At most one of those cities can be the real London, so the other city must be a merely possible city. Of course, one can balk at the prospect of describing Pierre's belief state by appealing to a city that does not really exist and is a merely possible entity; but, if one is ready to accept the thesis that propositions are sets of possible worlds, there does not seem to be any reason why one could not accept other advantages provided by the apparatus of possible worlds.

Stalnaker then shows how to use this assumption to explain how (1) and (2) can differ in truth value. He does this by appealing to the notion of a *derived context*, which he describes in the following way:

> ...[A] derived context will be determined by the basic context in the following way: for each possible situation in the basic context, [the subject] will be in a definite belief state which is itself defined by a set of possible situations—the ones compatible with what [the subject] believes in that possible situation. The union of all the possible belief states will be the set of all possible situations that might, for all the *speaker* presupposes, be compatible with [the subject's] beliefs. This set of possible situations is the derived context for interpreting the clauses that are intended to express the contents of [the subject's] beliefs. (Stalnaker (1986a), pp. 146–147)

We can now show that there is a difference between the truth conditions of (1) and (2) if we say that the denotation of the *that*-clauses of (1) and (2) is determined with respect to the derived context. The derivation of the truth conditions would go as follows. To begin with, in the possible worlds which accord with Lois' beliefs, there are two different people which are called, respectively, 'Superman' and 'Clark Kent.' The derived context associated with (1) and (2) will contain only these worlds. In all of these worlds, the names 'Superman' and 'Clark Kent' have different referents, so that (3) and (4) express different propositions, in relation with the derived context of (1) and (2). Call those propositions P and P', respectively. P and P' are different: P is about a man with superpowers, while P' is about a hapless, bespectacled reporter who does not have any superpowers. Thus, we can now say that (1) expresses the proposition that Lois is in the belief relation to P, and that (2) expresses the proposition that Lois is in the belief relation to P'.

Stalnaker's view can be supplemented with the claim that *that*-clauses have another possible denotation, which would be the proposition normally expressed by their embedded sentence. Because the English sentence 'Clinton is the US President' expresses the proposition that Clinton is the US President, and this is a proposition that María (the monolingual speaker of Spanish) believes, this denotation would explain how we can truly say, in English, that María believes that Clinton is the US President. Stalnaker does

not explicitly acknowledge that *that*-clauses (indeed, there is a sense in which the passage quoted above suggests that the denotation of a *that*-clause is always determined with respect to the derived context), but there is nevertheless good evidence that Stalnaker assumes that *that*-clauses do have these two kinds of denotations.

For familiar reasons, it is clear that none of the preceding would be enough to distinguish the truth conditions of our utterances of (5) at t_1 and t_2, or (6) and (7). To begin with, we cannot say that the *that*-clauses of those attributions denote the propositions that their sentences would normally express, for that would assign the same proposition to the *that*-clauses of our utterances of (5), and to (6) and (7). On the other hand, we cannot say either that the denotation of those *that*-clauses is determined with respect to what Stalnaker calls the derived context: Since neither Alfred nor Carlos are aware of our utterances, the respective derived contexts do not determine *any* interpretation for the *that*-clauses.

Stalnaker has also discussed, though briefly, the case of attributions whose utterance has not been witnessed by the subject of the attribution. Here is what he would say, for example, about the case of CARLOS[18]:

> In the case of Carlos, we cannot ask about the proposition that he believes that the sentence 'Hesperus is bright and Phosphorus is not' expresses, for he does not know English. Instead, we ask something like the following question: Were *we* to utter 'Hesperus is bright and Phosphorus not' in a possible world compatible with Carlos' beliefs, what would the content of our utterance be? If the Solar System were arranged so that planet appearing in the morning and the planet appearing in the evening were two different planets, then we would use the names 'Hesperus' and 'Phosphorus' to refer to two different planets. And so, according to the semantical rules in that possible world, the *that*-clauses 'that Hesperus is bright' and 'that Phosphorus is bright' denote two different propositions, propositions that are about different planets.

I think that Stalnaker's proposal is very attractive, but also that it faces some serious problems. In particular, this proposal relies on the truth of the following counterfactual:

(16) Were we to assert 'Hesperus is bright' and 'Phosphorus is bright' in a world that accords to Carlos' beliefs, we would express two different propositions.

It is not obvious to me that (16) is true. One could argue against the truth of (16) in two different ways. First, one could say that, in the world as Carlos takes it to be, English sentences do not mean anything. This

[18]The following has been adapted from Stalnaker (1986a), pp. 186–187. In that passage, Stalnaker presents his proposal by discussing an example that is different from the ones I have been discussing in this chapter. I have adapted the wording so that it applies to the examples I have been discussing.

is so because the semantical rules in this world accord to Carlos' beliefs, and Carlos does not know English. If this is right, then when we assert 'Hesperus is bright' and 'Phosphorus is bright' in the world as Carlos takes it to be, we fail to express any proposition at all. This would yield the result that the *that*-clauses of (6) and (7) fail to have a denotation, which is not right.

Second, one could say that what one means by one's assertion is determined by what one believes, and nothing else. Now, we believe that 'Hesperus' and 'Phosphorus' both refer to Venus; therefore, one could make the case that, were *we* to assert 'Hesperus is bright' and 'Phosphorus is bright' in a world as Carlos takes it to be, we would express each time the proposition that we would normally express, the proposition that Venus is bright. This would mean that the *that*-clauses in (6) and (7) would express the same proposition, which would not be right either.

My point here is not necessarily that Stalnaker is wrong about the truth value of (16); rather, my main point is that counterfactuals like (16) are difficult to evaluate. One has to be told much more before accepting (16) as true. In particular, one needs to have an idea of how to fill out the dots in the schema:

Were *we* to assert sentence S in a world that conforms to agent A's beliefs, we would express proposition...

Stalnaker has not told us how to do this, and for this reason it is hard to assess how well his suggestion handles the cases of THE STEAMROLLER and of CARLOS.

Nevertheless, I also think that there is an attractive core in Stalnaker's proposal. He is suggesting that the denotation of the *that*-clauses of, for example, (6) and (7), is determined with respect to a possible world that accords partly with what Carlos believes, and partly with what we believe. What we need to develop this idea is to figure out which of our beliefs and which of the beliefs of the subject of the attribution are relevant to determining the relevant possible world. I do not think that this can be done by appealing to counterfactuals like (16), but I nevertheless find this idea fascinating. The theory that I will present in the next chapter will probably look very different from Stalnaker's theory, but it can certainly be construed as a way of developing this suggestion.

Chapter 5

A Simulation Semantics for Belief Attribution

Why Can't I Be You?

— The Cure

5.1 The Simulation Semantics

Many philosophers are attracted to the idea that (1) and (2):

(1) Lois believes that Superman can fly

(2) Lois believes that Clark Kent can fly

attribute to Lois the beliefs which *Lois* thinks that the embedded sentences express. Unfortunately, all attempts to turn that idea into a full-fledged theory of belief attribution have failed. I too think that this idea is quite interesting, and I am now going to develop it in a new, and I think promising, direction.

One attractive feature of the semantics that I will offer is that it is neutral among competing accounts of the nature of the belief relation and of the object of belief. The only assumption about belief that is

important for the present semantics is the assumption that, whenever an agent assents to a sentence of the form:

$$X \text{ is } F$$

and dissents from a sentence of the form:

$$Y \text{ is } F,$$

(for coreferential proper names X and Y, and for some predicate F) there is a description of her belief state according to which she takes herself to have expressed different objects of belief each time. This assumption can be satisfied by several different accounts of the belief relation. It can certainly be satisfied by the traditional view that belief is a two-place relation between a person and a proposition, when supplemented by the account of belief in fictional entities supplied in chapter 2 (see chapters 1 and 2). But it can also be satisfied by other alternatives. For example, in chapter 4 we saw Schiffer and Crimmins' view, according to which belief is a *three*-place relation between individuals, propositions, and modes of presentation of propositions. On this revisionist view, the important *relatum* of the belief relation are the modes of presentation. On this view, for example, Lois associates with 'Superman can fly' a mode of presentation different from the one that she associates with 'Clark Kent can fly;' this will be all we need to get the semantics going.

Though, for the sake of convenience, in most of this chapter I will talk as if belief were a two-place relation between individuals and propositions, the theory presented here is really neutral between this view, and the view that belief is a three-place relation between individuals, propositions, and modes of presentation.

5.1.1 Ontological Disagreement

Let me begin with an attempt at diagnosis. I think that the failure to produce a theory of belief attribution that can handle cases like THE STEAMROLLER or CARLOS is a case of *misplaced emphasis*. Traditionally, philosophers have been interested in those pairs because they were puzzled about how substitution of coreferential names could affect the truth value of the embedding sentence. But I think that this is not the only problem raised by pairs like (1–2), our two utterances of (5), and (6–7). Indeed, I think that these

substitutivity failures are just a symptom of another phenomenon. It is my hypothesis that the failure to produce an adequate semantics for belief attribution is the result of the failure to identify and explain this other, deeper, phenomenon.

The phenomenon in question is this. In all the cases we have examined, it turns out that we disagree with the subjects of our attributions over which objects really exist. Take, for example, the case of Lois. Lois thinks that there are two different people, Superman and Clark Kent, while we think that there is only one. In the case of Alfred, he thinks that there are two women around, while we think there is only one. And in the case of Carlos, he thinks that there are ten planets, while we think there are only nine. In all these cases, we are in effect reporting the beliefs of someone with whom we have an ontological disagreement—that is, a disagreement over which objects exist. This realization should raise an interesting question: How do you report the beliefs of a person with whom you have an ontological disagreement?

When we report the beliefs of those with whom we have an ontological disagreement, we find ourselves in a quandary. On the one hand, we want to get as close as we can to the beliefs of the other person, so that, if she believes that there is a certain entity X, then we have to mention the fact that she believes in X. Otherwise, our description will seem incomplete. But, on the other hand, if we believe that there is no such thing as X, then it is a wonder that we ever get to describe the belief state of that person. On our view, X does not exist, and no word in our language can be used to refer to it. The problem is this: How can I use my language, which does not contain any expression that refers to X, to describe the beliefs of a person who believes in X?

I want to emphasize that the problem that I am trying to raise is not one about *ontological commitment*, but rather one about *meaning*. We can all agree, I think, that (3):

(3) Johnny believes that Santa Claus will bring him gifts,

does not commit us to the existence of Santa Claus. Even then, a puzzle remains: Since we do not think that the name 'Santa Claus' refers to anything, and presumably we do not have any notion of him —since we think there is no Santa Claus— how can we then say (3), and rest satisfied that we have said something meaningful? What is the contribution of *our* use of Santa Claus to the truth conditions of (3)?

5.1.2 Simulating Another Person

I am going to defend the thesis that, when we describe the beliefs of people with whom we have an ontological disagreement, we have to put ourselves in their shoes, as it were. The idea is that, when we put ourselves in another person's shoes, we can describe the entities that the other person believes in, even if we ourselves do not believe in them. On this view, what gives meaning to our use of 'Santa Claus' are not our beliefs, but rather the beliefs that we would have if we put ourselves in Johnny's shoes.

To put this account to work, we need an account of what it is to put oneself in another person's shoes. We all have some more or less intuitive understanding of what it is to put oneself in the shoes of another person. (Who has never tried it?) But it would be nice to have a precise understanding of what we do, when we put ourselves in another person's shoes.

On my view, to put oneself in another person's shoes is a two-step process. First, one adds the beliefs of the simulated person to one's own. But because the beliefs of the simulated person may be inconsistent with one's own, adding her beliefs to our own will typically create an unstable belief state. The second step of the simulation process is then to *revise* the resulting belief state, for the sake of removing any inconsistencies that may have been created. The result of this process will be a set of beliefs; if A is the person doing the simulation, and B is the simulated person, I will refer to this set as the *simulation* of B by A.

The process of simulating another person is very similar to the familiar phenomenon of belief revision. But note well an important difference: While the goal of the familiar process of belief revision is to yield true beliefs, the goal of the belief revision involved in simulating another person is to enable us to talk about the beliefs of the other person. For this reason, the process of belief revision that is part of the simulation process will be subject to some special constraints.

Because our goal is to get as close to the beliefs of the simulated person as we can, the inconsistencies that result from adding the beliefs of the simulated person to one's own will almost always be resolved in favor of the beliefs of the other person. But it will not always be so. There are certain aspects of a person's beliefs that cannot be coherently simulated, and talked about at the same time. For example, suppose that we want to talk about Juan's beliefs, and suppose that Juan does not speak English. We cannot simulate all of Juan's beliefs and talk about his beliefs in English at the same time, for we cannot simulate that we

do not know English and talk in English at the same time. Whenever this happens, this sort of beliefs will simply be left out of the simulation.

Except for these special constraints, the second step of the simulation process is just like any old ordinary process of belief revision. This means that, to decide which beliefs are included in the simulation of a certain subject, we can appeal to our intuitions about how we would revise our own beliefs, where we to learn what the other person believes (modulo the constraints mentioned in the previous paragraph). The topic of belief revision is, nowadays, a huge research topic; the good news is that we will only need to scratch its surface to get the semantics going[1].

One important point worth emphasizing is that, in general, to simulate another person is not to duplicate her beliefs. On my account, to simulate another person we add her beliefs to our own, and then revise our beliefs accordingly. If there is an issue on which the simulated person does not have an opinion but we do, then the simulation will contain propositions that we believe but the other person does not. As we will see later, this will be important in the cases of THE STEAMROLLER and of CARLOS.

Another important point that I want to emphasize is that this notion of simulation is compatible with several alternative accounts of the belief relation, and of the object of belief. To see this, let me spell the notion of simulation a bit further, entertaining different assumptions about the nature of the object of belief.

First, suppose one thinks that belief is a two-place relation between individuals and propositions. On this view, one simulates another person by adding the propositions that the simulated person believes to the propositions that one believes, and then revising the result to acommodate the newly added propositions. The important thing is that, since propositions have semantic properties (a proposition can imply another, and can be evidence for or against others), adding someone else's beliefs to our own will create some instability that will need to be resolved by means of a process of belief revision, along the lines described above.

One subtlety about this version of the proposal is that, if what was argued in chapters 1 and 2 was right, then to simulate another person one has to engage in a certain game of make-believe. This is so because, if the argument in chapter 1 is right, there are not enough propositions for us to describe the belief state of

[1] A useful introduction to it is given in Gärdenfors (1988).

agents with whom we have an ontological disagreement. Chapter 2 then proposed to appeal to games of make-believe to solve that problem. For our purpose in this chapter, what this means is that to simulate a person with which one has an ontological disagreement, one has to engage in a game of make-believe in which the entities that the simulated person believes in exist. Once this make-believe is appropriately set up, then we can have at our disposal as many propositions as are necessary to describe the belief state of the simulated person, and engage in a simulation of her.

On the other hand, and as we saw in chapter 4, some philosophers think that belief is a three-place relation between individuals, propositions, and modes of presentation. On this view, the important *relatum* of the belief relation is the mode of presentation, and we can also give an account of simulation based on this assumption. What is important for this version of simulation is that modes of presentation have semantic properties, and that one can say when two modes of presentation are incompatible, or when one is evidence for another.

There are several conceptions of modes of presentation which can be used to do this. For example, remember that in chapter 4 I proposed to understand modes of presentation as *file cards* in which one writes all the information one has about the individual or the property that the mode of presentation is about. On this view, the semantic properties of modes of presentation are gleaned from what is written in each of the file cards. For example, on this view, Lois is in the belief relation to the following modes of presentation of the proposition that Superman can fly, and that Superman cannot fly, respectively:

$$< m_{superhero}, m_{flying} >$$

$$< \neg, < m_{reporter}, m_{flying} >>$$

The idea is that Lois is not incoherent in being in the belief relation to these two modes of presentation, since these modes of presentation are *compatible*, and their compatibility is to be explained by looking at how the two different modes of presentation describe Superman: $m_{superhero}$ describes Superman as being the superhero defending Metropolis, and as not being the hapless, bespectacled reporter who works for the *Daily Planet*; while the mode of presentation $m_{reporter}$ describes Superman as being the hapless, bespectacled reporter working for the *Daily Planet*, and as not being the superhero defending Metropolis. Because

the modes of presentation $m_{superhero}$ and $m_{reporter}$ contain different descriptions of Superman, the two modes of presentation are indeed compatible.

On the view that modes of presentation are the important *relatum* of the belief relation, one simulates another person by adding the modes of presentation to which the other person is in the belief relation to the modes of presentation to which one is in the belief relation. Because modes of presentation can be in relations of compatibility and incompatibility to each other, this will likely cause some instability, that will need to be resolved by means of some kind of belief revision, along the lines described above.

5.1.3 Simulation and Semantics

Putting oneself in another person's shoes can affect the meaning of what we say: If we put ourselves in the shoes of a person who by sentence means that P, and we use sentence S to mean that Q, then putting oneself in that person's shoes means that one will use S to mean that P, not that Q. We can take advantage of this to formulate the following semantics for belief attribution:

> SIMULATION SEMANTICS: An utterance of $\ulcorner X$ believes that $P\urcorner$ in a context C expresses the proposition that X believes a certain proposition B, where B is determined in the following way:
> (i) Normally, B is the proposition expressed by the utterance of P in C
> (ii) If there is ontological disagreement between the utterer of the attribution and its subject, then B is the proposition which, according to the simulation of X by the utterer, is expressed by the utterance of P in C

I emphasize that I want to remain neutral between competing accounts of the nature of the object of belief. Though the semantic rule for belief attributions speaks of *propositions*, it could have been formulated equally well in terms of modes of presentation.

This semantics shows us how to describe the beliefs of Johnny. We disagree with Johnny over whether Santa Claus exists, and thus there is an ontological disagreement between Johnny and us. This means that, to describe Johnny's beliefs, we will have to engage in simulation. When we simulate Johnny, the simulation of Johnny will contain the belief that Santa Claus exists, and that 'Santa Claus' refers to him. We can then use something like (3) to say what it is that Johnny believes.

Let me now show how this idea can be used to deal with the problematic cases of belief attribution.

5.2 Applying Simulation to Recognition Failure

Let me now show how this proposal can be applied to solve the problematic cases. First, the case of Lois. Remember that we want to explain our intuition that (1) is true but (2) is false:

(1) Lois believes that Superman can fly

(2) Lois believes that Clark Kent can fly

Lois believes that Superman and Clark Kent are two different people; therefore, we will need to put ourselves in her shoes, to describe her belief state.

To represent Lois' beliefs, and because I want to be neutral on the question of the nature of belief, here and in what follows I will adopt two conventions. First, I will represent belief states by means of *sentences*. Second, to represent the fact that Lois' beliefs about the superhero defending Metropolis are different from her beliefs about the hapless, bespectacled reporter who works for the *Daily Planet*, I will introduce two new names, 'Superman$_{superhero}$' and 'Superman$_{reporter}$,' specifically for this purpose. When, in what follows, I use 'Superman$_{superhero}$,' it will be understood that I am intending a belief about the superhero protecting Metropolis; and when I use 'Superman$_{reporter}$,' a belief about the reporter. The reader should feel free to interpret these names in her favorite way. Those who are friends of modes of presentation, like Crimmins and Schiffer, will take them to denote two different modes of presentation. Those who are propositionalists will take them to denote two different people (at most one of which can be the real Superman; the other must be a purely fictional, or purely possible, entity). With this machinery at hand, we can say that Lois takes 'Superman can fly' and 'Clark Kent can fly' to express, respectively, (4) and (5):

(4) Superman$_{superhero}$ can fly

(5) Superman$_{reporter}$ can fly

We disagree with Lois on this. Therefore, when we add Lois' beliefs to our own, we will have to remove our own beliefs about the proposition expressed by 'Superman can fly' and 'Clark Kent can fly' to accomodate Lois' beliefs. This means that the simulation of Lois will imply that 'Superman can fly' expresses (4), and that 'Clark Kent can fly' expresses (5). The Simulation Semantics then yields that (1) expresses

the proposition that Lois believes (4), and that (2) expresses the proposition that Lois believes (5). Because (4) and (5) are different propositions, and moreover because Lois believes (4) but does not believe (5), the Simulation Semantics manages to assign the intuitively correct truth conditions to (1) and (2).

The case of Lois is easy, because Lois has an opinion about the proposition expressed by 'Superman can fly' and 'Clark Kent can fly.' Because we disagree with her on this, to simulate her means simply to substitute our beliefs by hers. The cases of THE STEAMROLLER and CARLOS are more challenging, because in those cases, the subject does not have an opinion about the proposition expressed by the embedded clauses of the relevant sentences. It is in connection with those cases that the Simulation Semantics proves its worth.

5.2.1 Alfred, Betty and the Steamroller

In the case of THE STEAMROLLER, Alfred did not realize that the woman she is seeing in the phone booth is the same woman as the one he is talking to on the phone. In that situation, suppose at t_1 we assert:

(6) Alfred believes that she is in danger

while pointing at the booth in the street; and that at t_2 we say (6) again, this time pointing towards the phone. Our assertion at t_1 expresses a truth, and our assertion at t_2 a falsehood. The challenge is to explain this, given that Alfred has not witnessed our utterances, nor the accompanying demonstrations.

It is clear that there is an ontological disagreement between Alfred and us: Alfred thinks that there are two different women in the scene, while we think there is only one. Because there is this disagreement, we will need to engage in simulation to describe Alfred's belief state. Let me now explain how the simulation of Alfred would go, first in general terms. Alfred believes that there is a woman standing inside the phone booth, and another, different woman who is speaking to him on the phone. Thus, when we simulate Alfred, we will have to adopt these beliefs of his. Further, if we are simulating Alfred, it seems clear that a demonstration of the phone booth will pick out the woman in the booth, but not the woman speaking on the phone; and also that a demonstration of the phone receiver will pick out the woman speaking on the phone, but not the woman standing in the booth.

Let us now use the name 'Betty$_{booth}$' to represent Alfred's beliefs about the person standing inside the phone booth, and 'Betty$_{phone}$' to represent Alfred's beliefs about the person speaking on the phone. We can now say that, when we are simulating Alfred, the utterance of 'she is in danger' that is accompanied by a demonstration of the booth expresses (7), and that the utterance of 'she is in danger' that is accompanied by a demonstration of the phone expresses (8):

(7) Betty$_{booth}$ is in danger

(8) Betty$_{phone}$ is in danger

According to the Simulation Semantics, our utterance of (6) at t_1 expresses the proposition that Alfred believes (7), and our utterance of (6) at t_2 expresses the proposition that Alfred believes (8). Since (7) and (8) are different propositions, and moreover Alfred believes (7) but does not believe (8), this explains how our utterance at t_1 is true and our utterance at t_2 false.

A more detailed explanation of this result makes explicit what I called the simulation of Alfred. The first step is to represent Alfred's beliefs and our own; the following chart summarizes them:

ALFRED'S BELIEFS:
(a) There is just one woman in the booth, and she is Betty$_{booth}$
(b) There is just one woman speaking on the phone, and she is Betty$_{phone}$
(c) The woman in the booth and the woman on the phone are different women
OUR OWN BELIEFS:
(d) The utterance of (6) at t_1 was accompanied by a demonstration of the booth
(e) The utterance of (6) at t_2 was accompanied by a demonstration of the phone
(f) There is just one woman in the booth, and she is Betty
(g) There is just one woman speaking on the phone, and she is Betty
(h) The woman in the booth and the woman on the phone are the same woman

To work out how the simulation would go, we have to add Alfred's beliefs to our own, correcting for any inconsistencies that may arise. Because I want to remain neutral between the hypotheses that (a–h) are

propositions and that they are modes of presentation, I will provide two different derivations.

First, suppose that (a–h) are specifications of modes of presentation. On this view, the three names 'Betty,' 'Betty_{booth}' and 'Betty_{phone}' are names for three different modes of presentation of Betty, all of which describe Betty in different ways. The important point is that 'Betty' is supposed to be a name of the normal mode of presentation of Betty, which describes Betty as the person who is both in the phone booth and speaking on the phone to Alfred. Because 'Betty' describes Betty in these two ways, (f) and (g) are incompatible with (c), and therefore do not form part of the simulation. Also, (h) is straightforwardly incompatible with (c), and is therefore also excluded from the simulation. On the other hand, there is nothing in Alfred's belief state that rules out (d) and (e); therefore, they will be part of the simulation.

Second, suppose that belief is a relation to a proposition, and that we describe Alfred as believing that a certain woman is talking to him on the phone, and that another, different woman is standing in the booth. On this view, 'Betty_{booth}' and 'Betty_{phone}' are names of different women. At most one of them can refer to the real Betty, so the other must refer to a different woman, perhaps a merely possible, or merely fictional, woman. Let us suppose that 'Betty_{booth}' is the one that refers to the real Betty[2]. Then (a) and (f) are really the same belief, and (g) and (b), and (h) and (c), are straightforwardly incompatible. (It is easy to see that the result is the same, if we take 'Betty_{phone}' to be the name that refers to the real Betty.)

The upshot is that, whatever your view about the object of belief, the simulation of Alfred will contain precisely the following beliefs:

SIMULATION OF ALFRED:
(a) There is just one woman in the booth, and she is Betty_{booth}
(b) There is just one woman speaking on the phone, and she is Betty_{phone}
(c) The woman in the booth and the woman on the phone are different women
(d) The utterance of (6) at t_1 was accompanied by a demonstration of the booth
(e) The utterance of (6) at t_2 was accompanied by a demonstration of the phone

[2]This decision need not be arbitrary. In chapters 1 and 2 I defend certain interpretive principles that determine when a subject has a belief about a fictional character, rather than about a real one.

It is easy to see that this set of beliefs implies that our utterance of 'she is danger' that is part of our utterance at t_1 expresses (7), and also that our utterance of 'she is in danger' that is part of our utterance at t_2 expresses (8); which is what we need to explain how our utterances at t_1 and t_2 can differ in truth value.

There is a subtlety about this treatment that is worth discussing. Focus, for example, on our utterance at t_1. One could object that, when we utter (6) at t_1, there are many other things that we believe about the object of the accompanying demonstration, besides (d). For example:

(9) The utterance of (6) at t_1 was accompanied by a demonstration of the woman who is speaking on the phone

It is clear that we believe this, because we know that the woman in the phone booth *is* the person speaking on the phone. One could feel curious about why (9) does not make it into the simulation. As in the case of (d), no belief of Alfred is inconsistent with (8), since Alfred has no opinion about whether or not we have uttered (6), or about which demonstrations accompanied it. Furthermore, note that the conjunction of (9) with (b) implies that our utterance of 'she is in danger' that is part of our utterance of (6) at t_1 expresses (8), rather than (7) —which would not be right. It is therefore vital for our purposes to show that there is a principled reason why (9) is excluded from the simulation, while (d) is not.

The answer to this question relies on a general point about belief revision. Suppose that we believe A and B, and that B depends on A for its justification. If in those circumstances we come to believe that $\neg A$, we will have to abandon A *and* B: A because it is inconsistent with what we have learned, and B because the support for it disappears, once we get rid of A. Something like this happens with (9). To justify (9), we need our visual experience about what the object of the demonstration is, plus (h), which says that the person in the phone booth is the person speaking on the phone. When we simulate Alfred, we simply have to renounce to (h), since it is inconsistent with (c). But if we have to renounce to (h), we will also have to renounce to all beliefs whose justification depends on (h), and this means that the simulation of Alfred will not contain (9). On the other hand, (d) does not depend for its justification on (h): The visual experience we have when we utter (6) is enough to justify (d). Therefore, the simulation of Alfred will contain (d). We thus see that, because (d) and (9) have different *pedigree*, the former should be included in the simulation of

Alfred while the latter should not.

This takes care of THE STEAMROLLER. Let me emphasize that the beauty of the Simulation Semantics does not lie in that it distinguishes the beliefs (7) and (8); that is a question for the Philosophy of Mind, and a question that has been answered by Schiffer, Crimmins, Stalnaker and many others. The beauty of the simulation semantics lies in that it manages to assign to our utterances of 'she is in danger' at t_1 and t_2 different propositions, *even if Alfred does not have any opinion about the proposition expressed by the utterances of the corresponding embedded sentences, and even if the utterances of those sentences express the same proposition.*

5.2.2 Carlos, Hesperus and Phosphorus

Let us now move to the case of CARLOS. Recall that Carlos is a monolingual speaker of Spanish who believes that he has discovered two new planets, one visible right after sunset, another right before dawn. Carlos uses the Spanish names 'Héspero' and 'Fósforo' to talk about these planets. As we described the case, (10) is true and (11) is false:

(10) Carlos believes that Hesperus is bright

(11) Carlos believes that Phosphorus is bright

The challenge is to explain this, bearing in mind that, because Carlos does not speak English, he does not associate any proposition with the English sentences 'Hesperus is bright' and 'Phosphorus is bright.'

Since Carlos thinks that there are ten planets, while we think there are only nine, we will need to simulate Carlos to describe his beliefs. The upshot is that, when we simulate Carlos, our names 'Hesperus' and 'Phosphorus' take on the semantic properties which Carlos associates with his names 'Héspero' and 'Fósforo,' respectively. If we can show this, then we are home free, since Carlos believes that those names have different referents. Our task is to show how the simulation semantics connects, so to speak, 'Hesperus' with Carlos' 'Héspero,' and 'Phosphorus' with Carlos' 'Fósforo.' There are two possible strategies to do so, depending on what our beliefs turn out to be.

The first strategy depends on our having beliefs about the circumstances in which the relevant English names were introduced. In this case, the connection would be established by taking advantage of our beliefs

about which those circumstances are, together with the beliefs of the simulated person about which objects would have been salient in those circumstances. Let me explain.

Suppose that we know that 'Hesperus' was introduced in a baptismal ceremony that occurred in the evening, and was accompanied of a demonstration of a certain planet, visible at that moment; and also that 'Phosphorus' was introduced in a baptismal ceremony that occurred in the morning, and was accompanied of a demonstration of a certain planet, visible at that moment. Because Carlos does not have any opinion about English names, this belief of ours will be part of the simulation of Carlos. On the other hand, the simulation will contain Carlos' beliefs that the planet that is visible right after sunset is different from the planet that is visible right before dawn. The conjunction of these propositions implies that 'Hesperus is bright' and 'Phosphorus is bright' express different propositions, which is what we need to distinguish the truth conditions of (10) and (11).

I will not work out this case in detail, because it is very similar to the one discussed in the previous section. Here the crucial point is that the denotation, within the simulation, of 'Hesperus' and 'Phosphorus' is determined by a demonstrative, and the example examined in the previous section already illustrated how the Simulation Semantics deals with this kind of demonstratives[3].

The second strategy handles the case in which we do not know the circumstances in which the names 'Hesperus' and 'Phosphorus' were introduced in the language. In that case, we have to suppose that our knowledge of the reference of those names comes from two other sources: First, beliefs about which properties the referent of 'Hesperus' and 'Phosphorus' are supposed to have, and about which objects instantiate those properties; and second, beliefs about which names, in other languages, translate 'Hesperus' and 'Phosphorus.' The following chart describes Carlos' belief state and our own in such a situation:

[3]Incidentally, note that this strategy does not require Carlos to have any names to refer to what he takes two different planets; it is enough if he believes that the planet visible in the morning is different from the planet visible in the evening.

CARLOS' BELIEFS:
(a) 'Héspero' refers to $\text{Venus}_{hesperus}$
(b) 'Fósforo' refers to $\text{Venus}_{phosphorus}$
(c) 'Héspero' and 'Fósforo' are not coreferential
(d) The body that hovers near the Sun after sunset is different from the body that hovers near the Sun before dawn
OUR OWN BELIEFS:
(e) The Spanish 'Héspero' is translated as 'Hesperus,' not as 'Phosphorus'
(f) The Spanish 'Fósforo' is translated as 'Phosphorus,' not as 'Hesperus'
(g) 'Hesperus' refers to the planet that hovers near the Sun before dawn
(h) 'Hesperus' refers to the planet that hovers near the Sun after sunset
(i) 'Phosphorus' refers to the planet that hovers near the Sun before dawn
(j) 'Phosphorus' refers to the planet that hovers near the Sun after sunset
(k) The planet that hovers near the Sun before dawn is Venus
(l) The planet that hovers near the Sun after sunset is Venus
(m) 'Hesperus' and 'Phosphorus' refer to Venus

Note well that the representation of our own belief state reflects the fact that we are aware that the planet that hovers near the Sun after sunset is the same as the planet that hovers near the Sun before dawn. Now, to work out the simulation of Carlos, we need to figure out the result of adding Carlos' beliefs to our own.

The conjunction of (a–m) implies two different things about the propositions expressed by 'Hesperus is bright' and 'Phosphorus is bright.' On the one hand, the conjunction (a–f) implies that they express different propositions. On the other hand, the conjunction (g–m) implies that they express the same proposition. These hypotheses cannot both be true; accordingly, we must revise our beliefs to avoid that result. On the face of it, there are two ways in which we can do so: One, renounce to our beliefs about translation, (e–f); another, renounce to our beliefs about which properties the referents of 'Hesperus' and 'Phosphorus' are supposed to have, (g–m). The question is, which of these two groups of beliefs should be eliminated by the

revision caused by adding Carlos' belief to our own.

Here it is important to realize that adding Carlos' beliefs to our own would force us to lower the probability that we assign to (g–m). To see this, suppose for a moment that you believed (h–l), and that then you learn (d). (d) is not inconsistent with any of (h–l), but it does imply that, of the pairs (g–h), (i–j) and (k–l), at most one of the members in each pair can be true. However, (d) does not tell us which member of those pairs is the one that is true. Therefore, the right reaction to learning (d) is to lower the probability that we assign to *all* of (g–l). On the other hand, learning (d) would not force us to revise our probability assignment to our beliefs about translation, (e–f). Therefore, if we believe (a–c) and (e–m), and we then learn (d), we will think that (g–m) are less likely to be true than (e–f). The simulation of Carlos would therefore contain the following beliefs:

SIMULATION OF CARLOS:
(a) 'Héspero' refers to $Venus_{hesperus}$
(b) 'Fósforo' refers to $Venus_{phosphorus}$
(c) 'Héspero' and 'Fósforo' are not coreferential
(d) The body that hovers near the Sun after sunset is different from the body that hovers near the Sun before dawn
(e) The Spanish 'Héspero' is translated as 'Hesperus,' not as 'Phosphorus'
(f) The Spanish 'Fósforo' is translated as 'Phosphorus,' not as 'Hesperus'

This implies that 'Hesperus is bright' and 'Phosphorus is bright' express, respectively, (12) and (13):

(12) $Venus_{hesperus}$ is bright

(13) $Venus_{phosphorus}$ is bright

According to the simulation semantics, (10) says that Carlos believes (12), and (11) says that Carlos believes (13). Because (12) and (13) are different beliefs, and Carlos believes (12) but does not believe (13), this yields the intuitively correct results.

Again, I emphasize that the beauty of the present approach does not lie in that it manages to distinguish

(12) from (13). The beauty of it lies in that it manages to assign to the English sentences 'Hesperus is bright' and 'Phosphorus is bright' different propositions, and propositions that Carlos believes, *even if Carlos does not have any opinion about the proposition expressed by those sentences, and those sentences express, in English, the same proposition.*

5.3 Saul on Confused Speakers

In a recent paper, Jennifer Saul has presented an interesting challenge for semantic theories on which the truth conditions of belief attribution depend, to some extent, on what the utterer of the attribution knows about the subject of the attribution. She makes her point by means of the following example[4]:

> THE PORTLAND BISTRO: Alice Metzinger, a Portland bistro chef, is in fact Katherin Ann Power, a fugitive bank robber and member of the FBI's Most Wanted List.
> Some employees at the bistro are fascinated by the FBI's Most Wanted List, and are having a conversation, before the revelation that Alice Metzinger is Katherin Ann Power. One of them, a strange sort who we will call Louie, makes the following claim about Ray, which he intends to be a wild allegation:
>
> (14) Ray believes that Alice Metzinger is wanted by the FBI
>
> He also declares:
>
> (15) Ray believes that Katherin Ann Power is wanted by the FBI
>
> Suppose, first, that things with Ray are such that he'd never suspect that the bistro has any employees who are wanted by a law enforcement agency. He'd turn in anyone who was wanted by the FBI, and he never makes any efforts to turn in any of them. He follows the Most Wanted List closely, however, and he assents to the sentence 'Katherin Ann Power is wanted by the FBI.'

As Saul says, this case evokes the intuition that (14) is false and (15) is true. Given the details of the case, it just seems clear that Ray does not know that Alice Metzinger is Katherine Ann Power, and thus that the truth of (15) does not warrant the truth of (14). This case raises the challenge of explaining how (14) and (15) can differ in truth value, given that their utterer —*i.e.*, Louie— is himself confused about the identity of Alice Metzinger and Katherin Ann Power.

Saul used this example to argue explicitly against Crimmins' theory. Crimmins distinguishes several providing conditions that determine, relative to a context, which mode of presentation is part of the truth

[4]From Saul (1999), pp. 361–362.

conditions of that attribution. But none of those providing conditions seems to get this case right. In particular, the one that looks more promising in this case is the one that operates in the case of *de dicto* belief attributions. *De dicto* belief reports are those in which the utterer of the report is expected to use the very same words which the subject of the report would use to report his own belief. The problem is that one would expect that a belief report is *de dicto* when there is some reason to believe that the subject of the attribution suffers from some kind of mistake, a mistake that makes it important to get right the exact words he would use to report her own belief. Saul argues that, in the case of the Portland bistro, there is no reason to expect (14) and (15) to be *de dicto*, since Louie does not believe that Ray is making any kind of mistake[5].

Saul's example raises a similar challenge for the Simulation Semantics. First, notice that there is no ontological disagreement between Ray and Louie: They are both confused about the identity of Alice Metzinger and Katherine Ann Power, and so both agree that Alice Metzinger and Katherine Ann Power are different people. Bearing this in mind, let us now recall what the Simulation Semantics says:

> SIMULATION SEMANTICS: An utterance of ⌜X believes that P⌝ in a context C expresses the proposition that X believes a certain proposition B, where B is determined in the following way:
> (i) Normally, B is the proposition expressed by the utterance of P in C
> (ii) If there is ontological disagreement between the utterer of the attribution and its subject, then B is the proposition which, according to the simulation of X, is expressed by the utterance of P in C

Since there is no ontological disagreement between Ray and Louie, the propositions expressed by (14) and (15) are calculated by following clause (i), rather than (ii). The problem is that clause (i) has the consequence that (14) and (15) express the same proposition: Since Alice Metzinger is Katherine Ann Power, the sentences 'Alice Metzinger is wanted by the FBI' and 'Katherine Ann Power is wanted by the FBI' express the same proposition, and clause (i) then implies that (14) and (15) both say that Ray is in the belief relation to that proposition. This does not seem right, since the intuition that (14) is false and (15) true is so natural.

However, I do not think that this example is a convincing argument against the Simulation Semantics (or against Crimmins' theory, for that matter). What the example suggests is that, when the utterer of a belief

[5]For Saul's discussion of this case, see Saul (1999), pp. 361–364.

attribution is mistaken about the proposition expressed by some sentence that she uses to describe a belief state, and there is no ontological disagreement between her and the subject of the attribution, the relevant *that*-clause will denote the proposition which the *utterer* of the attribution takes the embedded sentence to express. This suggestion suggests the following reformulation of the Simulation Semantics:

> SIMULATION SEMANTICS: An utterance of $\ulcorner X$ believes that $P\urcorner$ in a context C expresses the proposition that X believes a certain proposition B, where B is determined in the following way:
> (i) Normally, B is the proposition which, *according to the participants in the conversation in* C, is expressed by the utterance of P in C
> (ii) If there is ontological disagreement between the utterer of the attribution and its subject, then B is the proposition which, according to the simulation of X, is expressed by the utterance of P in C

This reformulation of (i) takes care of Saul's example. Let us use the names 'Alice$_{chef}$' and 'Alice$_{thief}$' to keep track of the difference between Ray and Louie's beliefs about Alice Metzinger the chef, and Katherine Ann Power the bank robber. We can then say that both Ray and Louie take 'Alice Metzinger is wanted by the FBI' to express (16), and 'Katherine Ann Power is wanted by the FBI' to express (17):

(16) Alice$_{chef}$ is wanted by the FBI

(17) Alice$_{thief}$ is wanted by the FBI

According to our convention, these are different beliefs; what is more, as the case has been described, it seems clear that Ray believes (16) but not (17). This is good news, because on the revised statement of the Simulation Semantics, clause (i) implies that (14) says that Ray believes (16), and that (15) says that Ray believes (17). Which is as it should be.

5.4 Simulation and Semantic Disagreement

Our Simulation Semantics exploits the idea that, to describe the belief state of subjects with whom we have an ontological disagreement, we have to put ourselves in their shoes to do so. It is a good question whether there are other situations in which the utterer and the audience of a belief attribution would put themselves in the shoes of the subject whose beliefs they are describing.

A particularly likely candidate for this are situations in which there is a *merely semantical disagreement* between the utterer of the attribution and its subject; where a *merely semantical disagreement* is a disagreement only about the meaning of certain words, but not about which objects exist. Here is an example:

> JOHNNY, SMITH AND CLINTON: Suppose that Johnny is mistaken about the reference of the name 'Bill Clinton.' He has heard the name used in conversation as applying to a white-haired individual, so when Johnny sees his neighbour Smith, who has white hair, he thinks that the name 'Bill Clinton' really applies to Smith. From that point on, he acts and talks as if the name 'Bill Clinton' referred to his neighbour Smith.
>
> It must be clear that this is not a case of recognition failure: There is nothing that suggests that Johnny believes that Clinton and Smith are the same person. We can suppose, for example, that Johnny has seen Clinton on TV, and that, though he has not heard his name, it is clear to him that that person is different from his neighbour Smith. For example, when he sees Clinton on TV, he would say 'There's the President;' but he would never say that when his neighbour Smith is present.
>
> We privately find this situation amusing. Now suppose that one day both Johnny and Smith come to visit us, and that Johnny realizes that it is Smith, and not Clinton, the one that came to visit. Suppose also that Johnny's peculiar use of 'Bill Clinton' is common knowledge between my audience and me (perhaps we have even had some fun at Johnny's expense, on this subject), and that then I say the following:
>
> (18) Johnny believes that Bill Clinton visited us today

I think that my assertion of (18) would be true in this situation, and would communicate to my audience that Johnny believes that *Smith* visited us today. But in order for this to be the case, the name 'Bill Clinton' has to be interpreted as referring to Smith, rather than Bill Clinton. In all likelihood, what is going on in this case is that, when (18) is asserted, we are putting ourselves in Johnny's shoes, making-believe that 'Bill Clinton' refers to Smith.

Note well that our disagreement with Johnny is *merely semantical*. We do not disagree with Johnny over which people exist: Johnny believes that Clinton and Smith are two different people, and so do we. If we wanted, we could describe Johnny's beliefs without simulating him; instead of (18), we could have said (19):

> (19) Johnny believes that Smith visited us today

(19) would have conveyed our meaning equally well; nevertheless, this does not make (18) any less appropriate, given the circumstances of the context in which it was uttered. The conclusion that I extract from this example is that, in situations in which there is a merely semantical disagreement between speaker and

hearer, on the one hand, and the subject of the attribution on the other, speaker and hearer can describe the beliefs of the subject by putting themselves in her shoes.

Having established that sometimes we put ourselves in the shoes of people with whom we have a merely semantical disagreement, we can now ask about the conditions under which we can expect that to happen. This is a much more difficult question. One thing that seems clear is that it would be unreasonable for a speaker to put himself in the shoes of another person, if the audience is not familiar with the beliefs of another person. This is shown by the following example[6]:

> MARTIN, ANN AND MARY: Suppose Martin has gotten Ann and Mary mixed up. He gives all the signs of believing, of Ann, that she is the referent of the name 'Mary', and of Mary, that she is the referent of the name 'Ann'. I want to communicate this to you, so I say:
>
> (20) Martin believes that Ann is the referent of 'Mary'

Martin thinks that the name 'Ann' refers to Mary, while you and I think that 'Ann' refers to Ann, so we disagree with Martin about the proposition expressed by the embedded sentence. It is clear, nevertheless, that the most natural interpretation of (20) is as expressing the proposition that Martin believes that Ann (rather than Mary) is the referent of 'Mary.' Here what bars the speaker from engaging in a simulation of Martin is surely that Martin's beliefs about the referent of 'Martin' are not known to the audience; for this reason, the *that*-clause in (20) denotes the proposition which speaker and audience take the embedded sentence to express.

On the other hand, it also seems clear that familiarity with the views of the subject whose beliefs are being reported about is not enough to trigger simulation. This is shown by the following example[7]:

> MONICA, FRED AND TED: Suppose that Fred and Ted are two different people, and that Fred is a good and peaceful doctor, while Ted is a dangerous criminal. Monica appears to have gotten their names mixed up. On the one hand, she says things like 'Fred is a dangerous criminal,' and 'There goes Fred' when Ted is in sight; on the other, she also says things like 'Ted is a good and peaceful doctor,' and 'There goes Ted' when Fred is in sight.
>
> Suppose also that Monica's mistake is common knowledge in our community: Everyone knows that Monica is mistaken in this way, Monica knows that everyone knows, and so on. Indeed, many people have tried to correct Monica's usage, to the point that her confusion is a common topic of conversation with her. However, she does not heed our advice.

[6]For this example, I am grateful to Dean Pettit
[7]From Moore (1999a), pp. 343-345.

In those circumstances, we receive the visit of a friend who wants to know whether the dangerous criminal is in town. Precisely a moment before, Monica and I had received the news that the criminal was indeed in town. Then Monica says, 'I believe that Fred is in town.' Our friend is acquainted with Monica's way of talking, so he understands that Monica is saying that, according to her, the criminal is indeed in town. Now suppose that, in this context, I say:

(21) No Monica, you believe that Ted is in town

Surely (21) is true, and what makes it true is that the *that*-clause of (21) denotes the proposition that *Ted*, the dangerous criminal, is in town. What is interesting is that this happens in circumstances in which it is common knowledge that Monica takes the name 'Ted' to refer to the good and peaceful doctor: We know that Monica believes that, Monica knows that we know, we know that Monica knows, and so on. Thus, this case shows that, in general, common knowledge that there is a semantic disagreement between the subject of an attribution and the speaker and his audience is not enough to force a *that*-clause to be interpreted according to the beliefs of the subject.

This suggests that in contexts in which there is common knowledge that there is a merely semantical disagreement between, on the one hand, the subject of the attribution, and on the other hand, speaker and audience, there will be at least two possible and relevant interpretations of the *that*-clause of a belief attribution: One, according to what speaker and audience believe, and another according to what the subject of the attribution believes. However, it seems difficult to give a rule that accurately characterizes *when*, in these contexts, we should expect one or another interpretation for the *that*-clause.

Perhaps what we find here is another instantiation of a more general problem, the problem of ambiguous proper names. We all know that a single proper name can be used to refer to many people. We also know that, in general, the audience of an utterance containing an ambiguous proper name will normally be able to figure out the denotation intended. Presumably, to do so, the audience relies on things like the purpose of the conversation, its general topic, expectations about what the speaker might have intended, and so on. It would be nice to have an accurate explanation of how these ambiguities are resolved, but at present we do not have one. My suggestion is that when we solve this problem, we will be able to answer our question about the interpretation of *that*-clauses in contexts in which there is common knowledge that there is a semantic disagreement between the speaker and the subject of the attribution.

5.5 Opinionated and Skeptical Subjects

Our treatment of the examples of THE STEAMROLLER and CARLOS relied on the fact that, since Alfred and Carlos had not witnessed our utterances, they had no opinion about what those utterances meant. Thanks to this, the simulations of Alfred and Carlos contained some of our own beliefs about the proposition expressed by those utterances, which was crucial to the derivation of the propositions expressed. But what if Alfred and Carlos actually *had* an opinion about our utterances, though a mistaken one?

For example: We have depicted Carlos as a monolingual speaker of Spanish; but what if, in addition to everything in the story, he believes that the English sentence 'Hesperus is bright' means, say, that the moon is made of cheese? There is nothing that would make this situation impossible. Let us call this the case of the OPINIONATED CARLOS. The problem is that, by our definition of the notion of simulation, the proposition that 'Hesperus is bright' means that the moon is made of cheese will form part of the simulation of the Opinionated Carlos. Of course, we disagree with Carlos on the meaning of these sentences; but, as we have defined the mode of presentation, we simply can't stop Carlos' belief from making it into the simulation. Therefore, the Simulation Semantics implies that (22):

(22) Opinionated Carlos believes that Hesperus is bright

expresses the proposition that Opinionated Carlos is in the belief relation to the proposition that the moon is made of cheese. Since Opinionated Carlos does not believe this, the Simulation Semantics implies that (22) is false, but this hardly seems the right result: Since the situation of Opinionated Carlos *vis à vis* Venus is the same as in the case of CARLOS, there is the intuition that (22) should really be true.

A similar problem could arise in the case of Alfred. As we have described the case, Alfred had no opinion on whether or not we were watching him from a van in the street. But suppose that instead, Alfred thought that no one was watching him in the street. Call this the case of the SKEPTICAL ALFRED. Suppose now that, while we are in the van, we point at the phone booth and utter (23):

(23) Skeptical Alfred believes that she is in danger

Intuitively, our assertion expressed a truth. The problem is that, by our definition of simulation, the simu-

lation of Skeptical Alfred will contain the information that there is no van parked in the street, and no one hearing his phone conversation. According to the simulation of Skeptical Alfred, our utterance never took place. Therefore, the Simulation Semantics does not assign any proposition to our utterance of 'she is in danger,' and as a result (23) does not express a proposition. Again, this hardly seems the right result.

Cases like the Opinionated Carlos and the Skeptical Alfred show that there is a tension between our Simulation Semantics and our definition of simulation: either our definition of simulation or our semantic rule for *that*-clauses has to go. I am going to argue that the right answer is to modify the definition of simulation, so that Opinionated Carlos' strange beliefs about English, and Skeptical Alfred's skeptical beliefs about who is watching him, are excluded from it.

I think there is a natural motivation for this course of action. Our definition of simulation, remember, says that when we simulate another person, we add *all* her beliefs to our own. This is unrealistic: When we utter a belief attribution, we normally are not interested in all of a person's beliefs, but only in some of them. If we are interested in explaining, say, Susie's knowledge of ancient history, Susie's beliefs about who will win the World Series will normally be out of our purview. Therefore, we have to amend our definition of simulation to take this into account.

When we utter a belief attribution, we are normally motivated to do so because we want to describe or explain some particular piece of behavior, or behavioral disposition, of a subject. For example, in the case of Alfred, when we utter (5) we are presumably motivated by our desire to explain his behavior *vis à vis* Betty. And in the case of Carlos, our utterances of (6) and (7) are presumably motivated by our desire to explain his behavior *vis à vis* Venus.

Suppose then that we find ourselves in a context in which we want to explain or describe some particular behavioral disposition of a subject. In that context, there will be some background knowledge about why the subject behaves in that way. That knowledge will normally not amount to a complete explanation of why the subject behaves in the way he does, but in occasion it can be quite significant. For example, in the case of Alfred, there is the background knowledge that Alfred thinks that the person in the booth is different from the person speaking on the phone. In the case of Carlos, there is the background knowledge that he believes that the planet that can be seen in the morning is different from the planet that can be seen in the

evening. These pieces of common knowledge are crucial, in that, if they are absent, the audience will likely find our utterances of (5), (6) and (7) puzzling, rather than informative. We can then use this background information in our definition of simulation:

> SIMULATION: Let us suppose that a belief attribution ⌜*X* believes that *P*⌝ is asserted in a context *C*, in which it is common knowledge that the topic of the conversation are certain behavioral dispositions of *X*. In that context, there is a set of propositions *S* such that there is the common knowledge that *X* believes the propositions in *S*, and also that the claim that *X* believes the propositions in *S* is part of an explanation for the behavioral dispositions of *X* that are the subject of the conversation. The simulation of *X* in *C* is obtained by adding the propositions in *S* to the set containing all the propositions believed by speaker and audience, and revising accordingly

This revised account of simulation manages to get the right results in the cases of Opinionated Carlos and Skeptical Alfred. In the case of the Opinionated Carlos, our utterance of (22) is uttered in a context in which we are interested in Carlos' behavior *vis à vis* Venus. To this purpose, Carlos' belief that the English sentence 'Hesperus is bright' means that the moon is made of cheese is clearly irrelevant. Therefore, this belief is left out of the simulation. In the case of Skeptical Alfred, our utterance of (23) is uttered in a context in which we are interested in Alfred's behavior *vis à vis* Betty. To this purpose, Alfred's beliefs about whether or not there is a van parked down the street are clearly irrelevant, and therefore must be left out of the simulation.

5.6 Motivation and Semantic Innocence

I argued in chapter 3 that the problem of substitutivity in belief attribution arises from the conjunction of four different theses:

> DIRECT REFERENCE: The only semantic function of proper names is to contribute their referents to the proposition expressed
>
> BELIEVES: The verb 'believes' always expresses the same two-place relation between a person and a proposition
>
> THAT-CLAUSES: The denotation of an embedded clause ⌜that *S*⌝ in a context *C* is the proposition that would be expressed by an assertion of *S* in *C*
>
> DISQUOTATION: Suppose English sentence ⌜*S*⌝ expresses proposition *P* with respect to context *C*; and suppose also that *A* is a competent speaker of English who, in *C*, is speaking sincerely. Then:

(i) if A has sincerely assented to $\ulcorner S \urcorner$ in C, then $\ulcorner A$ believes that $S \urcorner$ is true in C, and in any other context in which the embedded sentence would express C if asserted; and
(ii) if A has sincerely dissented from $\ulcorner S \urcorner$ in C, then $\ulcorner A$ believes that $S \urcorner$ is false in C, and in any other context in which the embedded sentence would express C if asserted

As we saw in chapter 3, Salmon and Soames propose to abandon DISQUOTATION. On their view, some of our intuitions about negative belief attributions are not reliable, but our intuitions about positive belief attributions are. I criticized their theory because their rejection of those intuitions seemed arbitrary. In particular, it seemed that they did not have any good reason to prefer their theory to an alternative on which our intuitions about negative belief attributions are reliable, while our intuitions about positive belief attributions are not. For this reason, I rejected their proposal as unmotivated.

On the view I am defending, the guilty party is not DISQUOTATION, but THAT-CLAUSES. According to the Simulation Semantics, a *that*-clause can denote a proposition different from the one expressed by its embedded sentence. To see this, we need not go further than our treatment of (1) and (2), repeated here:

(1) Lois believes that Superman can fly

(2) Lois believes that Clark Kent can fly

'Superman can fly' and 'Clark Kent can fly' express the same proposition; but, according to the Simulation Semantics, they contribute different propositions to the propositions expressed by (1) and (2). This is clearly a violation of THAT-CLAUSES.

In contrast with Salmon and Soames, I think that my proposal is adequately motivated. In particular, the Simulation Semantics relies on the following observation:

OBSERVATION: Most failures of substitution in belief attribution are cases in which there is a disagreement, ontological or semantical, between the speaker and the subject of the attribution

This led us to inquire about how one reports the beliefs of a person with whom one has an ontological or a semantical disagreement. We reached the conclusion that, to report the beliefs of people with whom one has an ontological or semantical disagreement, one sometimes puts oneself in their shoes, which led directly to the formulation of a semantical rule for *that*-clauses that is inconsistent with THAT-CLAUSES.

The upshot is that, once one accepts our OBSERVATION, one is led directly to reject *that*-clauses. And, of course, the OBSERVATION itself seems very plausible. Therefore, it seems to me that we have a very good reason to reject THAT-CLAUSES.

One could nevertheless harbor regrets towards our proposal, because it seems to run afoul of Davidson's doctrine of Semantic Innocence. Davidson formulates this doctrine as follows:

> ...If we could recover our Pre-Fregean semantic innocence, I think it would seem to us plainly incredible that the words 'The Earth moves', uttered after the words 'Galileo said that', mean anything different, or refer to anything else, than is their wont when they come in other environments... (Davidson (1968), p. 108)

Davidson's doctrine has been much discussed; nevertheless, there is a good question about what Davidson means here. In a celebrated paper, Mark Crimmins and John Perry have provided the following interpretation:

> ...[The principle of] *semantic innocence*... [says that] the utterances of embedded sentences in belief reports express just the propositions they would if not embedded, and these propositions are the contents of the ascribed beliefs... (Crimmins and Perry (1989), p. 686)

In effect, Crimmins and Perry's formulation suggest that Davidson's doctrine of Semantic Innocence is simply THAT-CLAUSES. They add that Semantic Innocence is "well-motivated by many considerations in the philosophy of language" (Crimmins and Perry (1989), p. 686), and in their paper they suggest a few times that compliance with Semantic Innocence is one of the virtues of their semantics for belief attribution. If Crimmins and Perry are right, then our theory is in trouble, because it runs afoul of a doctrine which is "well-motivated by many considerations in the philosophy of language." But are they right?

I think that Crimmins and Perry's claim is confused. I think that there is a doctrine of Semantic Innocence that is suggested by Davidson and is well-motivated, but that doctrine is not our old friend THAT-CLAUSES. Indeed, I think that when one gets clear on what Davidson is really saying in the passage above, one sees that our Simulation Semantics is actually compatible with Davidson's doctrine of Semantic Innocence. Let me explain.

To begin with, notice that there are some clear counterexamples to THAT-CLAUSES, even aside from the cases of attributions where there is a failure of substitution. Let me present two of them. The first

one concerns attributions that embed sentences with pronouns that are bound by an expression that occurs outside the *that*-clause. For example, suppose that there is a conversation between Jones, Smith and Mary, in which Mary is the topic of conversation. If Jones says:

(24) She deserves the best,

it is clear that he would express the proposition that *Mary* deserves the best. However, suppose that Jones had said instead:

(25) Mary is like every woman. Every woman believes that she deserves the best

In this case, there is no such thing as the proposition denoted by the *that*-clause, since what proposition would that be? Surely it is not the proposition that Mary deserves the best for this is not what is intended. On the face of it, this is a counterexample to THAT-CLAUSES, for it is clear that 'She deserves the best' expresses the proposition that Mary deserves the best, if unembedded, but does not contribute that proposition to the truth conditions of the attribution in (25).

The other example concerns attributions embedding sentences that contain definite descriptions which take *wide scope*. Consider, for instance:

(26) The shortest spy is about to cross the street

I expect many people would agree that assertions of (26) would express a general proposition, the proposition that the shortest spy is about to cross the street. Now, suppose Jones is the shortest spy, and suppose Smith sees Jones crossing the street. Then there is some intuition that I can say to you:

(27) Smith believes that the shortest spy is about to cross the street

even if it is clear between you and me that Smith does not know that Jones is the shortest spy. The explanation is that, in this case, the description 'the shortest spy' is best interpreted *de re*, or as having wide scope. On this view, (27) attributes to Smith a singular belief, one that is about Jones and that consists of the proposition that Jones is about to cross the street. Therefore, this is a case in which the *that*-clause of

(10) contributes to its truth conditions a proposition that is different from the one it would express, if unembedded. Because of examples like these, I think there is good *independent* evidence that THAT-CLAUSES is false[8].

If THAT-CLAUSES is false, does this mean that Davidson, in the above passage, was completely astray? I do not think so. I think that the most natural interpretation of Davidson's passage is one on which he is describing something like a default position about the denotation of *that*-clauses, one that we may have to resist abandoning it, but one that we *can* abandon, given good reason. This doctrine is best formulated as follows:

> SEMANTIC INNOCENCE: Any theory that implies that THAT-CLAUSES is false must provide an adequate justification for rejecting that thesis

This thesis is, I think, very plausible: When one starts thinking about belief attribution, THAT-CLAUSES seems like the most natural thing in the world. Therefore, Davidson is absolutely right in asking for a justification in case someone should feel tempted to abandon it. My position in this chapter is that THAT-CLAUSES has to be abandoned, *and* that there is a very good reason to do so; therefore, the view defended in this chapter does not run afoul of Davidson's Semantic Innocence.

5.7 Problems with Identity and Ignorance

The Simulation Semantics leaves open some questions that are worth noting. In the first place, the preceding examples were arranged so that the audience of the relevant attributions knew enough about the belief state of the subject they were talking about to know that they were in ontological disagreement with that subject. This assumption is important, for this knowledge is what prompts speaker and hearer to engage in simulation. Nevertheless, not all attributions of belief to agents confused about the identities will be uttered in this kind of situation. Consider, for example, the following example:

[8]Something puzzling about Crimmins and Perry's paper is that, while endorsing Semantic Innocence, in a footnote they acknowledge the existence of counterexamples to it (see Crimmins and Perry (1989), fn 14, p. 697). I find their dual attitude towards Semantic Innocence (enthusiastic endorsement in the text, skepticism in footnotes) rather hard to understand.

LUTHOR AND SUPERMAN: Superman is an alien from the planet Krypton, who leads a double life as the harmless reporter Clark Kent, who works as a journalist at the Daily Planet, and as Superman, the superhero protector of Metropolis. Luthor is a villain whose plans are always derailed by Superman; because of this, Luthor hates Superman, and is determined to eliminate him. However, up to this point, Luthor has not realized that Superman is Clark Kent.

Jones and Smith are two friends of Superman who are well-acquainted with Luthor, and who have a desire to protect their friend from whatever Luthor may do to him. One day Jones learns that Luthor is going to the Daily Planet. Jones is fearful that Luthor may have seen through Superman's disguise, and that he may be going to the newspaper to eliminate Superman. So he asks his friend: Does Luthor believe that their friend Superman is in the Daily Planet? Smith answers:

(28) Luthor believes that Clark Kent is in the Daily Planet

(29) Luthor does not believe that Superman is in the Daily Planet

On this occasion, the context in which the attributions are uttered is subtly different from the previous ones: Here Jones does not know whether Luthor believes that Superman and Clark Kent are the same person, and therefore does not know whether he disagrees with Luthor. Because he does not know this, he does not know whether or not he needs to engage in a simulation of Luthor to interpret (28–29). Therefore, according to our Simulation Semantics, Jones is not in a position to know the proposition expressed by (28–29). Nevertheless, it seems clear, first, that Jones finds the assertions of (28–29) informative, and moreover that those assertions communicate to him the knowledge that Luthor still has not realized that Superman is Clark Kent. The challenge is to explain how Jones manages to extract that information from (28–29) if, according to our semantics, he does not know how to interpret those sentences.

A challenge is raised as well by belief attributions that embed identity sentences. Suppose, for example, that this time Luthor discovers that Superman is Clark Kent, and that we want to communicate this to our audience. It seems that the best way to do it would be to say:

(30) Luthor believes that Superman is Clark Kent

On the face of it, our assertion of (30) would be informative, in the sense that it would tell our audience something that it did not know before. But it is not clear that the Simulation Semantics can account for this. If Luthor believes that Superman is Clark Kent, then there is no significant disagreement, semantical or ontological, between Luthor and us. Therefore the *that*-clause of (30) denotes the belief that both we and Luthor associate with the embedded sentence, a belief that we can represent as follows:

(31) Superman is Superman

The problem now is spelled out in two different ways, depending on what the reader thinks about the nature of the object of belief. If the reader thinks that belief is a relation between a person and a proposition, then he reads (31) as the proposition that Superman is Superman. The problem is that this proposition is trivial, and Luthor is expected to know all trivial propositions. In all likelihood, our audience already knew, before the assertion of (30), that Luthor believed (31). It is thus a puzzle, on this view, how our assertion of (30) could have communicated news to our audience.

The situation is similar, if the reader is a friend of the view that belief is a relation between a person and a mode of presentation. If Luthor knows that Superman is Clark Kent, then he has only one mode of presentation of Superman —*a fortiori*, the same mode of presentation that we have. The problem is that it would not be informative to be told that Luthor believes that Superman is Superman, under of a mode of presentation which has the structure:

$$m, m_{identity}, m$$

Whatever our audience knew about Luthor, they could surely figure out that he had a belief of that form.

I will defend the thesis that these two problems are really the same. In my view, assertions like (30) are informative only in contexts in which the audience does not know whether Luthor has realized that Superman and Clark Kent are the same person. But because this knowledge is crucial to the interpretation of (30), this is a context in which the audience does not know how to interpret (30). At this point, (30) becomes an instance of the same problem as (28–29). My claim is that when we find a solution to the problem of explaining how the audience can interpret (28–29), we will also see how they can find (30) informative. This will be our topic in the following chapter.

Chapter 6

An(other) Indirect Account of the Informativeness of Identity Sentences

6.1 Adequacy Conditions

It seems clear that identity statements can be informative. Remember, for example, the traditional problem of the evening and the morning star. The ancients gave the name 'Hesperus' to the first planet seen in the evening, and 'Phosphorus,' to the last planet seen in the morning. They did not know that by doing so they were naming the same thing twice—the planet Venus, as a matter of fact—, so that if someone had told them:

(1) Hesperus is Phosphorus

that would have been informative.

To explain how (1) can be informative has seemed particularly difficult if one assumes a very popular thesis about the semantics of proper names[1]:

[1]For arguments for Direct Reference, see Kripke (1980) and Donnellan (1970).

DIRECT REFERENCE: The only semantic function of a proper name is to contribute its referent to the proposition expressed by the sentence containing it

The problem is that, according to Direct Reference, (1) expresses the proposition that Venus is Venus (since both 'Hesperus' and 'Phosphorus' denote Venus). But this proposition appears to be trivial, and certainly something that the ancients already knew, whether or not someone asserted (1) to them. Friends of Direct Reference therefore find themselves in the unenviable position of having to reconcile their views about proper names with the uncontestable observation that (1) would have been news for the ancients. Nevertheless, in this chapter I will present a theory that purports to explain how identity statements can be informative, and that moreover is compatible with Direct Reference.

Let me begin by explaining what (I think) it would take for a theory of the informativeness of identity statements to be successful. It seems natural to assume that what it is for a speech act to be informative is for it to change the beliefs of the audience. Taking this seriously means accepting that the problem has two dimensions: One, the description of the speech act performed by means of the identity sentences; another, the description of the belief state of the audience, both before and after the performance of the speech act in question. I must emphasize the importance of the latter dimension: Because the problem is about describing the *new* belief (or beliefs) that an audience acquires from the assertion of an identity statement, there would be no solution to the problem unless we have some understanding of the belief state of the audience, for the purpose of assessing whether the information communicated is really *new*. Accordingly, a successful theory must tell us at least these two things: Which is the belief state of the typical audience of an informative identity statement, and which is the effect that the identity statement has on that audience.

Second, a constraint that we will take seriously in what follows is something we might call *Frege's Constraint*. Frege argued that (1) cannot simply communicate linguistic information (information that 'Hesperus' and 'Phosphorus' are coreferential, or that sentence (1) is true), but that it must communicate also *astronomical* information. This seems clear, because had someone asserted (1) to the ancients, they would have thereby learned, at least, that the planet seen in the morning was the planet seen in the evening, and perhaps also something about the number of planets too. The general point is that identity statements do not, or do not just, communicate linguistic information, but also information about facts that are non-

linguistic in nature[2].

Third, note that our problem resurfaces again in connection with identity sentences that occur embedded inside belief attributions. Just as (in certain circumstances) it would be informative to assert (1), it would also be informative (in certain, different circumstances) to assert (2):

(2) Luthor believes that Superman is Clark Kent

Direct Reference, plus certain attractive assumptions about the semantics of 'believes' and of *that*-clauses, implies that (2) says that Luthor believes the proposition that Superman is Superman. But surely anyone, Luthor included, knows that Superman is Superman. It would then seem that whatever (2) tells us about Luthor, it cannot be just that he believes that Superman is Superman. The problem, for the friend of Direct Reference, is the same as with (1): To square the theoretical result that (2) expresses a seemingly trivial proposition with the uncontestable observation that (2) could be news to someone. In this problem, as in the one about bare identity statements, identity sentences figure prominently, so it seems reasonable to assume that they are instances of the same problem.

Summing up, what we require of a satisfactory theory of the informativeness of identity statements is the following:

A theory of the informativeness of identity statements must:

- Describe the belief state of the audience of the statement, both before and after the statement
- Describe the information communicated by informative identity statements
- Justify the claim that the information communicated by the statement is new to the audience
- *Frege's Constraint:* Show that what an audience learns from an identity statement includes non-linguistic information (in the case of (1), it should include astronomical information)
- Show that the theory also explains the informativeness of belief attributions embedding identity sentences

Because our goal in this paper is to defend Direct Reference, we will aim for a theory that can do all this and is compatible with Direct Reference.

[2]See Frege (1892).

Note that, from our point of view, it is a mistake to take the problem of the identity statements to be the problem of explaining how they can *express informative propositions.* It may be that identity statements get to be informative by *conveying* informative propositions, rather than by *expressing* them. Indeed, given that DIRECT REFERENCE implies that (1) expresses the proposition that Venus is Venus, this may well be the only option to defend DIRECT REFERENCE. It would therefore beg the question against DIRECT REFERENCE to frame the problem as one about the proposition *expressed* by identity statements. The substantive task is to show that an indirect account of the informativeness of identity statements is plausible, and this is what I will do in this chapter.

6.2 Other Indirect Proposals

I said I will defend the view that identity statements get to be informative because of the information they convey, but not because of the information that they semantically encode. To be sure, I will not be the first in pursuing this strategy, as there already are several versions of it in the literature. Let me summarize briefly these earlier proposals, as well as the problems they face.

Nathan Salmon has advanced the view that (1) is informative because, even if it semantically encodes the uninformative proposition that Venus is Venus, it conveys the more informative proposition that the sentence 'Hesperus is Phosphorus' is true. Because this is something that the audience of (1) presumably did not know beforehand, this indirectly conveyed information would manage to change their belief states[3]. However straightforward this theory may initially seem, there are at least two problems with it.

First, it is not clear how this theory would answer Frege's challenge that the information that (1) conveys is, at least in part, astronomical, non-linguistic information. Salmon tells us only that (1) conveys the linguistic information that the sentence (1) is true, but he does not explain how, or whether, other pieces of information might be conveyed.

Second, if identity statements are informative because the convey informative propositions, then one would like an account of the *mechanism* by means of which the informative proposition gets conveyed.

[3]See Salmon (1986), pp. 78–79.

Salmon has not given any account of it. Perhaps Salmon's idea is that, in (1), the proposition that 'Hesperus is Phosphorus' is true gets to be conveyed by a mechanism of *conversational implicature*: Since the assertion of (1) is supposed to express a trivial proposition, perhaps this fact triggers a conversational implicature to the effect that the sentence 'Hesperus is Phosphorus' is true. To defend this line of thought, one would have to explain in detail how the conversational implicature in question is generated. But I do not think that this proposal could succeed. I think it is possible to show that there are cases in which the audience of (1) would find it informative, and yet they would not find (1) trivial, by paying attention to what those audiences believe *before* the statement is made. In what follows we will go at great length over this point, but for now we can say this: At least some of the contexts in which (1) can be asserted with informative results are of one of two kinds, and in neither case is the audience in a position to find (1) trivial. First, contexts in which the audience simply does not know, and would acknowledge that it does not know, the reference of at least one of the names occurring in (1). In this case, the audience will not find (1) trivial for the simple reason that they do not know the proposition expressed by it. Second, contexts in which the audience believes, and would so acknowledge, that the two names in (1) have different referents. In this case, the audience would not find (1) trivial, for, in their view, (1) would be false, since it predicates identity between what the audience takes to be two different planets[4].

There have also been attempts to explain the informativeness of belief attributions embedding identity sentences indirectly. For example, Scott Soames has argued that the informativeness of belief attributions embedding identity statements can be explained by appealing to the following maxim:

> Typically, when we report someone's attitudes in indirect discourse, we are expected to keep as close as the words he or she used, or would use, as is feasible... (Soames (1987b), pp. 117)

If this is right, then we can see that, in virtue of this requirement, an assertion of (2) could be taken to convey the information that Jones takes the sentence 'Hesperus is Phosphorus' to be true, which is something potentially informative. However, Soames' account faces two serious difficulties.

[4]Anne Bezuidenhout has presented additional criticisms of this line of argument in Bezuidenhout (1996), pp. 138–140. Bezuidenhout's point is reminiscent of one that Stephen Schiffer made earlier about the use of conversational implicatures in connection with belief attribution; see Schiffer (1987a).

The first problem is this. On Soames' view, the explanation for the informativeness of attributions like (2) lies in the existence of a pragmatic maxim about the typical intentions of people who assert belief attributions. But because the principle is limited to people who assert belief attributions, it is not clear whether this strategy could be extended to cover the case of the informativeness of bare identity statements. Perhaps the friend of Soames' ideas would endorse the story about the informativeness of bare identity sentences that I have been attributing to Salmon. However, this would violate our requirement that the explanation of the informativeness of bare identity sentences, and of belief attributions containing identity sentences, is the same. On the story I am attributing to Salmon, the informativeness of (1) is explained by saying that (1) expresses a trivial proposition, and that this fact triggers a conversational implicature to the effect that (1) is true. In Soames' account, the (alleged) triviality of attributions like (2) does not play any role, nor is there any mention of conversational implicatures. These two explanations are quite different.

The second problem is this. If there were a pragmatic maxim to the effect that, when we report a belief, we are expected to use the very same words that the subject would use to report her beliefs, then one could claim that an assertion of (2) conveys the proposition that Jones takes the sentence 'Hesperus is Phosphorus' to be true. But there is no such requirement; as the examples discussed in chapters 4 and 5 made clear, there are cases of attributions in which we do not expect the subject to even have an opinion about the proposition expressed by the embedded sentence. In cases like those, there is no reason to think that the subject of the attribution would use the same words that we are using to report her beliefs.

Salmon has given a different twist to the indirect proposal, in connection with belief attributions that embed identity sentences. His theory has two main elements. The first one is a revisionist account of the relation of belief. According to Salmon, belief is better regarded as a three-place relation between a person, a proposition, and something that he calls a mode of presentation, which represents *how* the proposition is taken by the believer. The second element in Salmon's proposal is a semantics for the verb 'believes' according to which it expresses a two-place relation, precisely the existential generalization (over modes of presentation) of the three-place relation of belief. Thus, on Salmon's view, there is a mismatch in adicity between the relation of belief, which is a three-place relation, and the semantics of 'believes,' which expresses only a two-place relation. According to Salmon, this mismatch plays a role in generating conversational

implicatures associated with belief attribution:

> ...In attributing beliefs, we are stating whether the believer is favorably disposed to a certain piece of information or a proposition... [But] we should, in... cases where the believer's disposition depends upon and varies with the way the proposition is taken, want to specify not only the proposition agreed to but also something about the way the believer takes the proposition when agreeing to it... (Salmon (1986), pp. 115-6)

The idea is that in cases when we would like to know the way in which a believer takes the proposition, a conversational implicature is generated, presumably an implicature that contains information which is new to the audience. Salmon's account faces two problems.

In the first place, Salmon's story has the problem that it cannot be generalized to explain the informativeness of bare identity sentences. On Salmon's view, the explanation of why attributions like (2) come to be informative relies on the hypothesis that there is a mismatch in adicity between the relation of belief and the semantics of 'believes.' Because this explanation relies on some particular features of the semantics of 'believes' and of the belief relation, it could never be generalized to the case of bare identity sentences, which do not contain the verb 'believes.'

In the second place, Salmon's theory is rather incomplete. Salmon does not say much about what is the information conveyed by informative assertions of belief attributions embedding identity sentences. Salmon does say that those assertions convey propositions about the mode of presentation under which the subject of the attribution thinks of a certain proposition; but he does not say what modes of presentation are (which he explicitly leaves as an open question), nor does he explain the mechanism by means of which a particular mode of presentation comes to be part of the proposition conveyed. To take Salmon's proposal seriously, we would have to fill out all these details, and Salmon has not indicated how.

However attractive the indirect account of the informativeness of identity statements may appear, it is clear that there is much work to be done to develop it. This will be my task in the remainder of this chapter.

6.3 The Indirect Account

My strategy will be a version of the indirect approach, but it departs from views like Salmon's or Soames' in some crucial aspects. In particular, I will assume two main theses, neither of which is suggested by Salmon or Soames. The first one is about the belief state of the typical audience of an informative identity statement; because Robert Stalnaker was (to my knowledge) the first one in suggesting it, I will give it his name[5]:

> STALNAKER'S THESIS: The audience of an informative statement containing an identity sentence is not in a position to know the proposition expressed by the statement

What I mean by saying that the audience is not in a position to know the proposition expressed is that, at the time the statement is made, the audience is in a belief state that determines at least two possible interpretations for the statement. Another way of putting this point is by saying that the audience of an informative identity statement is in a position analogous to the position of one who hears a sentence that is ambiguous between two or more readings, but does not know which of the readings is right.

The second thesis explains how, in those circumstances, statements containing identity sentences get to be informative:

> INFERENCE: Because of STALNAKER'S THESIS, the audience of an informative statement containing an identity sentence will react to it by drawing inferences from the fact that the statement was made and that the speaker was trying to be cooperative

In what follows I am going to show how these two theses help us explain how assertions of identity statements can be informative.

6.4 Explaining the Informativeness of Identity Statements

6.4.1 The Simple Case

Let us begin by considering a case of an informative assertion of an identity statement in which Stalnaker's Thesis is clearly borne out. Consider this situation:

[5]See Stalnaker (1978).

THE MAYOR: To the purpose of better governing his people, the Mayor of Athens has decided to learn about astronomy, and with this goal he attends a philosophy congress at Alexandria.

At the congress he discovers that the philosophical world is on the brink of a new discovery that will rock the astronomical world. It was already known that there were nine planets. The eight of Mercury, the Earth, Mars, Jupiter, Saturn, Uranus, Neptune, and Pluto, plus another planet which they called 'Hesperus' because it was thought to appear only in the evening.

But now it has been determined that something that was previously thought to be a star is a planet. The object in question was called 'Phosphorus,' because it was thought to appear only in the morning.

The philosophers are convinced, by transcendental deduction, that the number of planets is exactly nine. Therefore it must be that the planet called 'Phosphorus' is one of the other nine planets known before. They are able to rule out most of them, but for some time they cannot make up their minds whether Phosphorus is Hesperus or Mars. It is at this point that the Mayor arrives to Alexandria, learns about the discovery, and, like the astronomers, wonders whether Phosphorus is Hesperus or Mars.

But things soon get straightened out. At the end of the congress Aristarchus makes a long argument which ends with the assertion of:

(1) Hesperus is Phosphorus

which manages to convince everyone, the mayor included, that Hesperus is Phosphorus.

Obviously the assertion of (1) was informative, in the sense that it gave the mayor some new beliefs. Our task is to explain how this happened.

Let us begin by getting clear on what the mayor learns from the statement. One thing that he seems to learn is linguistic, namely that the name 'Phosphorus' refers to Hesperus, and not to Mars. But as Frege pointed out, this cannot be all. At the very least, the assertion of (1) also gave the mayor the beliefs that the planet seen in the morning really is Hesperus, the planet seen in the evening. This description of what the mayor learns suggests that, before Aristarchus' assertion of (1), the mayor entertained two possibilities: Either 'Phosphorus' refers to Hesperus and the planet seen in the morning is Hesperus, or else 'Phosphorus' refers to Mars, and the planet seen in the morning is Mars. This is summarized in the following chart:

THE MAYOR'S BELIEF STATE

Before	After
Either a: • The planet seen in the morning is Hesperus • 'Hesperus' and 'Phosphorus' refer to Hesperus • The planet seen in the evening is Hesperus • 'Mars' refers to Mars • Hesperus is Hesperus • Mars is Mars *Or b:* • The planet seen in the morning is Mars • 'Mars' and 'Phosphorus' refer to Mars • 'Hesperus' refers to Hesperus • The planet seen in the evening is Hesperus • Hesperus is Hesperus • Mars is Mars	*a:* • The planet seen in the morning is Hesperus • 'Hesperus' and 'Phosphorus' refer to Hesperus • The planet seen in the evening is Hesperus • 'Mars' refers to Mars • Hesperus is Hesperus • Mars is Mars

Given this description of the case, it is clear that the mayor is not in a position to know which is the proposition expressed by the astronomer's assertion. His initial belief state implies that (1) expresses either the proposition that Hesperus is Hesperus (possibility *a*), or that Hesperus is Mars (possibility *b*), but he does not know which. Thus we see that in this case Stalnaker's Thesis is borne out. Our task now is to show how the assertion of (1) could enable the mayor to rule out possibility *b*, given that he does not know which of two propositions (1) expresses.

Even if the mayor cannot know which is the proposition expressed by the astronomer's assertion, he can draw inferences from the fact that the astronomer has asserted (1). Because the mayor believes that the astronomer is cooperative, the mayor believes that the astronomer believes that the astronomer has expressed a truth. Because the mayor trusts the astronomer, the mayor himself comes to believe that (1) is true. Let us now notice that the truth of (1) is incompatible with possibility *b*, since this possibility implies that (1) expresses the proposition that Hesperus is Mars, which simply cannot be true. Learning that (1) is true would therefore allow the mayor to rule out *b*.

Notice our proposal satisfies Frege's constraint that not all that the mayor learns is information about his language. Linguistic information is part of what he learns, but not all. In ruling out *b*, the mayor *also* learns something substantive about astronomy, namely that the planet seen in the morning *is* Hesperus, the planet seen in the evening. It is crucial for this result to understand the structure of the possibilities that the mayor cannot rule out before the assertion of (1): In particular, that if 'Hesperus is Phosphorus' is true, then

the planet seen in the evening is the planet seen in the morning. It is because these possibilities are linked in this way that learning that (1) is true enables the mayor to know something about astronomy. We can therefore accomodate Frege's complaint that what the audience of (1) learns cannot be just purely linguistic information.

One could object that our description of the belief state of the mayor leaves out something important. The description we have proposed acknowledges that the mayor does not know whether 'Phosphorus' refers to Hesperus or to Mars, and that he does not know whether the planet that can be seen in the morning is Hesperus or Mars; but one could object that another thing that the mayor does not know is whether Phosphorus is Hesperus or Mars. Our description of the case simply does not say anything about what the mayor wonders about, when he wonders whether Phosphorus is Hesperus or Mars.

It is impossible to deny that before the assertion of (1) the mayor entertains two possibilities which can be informally described as the possibility that Phosphorus is Hesperus, and the possibility that Phosphorus is Mars. But we cannot take this informal description seriously. Our commitment to Direct Reference forces us to say that the possibility that Phosphorus is Mars is the possibility that two different planets are the same, and, *pace* Russell, surely we cannot charitably describe the mayor as wondering about *that*. What I urge is that, when we get down to describing more carefully the two possibilities which the mayor entertains, we find that the difference between those possibilities simply comes down to which planet is the one that appears in the morning and is the referent of 'Phosphorus.' While some may find this treatment surprising, it seems empirically adequate, at least for the case under discussion.

6.4.2 Defending Stalnaker's Thesis for Bare Identity Sentences

In the case examined in the previous section, our strategy relied on the fact that the audience of (1) entertained two possibilities, and that the truth of (1) was compatible with only one of those possibilities. It is easy to see that this strategy will work for all cases of informative identity statements in which the audience does not know the referent of at least one of the names in the statement. But one could wonder whether all contexts in which identity statements are asserted with informative results are contexts in which the audience suffers from this kind of ignorance. My goal in this section is to show that they are.

Let me begin by considering a variation of the case of the Mayor in which the audience finds an identity statement informative, and it seems, for all the world, as if they are in a position to interpret the proposition expressed by it:

> THE OPINIONATED MAYOR: Suppose that the facts are as explained in the story of The Mayor, except that this time the mayor does not share the astronomer's hesitation. He is an *opinionated* man, and he is sure that he knows what planet Phosphorus is: It is Mars, of course. If you ask him, he will say that Phosphorus is Mars, and that he does not understand what all the fuss is about.
>
> Nevertheless, he attends Aristarchus' speech. At the end, Aristarchus asserts:
>
> (1) Hesperus is Phosphorus,
>
> which manages to convince the mayor that he was wrong in holding that Phosphorus was Mars.

On the face of it, the assertion of (1) leaves the opinionated mayor in the same state that the mayor is in, in the case discussed in the previous section. But the problem for us is that, this time, previous to the assertion of (1), the opinionated mayor does not acknowledge ignorance concerning the proposition expressed by (1). The following chart summarizes the situation:

THE OPINIONATED MAYOR'S BELIEF STATE[6]

Before	After
• The planet that appears in the morning is Mars	• The planet that appears in the morning is Hesperus
• 'Phosphorus' refers to Mars	• 'Phosphorus' refers to Hesperus
• 'Hesperus' refers to Hesperus	• 'Hesperus' refers to Hesperus
• The planet that appears in the evening is Hesperus	• The planet that appears in the evening is Hesperus

We can now see that the opinionated mayor's initial belief state determines a unique interpretation for (1). On the face of it, this case is a counterexample to Stalnaker's Thesis, and to our general strategy.

But notice that this case is special in another way. In this case, both speaker and hearer disagree about the reference of 'Phosphorus': The opinionated mayor thinks it refers to Mars, while Aristarchus, the speaker, thinks it refers to Hesperus. So both disagree, and disagree about the meaning of the very words they are trying to communicate with. Though it is clear that communication does occur in this case, it is worthwhile to stop and think about how communication is at all possible between agents that have such importantly

[6]From now on we omit the propositions that Hesperus is Hesperus and that Mars is Mars, which belong to all the possibilities described.

different beliefs. My point is that, once we get clear on how there can be communication between these agents, we will see that this is a case in which, after all, Stalnaker's Thesis is borne out.

Peter Gärdenfors has explained how agents that have different beliefs can get down to discussing their differences[7]:

> ...One of the disputants may include A in her belief state K while the other includes $-A$ in her belief state K'. In order to avoid begging the question, both disputants should, at least temporarily, give up their beliefs in A and $-A$, respectively (and all the beliefs that imply those beliefs). From the remainder of their beliefs it may then be possible to carry on a debate, where the task of the debaters is to find arguments that support A and $-A$, respectively... (Gärdenfors (1988), p. 60)

I think that something like this happens whenever two agents who have different beliefs about the meaning of their words discuss who's right. In that case, both agents should update their belief states so that they are compatible with both A and $-A$.

In the case of the opinionated mayor, this means that, if the mayor is to make sense of what the speaker is saying, he must modify his belief state temporarily so that it is compatible with both his beliefs and the astronomer's beliefs. So, when it comes to interpret what the speaker is saying, the real situation is more similar to our first case:

THE OPINIONATED MAYOR'S BELIEF STATE

"In between"	After
Either a: • The planet that appears in the morning is Mars • 'Phosphorus' refers to Mars • 'Hesperus' refers to Hesperus • The planet that appears in the evening is Hesperus *Or b:* • The planet that appears in the morning is Hesperus • 'Phosphorus' refers to Hesperus • 'Hesperus' refers to Hesperus • The planet that appears in the evening is Hesperus	*b:* • The planet that appears in the morning is Hesperus • 'Phosphorus' refers to Hesperus • 'Hesperus' refers to Hesperus • The planet that appears in the evening is Hesperus

My thesis is that, if the opinionated mayor is to make sense of what the astronomer is saying, he must update his belief state to accomodate the astronomer's beliefs. This yields the result that, at the moment of interpreting (1), the mayor would be in the belief state that I called "In between," and which is basically the

[7]I am grateful to James Higginbotham for drawing my attention to the relevance of Gärdenfors writings on this point.

same as the initial state of the mayor in the case described in the previous section. This shows that this case is not, after all, a counterexample to Stalnaker's Thesis, and that, from this point on, it can be dealt with in the same way as the case described in the previous section.

Are there any other counterexamples to Stalnaker's Thesis? Some philosophers have suggested that there are subjects who would find statements like (1) informative, and who would know that both the names 'Hesperus' and 'Phosphorus' refer to. When these philosophers make this point, they seem to have in mind a set-up in which the subject's utterances of 'Phosphorus' are *caused* directly by Hesperus. The following story tries to make this situation vivid:

> THE DEMONSTRATIVE MAYOR: Suppose that the facts are as explained in the story of The Mayor, except that the Alexandrian conference takes place precisely at the time during which Hesperus is the last planet seen in the morning. The morning before Aristarchus' speech, the mayor says: "I know 'Phosphorus' refers to that," while pointing to Hesperus. The conference then goes on, Aristarchus asserts:
>
> (1) Hesperus is Phosphorus,
>
> after which the mayor learns what he wanted to know.

Again, it seems clear that the assertion of (1) was informative, in that it gave the demonstrative mayor some new beliefs—presumably, the same new beliefs as in the original case. But, on the face of it, the fact that the demonstrative mayor said "I know 'Phosphorus' refers to that," while pointing to Hesperus, is a good reason to think that the mayor knows that 'Phosphorus' refers to Hesperus. If this were right, this case would be a counterexample to Stalnaker's Thesis, and to our general strategy.

However, this reasoning cannot be accepted. To begin with, if the demonstrative mayor knows what 'Phosphorus' refers to, it follows that he knows that Phosphorus refers to Hesperus. So when he wonders whether *that* is Hesperus or Mars, what he is wondering, in effect, is whether Hesperus is Hesperus or Mars. But this does not make much sense: Whatever the mayor ignores, he surely knows that Hesperus is Hesperus, and not Mars!

Of course, one has to acknowledge that the demonstrative mayor says a truth when he says that " 'Phosphorus' refers to that," pointing to Hesperus. What we reject is the claim that he knows which truth

—which proposition— he thereby expresses. This does not seem completely unintuitive. After all, if we were to query the mayor, right after he said "I know that 'Phosphorus' refers to that," whether he was referring to Hesperus or to Mars, he would in all likelihood acknowledge that he did not know. The mayor knows that his assertion expressed either a proposition about Hesperus or one about Mars, and he also knows that, whatever proposition he expressed, it was one that was true; it is just that he does not know which proposition he expressed.

My conclusion is that Stalnaker's Thesis is true, that all informative assertions of identity statements occur in contexts in which the audience is not in a position to know the proposition expressed by the statement.

6.5 Belief Attribution and Identity Sentences

6.5.1 The Problem

Let me now turn to the problem of informativeness in connection with belief attributions. Here one might think that Stalnaker's Thesis has no chance whatsoever of being true. Consider, for example:

(2) Luthor believes that Superman is Clark Kent

On the face of it, (2) can be informative even if it is common knowledge between you and me that 'Superman' and 'Clark Kent' both refer to the same person, Superman. But the strategy used in the case of identity statements relies crucially on the hypothesis that informative identity statements are asserted in contexts in which the audience is not in a position to interpret them. On the face of it, that strategy seems useless for cases like (2).

Nevertheless, I will argue that there is good reason to think that essentially the same strategy can be applied to the embedded case. I will defend, in particular, that even if it is common knowledge between you and me that 'Superman' and 'Clark Kent' both refer to Superman, we might nevertheless *not* be in a position to interpret the embedded clause in (2). In arguing for this point, I will appeal to the *Simulation Semantics* for belief attribution which I have presented in the previous chapter. My claim is that if we

assume the (independently motivated) Simulation Semantics for belief attributions, we will be able to see how our indirect account of the informativeness of identity statements can be extended to the case of belief attributions.

6.5.2 The Simple Case

Let us begin by examining the following example of an informative belief attribution embedding an identity sentence:

> LUTHOR AND SUPERMAN: Superman is an alien from the planet Krypton. He leads a double life as Clark Kent, the hapless, bespectacled reporter who works as a journalist at the Daily Planet, and as Superman, the superhero protecting of Metropolis. Luthor is a villain whose plans are always derailed by Superman; because of this, Luthor hates Superman, and is determined to eliminate him. After much investigation, Luthor finally realizes that Superman is Clark Kent.
>
> Jones and Smith are two friends of Superman who are well-acquainted with Luthor, and who have a desire to protect their friend from whatever Luthor may do to him. One day Jones learns that Luthor is going to the Daily Planet. Jones is fearful that Luthor may have seen through Superman's disguise, and that he may be going to the newspaper to eliminate Superman. So he asks his friend: Does Luthor believe that their friend Superman is in the Daily Planet? Smith answers:
>
> (2) Luthor believes that Superman is Clark Kent
>
> After Smith's assertion, it is clear to Jones that Luthor's visit to the Daily Planet is a potential threat for his friend Superman, and decides to warn him.

Intuitively, the assertion of (2) carries news for Jones, in that it tells him something about Luthor's belief state that he did not know before. Our task is to explain how the assertion of (2) managed to do this.

Let us begin by describing how Jones would take the assertion of (2). The most salient feature of the context in which (2) is uttered is that Jones does not know whether Luthor has realized that Superman is Clark Kent. Because of this, our Simulation Semantics implies that Jones does not know whether this is a case in which simulation is necessary to interpret the embedded clause of (2). But Jones is in a position to determine that there are at most two possible interpretations of (2), depending on whether or not Luthor has realized that Superman is Clark Kent.

First, if Luthor has realized that Superman is Clark Kent, then there would be no disagreement between Luthor and Jones. In this case, Jones would interpret the embedded clause of (2) as expressing the proposition which, in Jones' view, 'Superman is Clark Kent' expresses—that is, the proposition that Superman is

Superman.

Second, if Luthor has still not realized that Superman is Clark Kent, then there would be a disagreement between Luthor and Jones about how many people there are: Where Jones only sees one, Luthor sees two. Because of this, it would be necessary to engage in simulation to describe Luthor's belief state. Keeping in mind the conventions introduced in chapter 5, the simulation of Luthor's belief state would contain the following propositions:

SIMULATION OF LUTHOR'S BELIEF STATE (BY JONES)

- 'Superman' refers to $\text{Superman}_{superhero}$
- 'Clark Kent' refers to $\text{Superman}_{reporter}$
- The superhero protecting Metropolis is $\text{Superman}_{superhero}$
- The hapless, bespectacled reporter who works for the *Daily Planet* is $\text{Superman}_{reporter}$
- The superhero protecting Metropolis is not the same as the hapless, bespectacled reporter who works for the Daily Planet

This is easy to see, since, on this possibility, the case of Luthor is essentially like the case of Lois, examined in the previous chapter. This simulation implies that the sentence 'Superman is Clark Kent' expresses (3):

(3) $\text{Superman}_{superhero}$ is $\text{Superman}_{reporter}$

An important point, for us, is that Luthor cannot believe (3). If you think that belief is a relation to propositions, then (3) specifies a proposition that says that two different people are the same. This proposition cannot be true, and surely Luthor does not commit the mistake of believing it. On the other hand, if you are a friend of modes of presentation, then (3) cannot be true either. Here the important point is that '$\text{Superman}_{superhero}$' denotes a mode of presentation that describes Superman as a person who is the superhero defending Metropolis, and does not work for the *Daily Planet* as a reporter; and that '$\text{Superman}_{reporter}$' denotes a mode of presentation that describes Superman as a person who works as a reporter for the *Daily Planet*, and is not the superhero defending Metropolis. It is clear that Luthor cannot be in the belief relation to (3), since (3) describes an object in incompatible ways. Whatever faults Luthor has, he does not commit the mistake of being in the belief relation to (3).

We thus see that, on the assumption that our simulation semantics is right, we can show that Jones is not in a position to interpret (2): There are two strategies available to Jones to interpret (2), but given that

he does not know whether Luthor believes that Superman is Clark Kent, he does not know which one of them to use. And there is a big difference between using one or another of these strategies: According to the former, what (2) says is that Luthor believes that Superman is Superman, something that is quite likely; while according to the latter, what (2) says is that Luthor believes (3), something that cannot be true. There simply could not be any more difference between these two interpretations. We thus we see that Stalnaker's Thesis is borne out, at least in this case.

But if this is right, how does Jones manage to extract interesting information from (2)? To ascertain this, we will have to get clearer on Jones' belief state, both before and after the assertion of (2). It seems clear that, before the assertion of (2), Jones contemplates two possibilities: Either Luthor has realized that 'Superman' and 'Clark Kent' are coreferential, and that the superhero defending Metropolis is the hapless bespectacled reporter working for the *Daily Planet*, or he has not. The following chart summarizes these possibilities:

JONES' BELIEF STATE, BEFORE (2)

Either a:
- Luthor believes that 'Superman' and 'Clark Kent' refer to Superman
- Luthor believes that the superhero defending Metropolis is Superman
- Luthor believes that the hapless, bespectacled reporter who works for the *Daily Planet* is Superman
- Luthor believes that the superhero protecting Metropolis and the hapless, bespectacled reporter who works for the *Daily Planet* are the same person
- (2) says that Luthor believes that Superman is Superman

Or else b:
- Luthor believes that 'Superman' refers to $\text{Superman}_{superhero}$
- Luthor believes that 'Clark Kent' refers to $\text{Superman}_{reporter}$
- Luthor believes that the superhero defending Metropolis is $\text{Superman}_{superhero}$
- Luthor believes that the hapless, bespectacled reporter who works for the *Daily Planet* is $\text{Superman}_{reporter}$
- Luthor believes that the superhero protecting Metropolis and the hapless, bespectacled reporter who works for the *Daily Planet* are different people
- (2) says that Luthor believes that $\text{Superman}_{superhero}$ is $\text{Superman}_{reporter}$

As argued above, each of the possibilities Jones entertains has different consequences concerning the interpretation of (2): Possibility (a) implies that (2) says that Luthor believes that Superman is Superman, and possibility (b) implies that (2) says that Luthor is in the belief relation to (3). This means that, as we saw above, possibility (b) implies that (2) cannot be true.

We can now apply the same reasoning that we applied in the case of bare identity statements. Jones is not in a position to know the proposition expressed by (2), but he does know that Smith has asserted (2). Because he believes that Smith is cooperative, he believes that Smith believes himself to have expressed a

truth. Because he trusts Smith, he believes that Smith has asserted a truth. Once Jones learns this, he can rule out possibility (b), since possibility (b) implies a reading of (2) which simply cannot be true. Thus Jones can figure out that Luthor has realized that the superhero defending Metropolis is the hapless, bespectacled reporter who works for the *Daily Planet*.

What is really crucial for our strategy is the claim that each of the two possibilities that Jones entertains about Luthor's belief state would result in a different truth value for (2): Were it not for the fact that (2) is true only in one of the two possibilities that Jones entertains at the beginning, the fact that (2) has been asserted would not offer Jones any possibility of deriving interesting information. For this reason, our appeal to the simulation semantics is really crucial, for it is our justification for the claim that (2) is true only in one of the two possibilities that Jones entertains.

6.5.3 Ontological and Semantical Disagreement

In the case just discussed, it was assumed that there was an ontological disagreement between speaker and audience, on the one hand, and the subject of the attribution, on the other. Our Simulation Semantics makes a distinction between cases in which there is an ontological disagreement between speaker and subject, and cases in which there is a merely semantical disagreement (see chapter 5, section 5). It would be interesting to see how our theory would deal with the case of Luthor, if the disagreement between Jones, Smith and Luthor were a merely semantical disagreement. The following is an example of this situation:

> LUTHOR, PERRY WHITE AND SUPERMAN: Suppose that the facts are as explained in the story of LUTHOR AND SUPERMAN, except for the following. When the story begins, Luthor still believes that Superman and Clark Kent are two different people, but in the following, peculiar way: He gives every sign of believing that Superman really is Perry White, the chief editor of the *Daily Planet*[8].
>
> After some time, Jones fears that Luthor may have realized his mistake, and come to believe instead that Superman really is Clark Kent, and not Perry White. He asks Smith, who says:
>
> (2) Luthor believes that Superman is Clark Kent,
>
> which tells Jones what he wanted to know.

In this case, there is no ontological disagreement between Jones, Smith and Luthor. At the beginning of the story, they all believe that there are two different people, Perry White and Clark Kent, and they only

[8]This is basically one of the cases discussed in §1.5, what we called a *mixed case*.

disagree about which one of them is the referent of 'Superman,' and which is the one that is the person defending Metropolis; this is a disagreement about the referent of 'Superman' and about the facts, but not a disagreement about which people really exist.

In chapter 5 I argued that, whenever there is a merely semantical disagreement between speaker and audience, and the subject of the attribution, the speaker and the audience are not *forced* to simulate the subject to describe her belief state. Because there is agreement between them about which objects really exist, speaker and audience can describe the belief state of the subject without need to put themselves in the shoes of the agent. Nevertheless, I drew attention to the fact that there were cases in which speaker and audience could nevertheless choose to describe the belief state of the subject by *simulating* her, and using words as the subject herself would use them. One consequence of this proposal is that, whenever we describe the beliefs of a subject with whom we have a merely semantical disagreement, the *that*-clause could be interpreted in two ways, depending on whether or not we choose to simulate the subject of the attribution. I concluded that speaker and audience rely on contextual clues to resolve this ambiguity.

Jones does not know whether or not Smith is simulating Luthor, when he asserts (2). But I think Jones can work out that he is. Suppose first that when Smith asserts (2), he is not simulating Luthor. Then the proposition expressed by (2) is determined by Jones' and Smith's own beliefs about the sentence 'Superman is Clark Kent,' and those beliefs are not affected by what Luthor believes. Thus, if Smith were not simulating Luthor, the assertion of (2) could not tell Jones whether Luthor has realized that Superman is Clark Kent, and not Perry White.

On the other hand, if Smith were simulating Luthor, then the interpretation of (2) would change, depending on whether or not Luthor has realized that Superman is Clark Kent, and not Perry White. As we have seen, if Luthor has realized that Superman is Clark Kent, then (2) would say that Luthor believes that Superman is Superman, something that is surely true. But if Luthor has not realized that Superman is Clark Kent, then (2) would say that Luthor believes that *Perry White* is Clark Kent (since, on this hypothesis, Luthor believes that the name 'Superman' refers to Perry White), which is a necessary falsehood, and something that Luthor surely does not believe. Therefore, if Smith were simulating Luthor, he could use his assertion of (2) to communicate to Jones that Luthor realized that Superman is Clark Kent, and not Perry

White.

The situation seems to be this: on the hypothesis that Smith is not simulating Luthor, (2) does not communicate anything to Jones; while on the hypothesis that Smith is simulating Luthor, (2) does communicate to Jones something that he did not know before. The assumption that Smith is a cooperative speaker who wants to communicate something interesting to Luthor would then allow Jones to draw the conclusion that Smith is, after all, simulating Luthor, since this is the only hypothesis on which (2) communicates something interesting. Once Jones reaches this conclusion, the case becomes basically the same as the one discussed in the previous section, and can be handled in a similar fashion.

6.5.4 Stalnaker's Thesis and Belief Attribution

At the beginning of this section (§6.5.1) we noted that Stalnaker's Thesis seemed to have no plausibility, in the case of belief attribution. What we have discovered is that this is not true. Every belief attribution $\ulcorner A$ believes that X is $Y\urcorner$ occurs in a context in which there is an open question over whether A thinks that X and Y are the same, and that the names X and Y are coreferential. The Simulation Semantics (plus some assumptions about how rational, cooperative speakers communicate with each other) implies that those are contexts in which there will be some uncertainty about how the attribution will be interpreted, which is just what Stalnaker's Thesis says. Thus we see how, with the help of the Simulation Semantics, we can show how Stalnaker's Thesis is borne out in the case of belief attribution too.

6.5.5 Simulation Semantics, Ignorance and Identity

At the end of the previous chapter (§5.8) we noticed that the Simulation Semantics seemed to have two sorts of problems. First, in cases of attributions uttered in contexts in which the audience does not know whether or not they are in disagreement with the subject of the attribution, the Simulation Semantics seems to imply that those audiences will be unable to interpret the attribution in question, which does not seem right. Second, in cases of attributions embedding identity sentences, the Simulation Semantics seems to imply that those attributions are trivial, while they seem to be news to the audience.

The theory presented in this chapter shows how these two difficulties can be solved at the same time.

To begin with, we have shown how that the audience of an informative belief attribution embedding an identity sentence is not in a position to interpret the attribution in question, before the statement is made. Typically, before the statement is made, the audience will entertain two different possibilities about the belief state of the subject, and each of those possibilities will suggest a different interpretation of the relevant attribution. Nevertheless, before the statement is made, the audience is not able to tell which one of those interpretations is the right one.

We have also shown how the audience can nevertheless extract some valuable information, in situations like this one. The general idea is that the attribution will be true only on one of the possibilities that the audience entertains. The fact that the statement is made signals to the audience that the speaker believes that this possibility is the one that is actual. The audience can then gain a new belief by adopting this belief of the speaker. We have shown in detail how this process works in the case of attributions embedding identity sentences. It is easy to see how it would apply to other cases, like the case of LUTHOR AND SUPERMAN discussed in §5.8, and which can be safely left to the reader as an exercise.

6.6 An Objection: Speech Reports and Identity Sentences

According to our account, an assertion of 'Hesperus is Phosphorus' expresses the trivial proposition that Hesperus is Hesperus, and one of 'Hammurabi believes that Hesperus is Phosphorus,' the equally trivial proposition that Hammurabi believes that Hesperus is Hesperus. Assertions of those sentences are informative allright, not because of their content, but because of the inferences that their audiences are willing to draw from the fact that the assertions have been made.

One could object at this point that that is not the way we talk in English. We tend to say, for example, that when the astronomer said that Hesperus is Phosphorus, he said something *interesting*. Our theory does not take this way of talking at face value, since it assigns a trivial proposition to the astronomer's assertion. But it could be urged that it must provide, at least, some explanation of why we talk about the informativeness of English sentences in the way we do.

Actually, the above suggests two different difficulties, worth distinguishing. One is the claim that (4):

(4) The mayor said that Hesperus is Phosphorus

can be informative. Another is the claim that (5):

(5) What the mayor said when he asserted 'Hesperus is Phosphorus' was informative

is true. The challenge for us is to explain how (4) can be informative, and (5) true, in a way that is compatible with our general theory of the informativeness of identity sentences.

Our proposal to explain the informativeness of (4) consists in explaining the functioning of *that*-clauses in speech reports in a way completely parallel to the functioning of *that*-clauses in belief reports. As we sometimes need to engage in simulation to explain what an agent believes, we may also need to engage in simulation to explain what an agent says, whenever what she believes is sufficiently different from our own beliefs. Once this is granted, it is easy to see that the explanation of the informativeness of (4) can go like the case of belief attributions. Any informative assertion of (4) will take place in a context in which we are entertaining two possibilities about what the mayor said, and the truth of (4) will be compatible with only one of them. The fact that (4) was asserted will indicate to the audience which of these possibilities is the one that obtains.

The problem about (5) is more challenging. On the face of it, (5) says that there is one proposition that the mayor asserted, and that *that* proposition is informative. And this is trouble, because on our view, the proposition that the mayor asserted is the proposition that Hesperus is Hesperus, which is not informative.

I nevertheless think that the problem does not lie with the theory defended here, but with phrases of the form ⌜What *X* said⌝, which are notoriously promiscuous. In this connection, David Lewis has remarked:

> ...Unless we give it some special technical meaning, the locution 'what is said' is very far from univocal. It can mean the propositional content, in Stalnaker's sense (horizontal or diagonal). It can mean the exact words. I suspect that it can mean almost anything in between... (Lewis (1980), p. 97)

We can even add some more to Lewis' list. For example, sometimes speakers report on what was conversationally implicated as if it were what was said. This is by no means rare. Suppose you ask: 'When has Jones been on time?', meaning to imply that he has never been. I may then say that what you said was true,

right and interesting, even if, strictly speaking, you did not *say* anything, but rather asked a question. What is going on seems to be this: My phrase 'what you said' does not pick out what you strictly and literally said, but rather some other proposition that was salient in the context in which you were speaking.

Something similar is true of speech that involves some fictional element. Mark Crimmins has given the following example. Suppose that we want to describe Ann's cleverness and modesty by comparing her to someone else, but that no actual person will do. Then I can say:

(6) Ann is as clever as Holmes and more modest than Watson

According to Crimmins, *what I said* could afterwards be correctly described as the proposition that Ann is very clever and modest[9]. Nevertheless, the proposition that I strictly and literally said was one that, first, contained some reference to the fictional characters Holmes and Watson, and second, was probably uttered in a context of make-believe. However, on the face of it, when we say that (6) says that Ann is very clever and modest, we do not say anything about either of those things. Therefore, our phrase 'what I said' does not denote the proposition that I strictly and literally asserted when I asserted (6).

These examples suggest that reports about what is said do not always track the information semantically encoded in some assertion. If Lewis is right, our reports about what is said sometimes track the words used in a certain assertion, rather than the proposition expressed by it. Other times they track Stalnaker's diagonal proposition, rather than the ordinary horizontal proposition expressed in the assertion. Other times they track the content of a conversational implicature, rather than the content of the assertion in question (if there is one). And other times they track the epistemic value of speech about fictional characters, rather than the proposition strictly expressed by those assertions.

The upshot is this: Reports about what is said may track other things, besides the proposition expressed by an assertion. And this is, in our view, what is going on with (5). (5) is true, but in our view, the denotation of the phrase 'what the mayor said when he asserted (1)' does not denote the proposition semantically encoded in the sentence the mayor asserted—that is, the proposition that Hesperus is Hesperus. Rather, it denotes the proposition which, on the occasion of the mayor's assertion, his audience learnt: That Hesperus

[9]See Crimmins (1998).

is the planet seen in the morning, that the planet seen in the morning is the planet seen in the evening, and that 'Phosphorus' denotes Hesperus.

Of course, there is much that remains to be done until this sketch of the semantics of phrases like 'what is said' can be taken seriously; but even at this preliminary stage, I think that it cannot be overlooked, for it provides us with a way of explaining how our indirect account of the informativeness of identity statements is compatible with the truth of (5).

6.7 A Comparison with Stalnaker's Diagonalization

In his paper 'Assertion,' Robert Stalnaker presented his so-called *diagonalization* theory, a theory that could be used to explain, among other things, the informativeness of identity statements. In a later paper, he applied the same device to explain, among other things, the informativeness of belief attributions embedding identity sentences[10]. Stalnaker's theory bears some interesting similarities and differences to the theory presented here, and it would be interesting to compare them.

6.7.1 Outline of Diagonalization

Stalnaker's diagonalization strategy assumes what I have been calling STALNAKER'S THESIS, which says that informative assertions of sentences containing identity statements always occur in contexts in which the audience is not in a position to know the proposition expressed by the statement. But the diagonalization strategy makes a different proposal about how communication proceeds in those circumstances. Simplifying considerably, the idea is this. In any context in which there is a conversation going on, speaker and audience have certain presuppositions about how the utterances made in the conversation are to be interpreted[11]. These presuppositions determine a set of possible worlds that I am going to call the *Interpretation Set*, and which contains all and only the possible worlds that are compatible with the presuppositions of speaker and audience[12]. By using the notion of an Interpretation Set, we can now define the notion of a

[10] See Stalnaker (1978) and Stalnaker (1987).

[11] For the notion of *presupposition*, see Stalnaker (1974).

[12] The notion of an *Interpretation Set* is a simplification that I introduce only for ease of exposition. Stalnaker distinguishes between the main and the derived context, and argues that one or another, or both of these contexts, can be relevant to the interpretation of an

diagonal proposition:

> DIAGONAL PROPOSITION: The diagonal proposition of S in C is the proposition that is defined for all and only worlds in the Interpretation Set of C, and assigns to each world w in the Interpretation Set of C the truth value that S would have, if S were to be asserted in w

The gist of the diagonalization strategy is that, in contexts in which an utterance is made and the audience is in a belief state that determines more than one interpretation for the utterance, they should assume that the the proposition expressed by the utterance is the diagonal proposition associated with that utterance.

The diagonalization strategy can handle all the cases discussed in the foregoing, and can do it at least as well as our indirect account. Let me illustrate this briefly. Remember, to begin with, the case of the mayor (§6.4.1). Prior to the assertion of 'Hesperus is Phosphorus,' the mayor entertains two possibilities, which have different consequences about the proposition expressed by 'Hesperus is Phosphorus:'

THE MAYOR'S BELIEF STATE

Before	After
Either a: • The planet seen in the morning is Hesperus • 'Phosphorus' refers to Hesperus *Or b:* • The planet seen in the morning is Mars • 'Phosphorus' refers to Mars	a: • The planet seen in the morning is Hesperus • 'Phosphorus' refers to Hesperus

Therefore, the Interpretation Set associated with 'Hesperus is Phosphorus' will contain two sets of worlds, those in which possibility *a* is realized, and those in which possibility *b* is realized. Because (a) and (b) determine different interpretations for 'Hesperus is Phosphorus,' the diagonalization strategy implies that, in this case, 'Hesperus is Phosphorus' expresses the diagonal proposition. To calculate this diagonal proposition, let us take a look at the following chart, whose horizontal rows represent the proposition that 'Hesperus is Phosphorus' would express at worlds *a* and *b*:

	a	*b*
a	T	T
b	F	F

utterance. My Interpretation Set runs both contexts together. Nothing in what follows hinges on this simplification.

It is now easy to see that the diagonal proposition associated to 'Hesperus is Phosphorus' in this context is the proposition that assigns truth to *a* and falsehood to *b*. The diagonalization strategy then says that the astronomer's assertion of 'Hesperus is Phosphorus' expressed this diagonal proposition. And this is interesting, because the diagonal proposition rules out possibility *b*, and, as I argued in the preceding, the astronomer's assertion has the effect of allowing the mayor to rule out (b).

Note well that one strength of the diagonalization strategy is that it can accommodate Frege's point that the astronomer's assertion does not communicate just linguistic information. In the case of the mayor, the diagonal proposition associated with the astronomer's assertion encodes both the fact that 'Phosphorus' refers to Hesperus and not to Mars, and the fact that the planet seen in the morning is Hesperus, and not Mars.

In a subsequent paper, Stalnaker explained how the diagonalization device could be applied to solve the problems surrounding belief attributions embedding identity sentences. As we saw in chapter 4, Stalnaker assumes that for an attribution like (2):

(2) Luthor believes that Superman is Clark Kent,

the denotation of the *that*-clause is determined with respect of what he calls the *derived context*, which contains all the possible worlds which, for all speaker and audience believe, are compatible with what Luthor believes. Therefore, the context in which (2) is asserted is a context in which the Interpretation Set will contain all and only worlds which, for all speaker and audience presuppose, are compatible with what Luthor believes. As we saw earlier, previous to assertion of (2), the audience entertains two possibilities about the nature of Luthor's belief state:

LUTHOR'S BELIEF STATE, BEFORE (2)

Either a: • 'Superman' and 'Clark Kent' refer to Superman • The superhero defending Metropolis is Superman • The hapless, bespectacled reporter who works for the *Daily Planet* is Superman • The superhero protecting Metropolis and the hapless, bespectacled reporter who works for the *Daily Planet* are the same person
Or else b: • 'Superman' refers to $\text{Superman}_{superhero}$ • 'Clark Kent' refers to $\text{Superman}_{reporter}$ • The superhero defending Metropolis is $\text{Superman}_{superhero}$ • The hapless, bespectacled reporter who works for the *Daily Planet* is $\text{Superman}_{reporter}$ • The superhero protecting Metropolis and the hapless, bespectacled reporter who works for the *Daily Planet* are different people

Each of these possibilities determines a different interpretation for the sentence 'Superman is Clark Kent:' According to possibility (a), 'Superman is Clark Kent' expresses the proposition that Superman is Superman, while according to possibility (b), it expresses the proposition that $\text{Superman}_{superhero}$ is $\text{Superman}_{reporter}$. Therefore, this will be a case in which the *that*-clause of (2) denotes the diagonal proposition associated with the utterance of 'Superman is Clark Kent.' To see what the diagonal proposition will be, let us take a look first at the following chart, which represents the propositions 'Superman is Clark Kent' expresses, relative to *a* and *b*:

	a	*b*
a	T	T
b	F	F

It is now easy to see that the diagonal proposition associated with 'Superman is Clark Kent' in this context is the proposition that is true at *a* and false at *b*. Again, this is interesting, for this proposition is incompatible with one of the possibilities about Luthor's belief state that Jones entertains, possibility (b). The diagonalization strategy would therefore manage to explain the evolution of Jones' belief state caused by the assertion of (2).

Basically, the diagonalization handles all the examples that our indirect account can handle, and does it in a very similar way. Both theories agree on Stalnaker's Thesis, the thesis that the typical audience of an informative assertion containing an identity sentence is not in a position to interpret the proposition expressed by the statement. Both theories can respond to Frege's objection that identity statements com-

municate non-linguistic information. Both theories can be applied to the case of bare identity sentences, and to the case of identity sentences embedded in belief attribution.

The only respect in which the diagonalization strategy and the indirect account I have presented disagree is on the issue of the *proposition expressed* by identity sentences. The diagonalization strategy implies that the proposition expressed by identity statements, and denoted by *that*-clauses embedding identity sentences, is the diagonal proposition associated with the assertion. The indirect account that I have been defending implies that the proposition expressed by identity statements, and denoted by *that*-clauses embedding identity sentences, are trivial propositions that assert the identity of an object with itself. Let us now run through a couple of arguments that test this difference between the diagonalization strategy and our own account.

6.7.2 Speech Reports Revisited

One argument that could be made in favor of diagonalization, and against our theory, is that diagonalization accords better with the way we talk about informativeness. For example, in the previous section we remarked that, with respect to the case of the mayor, it would seem natural to say:

(5) What the mayor said when he asserted 'Hesperus is Phosphorus' was informative

Our theory accomodates the truth of (5), but only by means of some fancy footwork about the denotation of the phrase 'what is said.' Diagonalization, on the other hand, can explain the truth of (5) easily: Since the diagonal proposition associated with 'Hesperus is Phosphorus' is a contingent proposition, the friend of diagonalization can let the phrase 'what the mayor said' denote the proposition which, according to diagonalization, the mayor strictly and literally said. This result is very satisfying, and it definitely seems to tell in favor of the diagonalization strategy.

Nevertheless, there are other kinds of reports which our theory can handle easily, but which are problematic for the diagonalization strategy. For example, suppose that in the philosophy conference at Alexandria it is a common assumption that the Earth is flat. Then all the worlds in the context set at the moment of the assertion of (1) will be worlds in which the Earth is flat. In this situation it would still seem natural

to report on what the astronomer said by saying:

(7) What the astronomer said when he asserted 'Hesperus is Phosphorus' was true

In order for (7) to be true, what the astronomer said has to be true at the actual world. But it is not clear that the diagonalization strategy can accommodate this. The diagonal proposition associated with the astronomer's assertion is defined only for worlds in the Interpretation Set of that context, and the Interpretation Set in that context contains only worlds in which the Earth is flat. Because in the actual world the Earth is round, the diagonal proposition associated with the astronomer's assertion of 'Hesperus is Phosphorus' is not defined for the actual world. This result makes it difficult for diagonalization to explain how (7) is true. On the other hand, our theory can accomodate the truth of (7) quite easily, since, in our view, what the astronomer strictly and literally said was that Hesperus is Hesperus, and this proposition is true at the actual world.

To solve this challenge, the friend of diagonalization will have to recur, in all likelihood, to Lewis' suggestion about the promiscuity of phrases of the form 'what is said.' The idea would be to provide a semantics for this expression according to which what 'what the astronomer said' denotes, in the context described above, is not the proposition that the astronomer strictly and literally said, but some other proposition which is true at the actual world and is moreover salient in the context in which the astronomer made his assertion. (The proposition that Hesperus is Hesperus could certainly play this role.) But if the friend of diagonalization pursues this line, then he is in the same situation as the friend of our indirect account, in that both need to do some fancy footwork to accomodate certain speech reports.

6.7.3 Lau's Problem

In his Ph.D. dissertation, Joe Lau raised a different problem for the diagonalization strategy, and in particular for its application to the case of identity sentences embedded in belief attribution. According to Stalnaker, the denotation of the *that*-clause of a belief attribution is determined with respect to the derived context, the set representing what speaker and audience know about the beliefs of the subject of the attribution. Lau pointed out that this strategy can yield incorrect results, when coupled with diagonalization.

As an illustration of Lau's argument, consider the following example:

> FRENCH LUTHOR: Suppose that the story of Luthor is as explained in the case of Luthor and Superman, except for the fact that this time the Superman story occurs in Paris, where everyone (Luthor included) speaks French, and no one speaks English. Suppose that this is common knowledge between Jones and Smith, who are talking, in English, about French Luthor's beliefs. One day, Jones becomes fearful that French Luthor may have seen through Superman's disguise. He asks his friend Smith, who says, in English:
>
> (8) French Luthor believes that Superman is Clark Kent
>
> After hearing to Smith's assertion, Jones learns that French Luthor has indeed guessed Superman's hidden identity.

Lau argued that, in this case, diagonalization cannot explain how (8) can be true. To see this, begin by considering a world that we will call w^*. In w^*, the linguistic facts are such that the sentence 'Superman is Clark Kent' expresses a necessary falsehood. The rest of the facts are as French Luthor takes them to be: There is one person who is both the superhero defending Paris, and the hapless, bespectacled reporter who works for the Parisian branch of the *Daily Planet*.

It seems clear, in the first place, that French Luthor's beliefs are compatible with w^*. In particular, since French Luthor does not speak English, any hypothesis about what English sentences mean is compatible with what French Luthor believes. Therefore, any belief attribution that attributes to French Luthor a proposition incompatible with w^* is *ipso facto* false.

In the second place, it is also clear that w^* is part of the Information Set of the context in which (8) is asserted. In particular, remember that it is common knowledge between Jones and Smith that French Luthor does not speak English. Therefore, the Information Set will contain possible worlds that reflect all possibilities about the proposition expressed by the English sentence 'Superman is Clark Kent,' and in particular the possibility that it expresses a necessarily false proposition.

Intuitively, (8) is true. Because (8) is asserted in a context in which it is common knowledge that there are several hypothesis about how French Luthor takes the English sentence 'Superman is Clark Kent,' this is a context where the diagonalization strategy implies that the *that*-clause of (8) denotes the diagonal proposition associated with 'Superman is Clark Kent.' But notice that the truth value of the diagonal at w^* is false: By assumption, 'Superman is Clark Kent' expresses a necessary falsehood at w^*, a proposition that is

false at every possible world and therefore also at w^*. Therefore the diagonalization strategy has the consequence that (8) is false, because (8) says that French Luthor believes the diagonal proposition associated with 'Superman is Clark Kent,' and this proposition is incompatible with w^*, which is a possibility that is compatible with French Luthor's beliefs. But the problem is that, intuitively, (8) is true[13].

The problem arises because the diagonal of the propositional concept is defined as a function of what Luthor believes, and his lack of knowledge of English gets in the way, so to speak. On the face of it, what we need is an account of the denotation of the *that*-clause of (8) that is *immune* to French Luthor's lack of knowledge of English. Our indirect account, coupled with the Simulation Semantics, can provide just that. (8) is asserted in a context in which there is an open question about whether Luthor thinks that the superhero defending Paris and the hapless, bespectacled reporter who works for the Parisian branch of the *Daily Planet* are the same. If he thinks they are the same, then no simulation is necessary to describe his belief state. If he thinks they are different, then simulation will be called for. Because he thinks that they are the same, no simulation is necessary to interpret it, and what (8) says is that Luthor is in the belief relation to the proposition that Superman is Superman. This proposition is certainly compatible with w^*. Thus we see that our theory really has no trouble with (8).

It is not clear how the diagonalization proposal could be modified to meet this argument. The problem for the diagonalization strategy arises because the denotation of the *that*-clause of (8) is determined with respect to what I called the *Interpretation Set*, which represents what speaker and hearer presuppose to be French Luthor's beliefs. In particular, the Interpretation Set represents the speaker and hearer's common knowledge of the fact that French Luthor does not speak English, which ends up getting in our way, so to speak. It is worth noting, however, that the diagonalization proposal is compatible with other proposals about what information there is represented in the Interpretation Set in cases of attributions to people like French Luthor. Perhaps diagonalization could be saved by proposing that the Interpretation Set, in the case of Luthor, contains different information.

The notion of simulation that I have defended in the previous chapter could be used to provide some relief, but, I think, not enough. The idea behind the Interpretation Set is that it must represent all the in-

[13]For the original argument, see Lau (1994), chapter 1, esp. pp 24–28.

formation that is relevant to the interpretation of a sentence, be it bare or embedded in a *that*-clause. Our proposal in the previous chapter is that there are two relevant possibilities to interpret the *that*-clause of (8), depending on whether or not it is necessary to simulate Luthor. Very schematically, the possibilities are the following (see §5.3.2 for details on the similar case of CARLOS):

POSSIBILITY *x*:
•The superhero defending Paris and the hapless, bespectacled reporter who works for the *Planet* are the same people •'Superman' refers to Superman •'Clark Kent' refers to Superman
POSSIBILITY *y*:
•The superhero defending Paris and the hapless, bespectacled reporter who works for the *Planet* are different people •'Superman' refers to $\text{Superman}_{superhero}$ •'Clark Kent' refers to $\text{Superman}_{reporter}$

We could now claim that the Interpretation Set contains all and only worlds that are compatible with these two possibilities, on the grounds that, according to the Simulation Semantics, these are the possibilities that are relevant to the interpretation of (8). This would have the effect that w^* is no longer included in the Interpretation Set. Remember that w^* had two main features: At w^*, the superhero defending Paris and the hapless reporter are the same person, and moreover 'Superman is Clark Kent' expresses a necessary falsehood. The first one of these possibilities is incompatible with possibility (y), while the second is incompatible with possibility (x). Because the Interpretation Set for (8) would contain, on this view, all and only possible worlds that are compatible with either (x) or (y), and w^* is compatible with neither, w^* would be excluded from the Interpretation Set. This is definitely a relief, since the problem was caused by the presence of the world w^* in the Interpretation Set.

However, the relief is only temporary. According to this revised understanding of the diagonalization strategy, the *that*-clause of (8) would denote the diagonal proposition associated with 'Superman is Clark Kent,' relative to the Interpretation Set just described. The following chart allows us to work out this proposition:

	x	y
x	T	T
y	F	F

Thus we see that the diagonal proposition of 'Superman is Clark Kent' is the proposition that assigns truth to x and falsehood to y.

Nevertheless, the hypothesis that (8) says that Luthor believes this diagonal proposition cannot explain the evolution of the belief state of the audience of (8). To see this, begin by considering how we could represent, very schematically, the belief state of the audience before the assertion of (8):

EITHER *a:*
•French Luthor believes that the superhero defending Paris and the hapless, bespectacled reporter who works for the *Planet* are the same people •French Luthor does not know English
OR ELSE *b:*
•French Luthor believes that the superhero defending Paris and the hapless, bespectacled reporter who works for the *Planet* are different people •French Luthor does not know English

The problem can now be spelled out as follows. In the first place, note that the belief state of the audience, prior to the assertion of (8), includes a world —call it w'— where, first, French Luthor believes that the superhero defending Paris and the hapless, bespectacled reporter who works for the *Planet* are *different* people, and second, where French Luthor believes that 'Superman is Clark Kent' means that the moon is made of cheese. This is so because, first, the audience does not know whether French Luthor has realized that the superhero defending Paris is the reporter who works for the *Planet*; and second, the audience knows that French Luthor does not know English, and therefore that any possibility about what any English sentence means is compatible with what French Luthor believes.

Intuitively, the assertion of (8) should be incompatible with w'. This is so because, once (8) has been asserted, we learn that French Luthor knows that the superhero defending Paris is the reporter working for the *Planet*, and w' is a world in which French Luthor believes that the superhero defending Paris and the reporter working for the *Planet* are two different people.

The problem is that, on this revised understanding of the diagonalization strategy, the proposition expressed by (8) is compatible with w'. The reason is that the diagonal proposition associated with 'Superman is Clark Kent' in this context is incompatible *only* with worlds where Luthor believes the proposition that

is true at x and false at y, and that would not allow us to rule out the world w'. In particular, learning that Luthor believes a proposition that is false at y would allow us to rule out all those worlds in which Luthor believes *both* that the superhero defending Paris is different from the reporter working for the *Planet*, and in which 'Superman is Clark Kent' expresses the proposition that $\text{Superman}_{superhero}$ is $\text{Superman}_{reporter}$. But w' is not one of these worlds, since in w' 'Superman is Clark Kent' expresses the proposition that the moon is made of cheese. Thus this revised understanding of diagonalization cannot explain how the audience of (8) manages to rule out w', and therefore also how the assertion of (8) enables the audience to learn that French Luthor has realized that the superhero defending Paris is the reporter who works for the *Planet*.

It is not clear to me how the diagonalization should be further amended, to deal with this problem. My conclusion is that, while it is true that there are many similarities between Stalnaker's diagonalization theory and the account defended in this chapter, Lau's problem gives us a reason to prefer our account.

Chapter 7

Points of View, Simple Sentences and Substitutivity

7.1 Points of View

In the ordinary course of life, opportunities often arise to consider things *from another point of view* — sometimes it is even fun to do so. But what do we say, when we say that, from the point of view of so-and-so, things are thus-and-so? That is the question that I will address in this chapter.

Let us say that a sentence containing the prefix 'from the point of view of' is a *Point-of-View sentence*, or *POV*, for short. The semantics for POVs presents some interesting features worth studying. To begin with, the prefix can create contexts in which substitution of coreferential names does not preserve truth value. For example, everyone acquainted with the Superman stories knows that (1) is true and (2) is false:

(1) From the point of view of the inhabitants of Metropolis, Superman can fly
(2) From the point of view of the inhabitants of Metropolis, Clark Kent can fly

But (1) and (2) only differ in that coreferential names have been intersubstituted.

The second interesting feature is that 'From the point of view of' is not synonymous with 'believes that.' To see this, consider the following story:

> THE ABDUCTION: A woman from New York is abducted, and no one in Metropolis finds out. The woman, left on some rail tracks outside Metropolis, cries for help. Superman, with his superhearing, hears the cries, puts his superhero costume on, and saves the woman.

Suppose that now we ask whether (3) is true:

(3) From the point of view of the inhabitants of Metropolis, Superman rescued the woman on time

The scenario has been designed so that no one in Metropolis will know about it. The abduction is performed outside Metropolis, the woman is placed on some tracks outside Metropolis, and Superman does not explain to anyone in Metropolis what he is about to do. Given this, the first reaction is surely to say that (3) is false: Since the people in Metropolis do not know what happened, they do not have any opinion about it, and therefore (3) should be false. This reasoning suggests that (3) implies (4):

(4) The inhabitants of Metropolis believe that Superman rescued the woman in time,

However, I think that (3) does not imply (4). Rather, I think that (4) is an *implicature* of (3), rather than an *implication*. This can be shown by figuring out a context in which the implicature (4) is *cancelled*. Suppose, for example, that the assertion of (4) occurs in the following conversational context:

> We have just heard what happened in ABDUCTION. Let me now ask what happened here, from the point of view of the inhabitants of Metropolis. Of course, no one in Metropolis knows what happened, but let me put that aside. What we want to know is how the people in Metropolis would describe the situation, if they were to know it. From their point of view, who did the rescuing? Was it Superman? Was it Clark Kent? Was it Lois Lane?

I think that we would answer this question by asserting (3), and that this shows that, at least in this context, we think that (3) is true. Because in this context it is common knowledge that (4) is false, we have to conclude that assume that (4) is nothing more than an implicature of (3), which in this particular case is explicitly cancelled by our remark that we are going to "put aside" the fact that the people in Metropolis do not know what happened. Therefore, examples like this show that 'From the point of view of' is not synonymous with 'believes that.'

It would be nice to have a semantics for the prefix 'from the point of view of' that can explain why this prefix can create opaque contexts, and how exactly this prefix is different from 'believes that;' and this is

precisely what I will try to do in this chapter. I will also argue that the semantics I will offer can be applied to solve an interesting puzzle about substitutivity in simple sentences presented by Jennifer Saul in a recent paper.

7.2 The Sentential Theory

On the face of it, POVs report how a person would describe a certain situation. Because one describes situations by using *sentences*, this suggests that POVs report a relation between a person, a situation, and a sentence. Here is an initial proposal:

> SENTENTIAL THEORY: An utterance of ⌜From the point of view of X, P⌝ is true if, and only if, there is a relevant situation S such that X would use ⌜P⌝ to describe S

On this view, (1) says that the inhabitants of Metropolis would use the sentence 'Superman can fly' to describe the state of affairs in their city, which is surely true. On the other hand, (2) would say that the inhabitants of Metropolis would use 'Clark Kent can fly' to describe the situation in their city, which is surely false. Thus this proposal explains how substitution of coreferential names can affect the truth value of a POV: Substitution of coreferential names will change the sentence that is being attributed to the subject of the POV, and the subject may bear different attitudes towards those sentences.

Also, the Sentential Theory accomodates the observation that 'from the point of view of' is not synonymous with 'believes that.' On this view, (3) says that the inhabitants of Metropolis *would* describe THE ABDUCTION by means of the sentence 'Superman rescued the woman in time.' This does not imply that the inhabitants of Metropolis believe that Superman rescued the woman in time.

Now, promising as this proposal may seem, there are several difficulties for it. The first one is this. It is a familiar fact that, even from a single point of view, many situations can be described equally well by more than one sentence. Notice, however, that we cannot assume that the subject of a POV would describe the relevant situation by using the very same sentence that we use to report how things look like from her point of view. There are many things that can get in the way. To begin with, the subject of the POV may not speak our language. Suppose, for example, that the Superman story happened in Paris, and that the people

in Paris were all monolingual speakers of French. Now consider (5) and (6):

(5) From the point of view of the people in Paris, Superman can fly

(6) From the point of view of the people in Paris, Clark Kent can fly

Surely, in this situation, (5) is true and (6) is false. But a problem with the current proposal is that we cannot say that (5) is true if and only if the people in Paris would use the sentence 'Superman can fly' to describe the state of affairs in their city, for they do not know English, and therefore would not use any English sentence to describe that.

Perhaps, on the face of this example, we could introduce the following modification of the Sentential Theory:

> An utterance of ⌜From the point of view of *X*, *P*⌝ is true if, and only if, there is a relevant situation *S* such that *X* would use ⌜*P*⌝ to describe *S*, if *X* knew the language to which *P* belongs

However, it is not clear that this reformulation can assign the right truth value to (6). Those of us who are committed to Direct Reference think that the sentences 'Superman can fly' and 'Clark Kent can fly' express, in English, the same proposition. If the people in Paris knew English, they would presumably know this fact, and therefore would also use the sentence 'Clark Kent can fly' to describe what is going on. Therefore, this proposal would imply that (6) is true, when it seems clearly false.

At this point, one could try to modify the semantics for POVs by invoking the notion of translation instead:

> An utterance of ⌜From the point of view of *X*, *P*⌝ is true if, and only if, there is a relevant situation *S* such that:
>
> (i) Either *X* knows the language of the POV, and *X* would use ⌜*P*⌝ to describe *S*
>
> (ii) Or else *X* does not know the language of the POV, and *X* would use a sentence of *X*'s language which is a proper translation of ⌜*P*⌝ to describe *S*

As I argued above, I do not see any reason why those of us who are committed to Direct Reference could not accept a notion of translation on which 'Superman can fly,' but not 'Clark Kent can fly,' is a proper translation of the French sentence 'Superhomme peut voler;' and similarly, that 'Clark Kent can fly,' but

not 'Superman can fly,' is a proper translation of 'Clark Kent peut voler.' Given this, clause (ii) of this reformulation of the Sentential Account would manage to assign the right truth values to (5) and (6).

Nevertheless, this proposal still has problems. We could imagine a version of this case in which the people in Paris did not have names for the two personalities of Superman, but they still believe that the superhero defending Paris is different from the bespectacled, hapless reporter working for —say— the Parisian branch of the *Daily Planet*. In this situation, they would describe the situation in their city by means of the following sentence:

(7) Le superhero qui defend Paris peut voler

In this case, it seems that we can still use (5) to truly describe how things look like from their point of view. The problem is that there is no sentence which translates 'Superman can fly' which the people in Paris would use to describe that situation. Certainly, (7) is not a proper translation of 'Superman can fly,' but rather of something like 'The superhero defending Paris can fly.' It is not clear to me how the Sentential Theory should be fixed to deal with this problem.

A different problem arises whenever the relevant POV contains demonstratives. For example, it seems clear that if we were to point at Superman, when he is wearing his red cape and his blue tights, and then say:

(8) From the point of view of the people in Metropolis, he can fly,

we would say a truth. But the truth conditions of (8) cannot depend on whether or not the people in Metropolis would use the *sentence* 'he can fly' to describe the situation, for that sentence can be used to describe many different situations, when accompanied by different demonstrations.

A possible amendment of the Sentential Theory to account for this possibility is this:

> An utterance of ⌜From the point of view of X, P⌝ in a context C is true if, and only if, there is a relevant situation S such that X would recognize that our utterance of ⌜P⌝ in C is a good description of S

This alternative account would manage to assign the right truth conditions to (8). However, on this view the opinion of the subject of the POV about the utterance of the complement sentence is crucial to the

truth conditions of POVs, and in occasion this may yield the wrong results. Suppose for example that one day we see Superman coming down the street, wearing his glasses and his business suit. Suppose also that there is a big banner hanging from the sky with an arrow pointing directly at him. In these circumstances, it seems that if we were to say:

(9) From the point of view of the people in Metropolis, Superman can fly and he cannot,

we would say a truth; and moreover that our amendment of the Sentential Theory can handle it.

But now suppose that, in this same scenario, the banner containing the arrow that points at Superman contains, in big letters, the inscription 'Superman is Clark Kent.' If we were to say (9) in this situation, we would still say a truth, since the banner would still raise Superman to saliency, under his reporter mode of presentation. However, our amendment of the Sentential Theory cannot handle this case. It seems clear that, if any of the inhabitants of Metropolis were to witness our utterance, they would immediately learn that Superman and Clark Kent are the same person, and therefore that Clark Kent fly. Thus, the inhabitant of Metropolis put in this situation would no longer acknowledge that the sentence 'Superman can fly and Clark Kent cannot' can describe the situation in his city. By the same reasoning, this person would not acknowledge either that the utterance of 'Superman can fly and he cannot' that is part of our utterance of (9) in that context accurately describes the situation in his city. On the face of it, the truth conditions of (9) cannot depend on whether or not the people in Metropolis would find our utterance of 'Superman can fly and he cannot' an accurate description, were they to witness the utterance. But it is not clear how to further modify the Sentential Theory to avoid this result.

A third problem arises because the Sentential Theory fails to make some distinctions that seem quite intuitive. To see this, consider the following situation:

> THE HELICOPTER: A helicopter flying over Metropolis has engine failure, and starts to fall towards a busy street. When it seems as if many people will die, Superman flies to the scene, sporting his Superman suit. He grabs the helicopter in mid-air, gently deposits it on the street, and flies away. After a few minutes, Superman returns dressed as Clark Kent and asks questions about what happened.

It seems clear that we and the inhabitants of Metropolis do not agree about what happened. From our point of view, a single person both rescued the helicopter and then returned and asked questions. On the other

hand, from the point of view of the inhabitants of Metropolis, *two* different people did the saving and the questioning. Now consider the following sentences, both of which seem true:

(10) From the point of view of the inhabitants of Metropolis, Superman rescued the helicopter, and afterwards Clark Kent started asking questions about what happened

(11) From our point of view, Superman rescued the helicopter, and afterwards Clark Kent started asking questions about what happened

Intuitively, (10) implies that, from the point of view of the inhabitants of Metropolis, it was two different people who did all those things; while (11) implies that, from our point of view, it was the same person who did them all. The problem is that the sentential account cannot account for these implications. According to the sentential account, all (10) and (11) say is that the inhabitants of Metropolis and us would describe THE HELICOPTER by means of the sentence 'Superman rescued the helicopter, and Clark Kent came afterwards asking questions.' This is true of both, but it does not suffice to capture the difference between how the people in Metropolis see what happened. On the face of it, (10) and (11) say something more than what, according to the Sentential Theory, they say.

Perhaps there is a fix that can solve all these difficulties, and is compatible with the spirit of the Sentential Theory, but I do not know what such a fix would be like. In any case I suspect that it would be very hard to make the fix appear natural, and not *ad hoc*. It seems to me a better strategy to scrap the idea that sentences are a relatum of the relation predicated by POVs, and look for another object to play that role.

7.3 The Belief Theory

There is a familiar view according to which belief is a two-place relation between a believer and something that some people take to be a proposition (understood in one or another way), and other people take to be something that they call a mode of presentation. Propositions and modes of presentation are very promising candidates to play the role of the third relatum of the relation predicated by POVs: They are not ambiguous; they do not belong to English, or to any other natural language; a single proposition or mode of presentation can be communicated by different English sentences; and a single English sentence can be associated with several beliefs at the same time. Perhaps we can use them in the semantics of POVs

There immediately arises a problem that I want to put aside, and that is that the debate about which of these views is right is, of course, very controversial. Because I want to remain neutral on it, I will present a semantics for POVs that is compatible with both hypotheses. For the sake of convenience, in what follows I will talk as if the belief were a relation to propositions, but that should be construed as arising out of convenience, rather than arising out of a preference for one of the sides in the debate.

Having said this, here is a preliminary suggestion about how propositions could be used in the analysis of POVs:

> An utterance of ⌜From the point of view of *X*, *P*⌝ in context *C* is true if, and only if, there is a relevant situation *S* and a certain proposition *B* such that *X* would start believing *B*, were *X* to learn about *S*

The challenge for this style of analysis is to explain how the proposition *B*, that plays such a crucial role in it, is determined. Notice that we cannot say that *B* is the proposition that *P* expresses, according to the utterer of the POV: 'Superman can fly' and 'Clark Kent can fly' express the same proposition, according to us; so that this proposal would fail to distinguish the truth conditions of (1) and (2), when we assert them. Neither can we say that *B* is the proposition that *P* expresses according to *X*. To be sure, this proposal would distinguish the truth conditions of (1) and (2)—since the inhabitants of Metropolis do believe that 'Superman can fly' and 'Clark Kent can fly' express different propositions. But this proposal cannot handle cases in which *X* does not have any opinion about the proposition expressed by the complement sentence—as, for example, when the subject of the POV does not speak the language to which the POV belongs.

My proposal will be different from these two. On the face of it, POVs do not tell us precisely what their utterers think, or what their subjects think. Rather, they seem to be urging their utterers, and their audiences, to describe a certain situation from the point of view of someone else. This suggests that the notion of *simulation* developed in the previous chapter can be used to provide a semantics for POVs, in the following way:

> BELIEF THEORY: An utterance of ⌜From the point of view of *X*, *P*⌝ in context *C* is true if, and only if, there is a relevant situation *S* and a certain proposition (or mode of presentation) *B* such that

(i) B is the proposition which, according to the simulation of X by the utterer, $\ulcorner P \urcorner$ expresses in C

(ii) X would start believing B, were X to learn about S

It is easy to see, to begin with, that this proposal manages to accomodate the observation that 'from the point of view of' is not synonymous with 'believes that.' On this proposal, a POV does not say what the subject believes, but rather that he would start believing a certain proposition, where she to know certain facts.

It is also easy to give an explanation of why 'from the point of view of' should give rise to opaque contexts. To begin with, consider (1) and (2):

(1) From the point of view of the inhabitants of Metropolis, Superman can fly

(2) From the point of view of the inhabitants of Metropolis, Clark Kent can fly

We know that the people in Metropolis associate different propositions (or modes of presentation) with 'Superman can fly' and 'Clark Kent can fly.' If we use the convention established in chapter 5 to represent those propositions, we can represent those propositions as follows:

(12) $\text{Superman}_{superhero}$ can fly

(13) $\text{Superman}_{reporter}$ can fly

The upshot is that we disagree with the people in Metropolis over which are the propositions (or modes of presentation) expressed by (1) and (2); therefore, the simulation of the people in Metropolis will contain the information that 'Superman can fly' expresses (12), and that 'Clark Kent can fly' expresses (13). Thus, on the Belief Theory, (1) says is that the people in Metropolis would believe (12) were they to learn about the situation in their city; and (2) says that they would believe (13). Well, they do know about the situation in their city, and they do believe (12), and they do not believe (13). Therefore, this Belief Theory manages to assign the intuitively correct truth conditions to (1) and (2).

The rest of the problematic cases mentioned in the previous section are very similar to the cases of belief attributions discussed in chapter 5, and can be handled in a very similar way. Since it would be tedious and repetitive to go over those cases again, they will be left for the reader as an exercise.

7.4 Substitution and Simple Sentence

7.4.1 Saul's Problem

This semantics for POVs can be applied to solve a puzzle about substitution in simple sentences presented by Jennifer Saul in a recent paper[1]. Saul pointed out that there is a strong intuition that (14) is true and (15) is false:

(14) Clark Kent always arrived at the scene after one of Superman's daring rescues

(15) Superman always arrived at the scene after one of Clark Kent's daring rescues

The existence of this intuition raises a problem on its own right, since there does not seem to be any obvious way of accounting for it. We are all familiar with failures of substitutivity in belief attribution contexts, and many philosophers are willing to agree that, in a sense or another, coreferential proper names can acquire different semantic values *when embedded inside belief attribution contexts*. The problem is that in (14) and (15), 'Superman' and 'Clark Kent' are not embedded within a belief attribution context. Furthermore, it hardly seems as if any of the other words in (14) and (15) can create opaque contexts. Thus explaining why (14) should seem true, and (15) false, is a genuine challenge.

Saul added that this problem has repercussions for the current debate on the semantics of belief attribution. It is agreed that we have the intuition that, for example, (16) and (17) differ in truth value:

(16) Lois believes that Superman can fly

(17) Lois believes that Clark Kent can fly

But there is an ongoing debate about whether that intuition is reliable, and thus whether a systematic semantics for belief attribution should reflect it. Many philosophers and semanticists think that our intuitions are reliable, and that a satisfactory semantics should imply that (16) and (17) have different truth value; I am going to call these people *contextualists*[2]. But the intuition that (14) and (15) differ in truth value raises a dilemma for contextualists, and neither of its horns appears attractive.

[1] See Saul (1997)

[2] Examples of contextualists are the views discussed in chapter 4 and chapter 5

First, suppose that, after substantive philosophical investigation, we conclude that we cannot explain how (14) and (15) can differ in truth value. Then we would have to say that (14–15) do not really differ in truth value, in spite of our intuitions. The problem is that the contextualist would then be hard pressed to explain why our intuitions should be respected in the case of (16–17), but not in the case of (14–15).

Secondly, suppose that we manage to give some sort of explanation of how (14–15) can differ in truth value. There is a good chance that this explanation will be substantially different from the traditional contextualist explanations of how (16–17) can differ in truth value, which all exploit the fact that, in (16–17), the problematic names occur embedded. The problem is that in (14–15) the relevant names are *not* embedded, and therefore this style of explanation does not seem applicable to (14–15). Since one would expect that similar solutions can be given to both problems, there is a considerable chance that if we ever get to explain how (14–15) can differ in truth value, that explanation could be turned into an argument against contextualists.

Neither horn appears promising, and Saul's tentative conclusion is that the existence of pairs like (14–15) is an argument in favor of the position of philosophers like Nathan Salmon and Scott Soames, who think that, in spite of our intuitions, (16) is true if and only if (17) is[3].

I think that Saul's challenge about (14) and (15) can be solved, and that her argument in favor of Salmon and Soames' ideas can be defused, by appealing to the semantics for POVs that I have presented above. My proposal to meet her challenge has two parts. The first part is the claim that, when we feel inclined to say that (14) is true and (15) false, we take them to mean (18) and (19), respectively:

(18) From the point of view of the inhabitants of Metropolis, Clark Kent always arrived at the scene after one of Superman's daring rescues

(19) From the point of view of the inhabitants of Metropolis, Superman always arrived at the scene after one of Clark Kent's daring rescues

The second part consists in explaining how (18) and (19) can differ in truth value. Because, on my view, the problematic names occur embedded, Saul's challenge against contextualism simply disappears.

[3]See Soames (1987a), Soames (1987b), and Salmon (1986)

The second part of my proposal is supported by the semantics for POVs just presented. On the Belief Theory, (18) says that the inhabitants of Metropolis believe (20), and (19) says that the inhabitants of Metropolis believe (21):

(20) Superman$_{reporter}$ always arrives at the scene after one of Superman$_{superhero}$'s daring rescues

(21) Superman$_{superhero}$ always arrives at the scene after one of Superman$_{reporter}$'s daring rescues

The inhabitants of Metropolis do believe (20), and they do not believe (21); this accounts for the difference in truth value between (18) and (19).

What needs argument is the first part of my proposal, the claim that we are inclined to say that (14) is true and (15) is false because we interpret them as (18) and (19). I will now present two different arguments in favor of this claim.

7.4.2 Storytelling

To explain a story, some points of view are better than others. For example, Arthur Conan Doyle wrote much of the Sherlock Holmes stories from the point of view of Holmes' companion, Dr. Watson. Because generally Watson does not learn the drift of Holmes' thoughts until late in the story, this made for quite a good deal of intrigue, which is one of the main attractions of the novels.

This also happens with the Superman stories, although in a different way. Remember that Superman hides from the general public the fact that he leads a private life as the reporter Clark Kent, and that he succeeds in fooling everyone. As a result, everyone in Metropolis thinks that Superman and Clark Kent are two different people. One consequence of this is that, when we explain some episode of the Superman stories in which there is an interaction between an inhabitant of Metropolis and Superman, we will have to keep track of whether the person in question thinks she is interacting with Superman or with Clark Kent; otherwise, we would probably miss the point of the episode.

Perhaps this point could also be made with the help of an example. Everyone acquainted with the Superman stories knows that Lois is in love with Superman, and that she does not know that Superman is, in fact, her workmate Clark Kent. Suppose that one day, Superman, under his Clark Kent identity, invites

Lois to dinner, and that Lois rejects him. Later on, Superman, under his Superman identity, invites Lois to dinner, and this time she accepts. How are we going to explain what happened? Not with (22):

(22) Lois rejected Superman's invitation, but when he tried again, later on, she accepted

Though strictly speaking (22) is true, it is bad storytelling, since (22) does not reflect the fact that Lois believes that each of the invitations came from a different person. To explain the story well, we have to keep track of who Lois thinks that is inviting her each time.

There are many ways to fulfill this desideratum, but the most straightforward and economical way is to explain the Superman stories *directly* from the point of view of the inhabitants of Metropolis. Because the inhabitants of Metropolis believe that Superman and Clark Kent are different people, explaining the stories from their point of view guarantees that our storytelling keeps track of when Lois believes she is being invited by the superhero protecting Metropolis, and when she believes she is being invited by her coworker.

If this is right, then our inclination to say that (14) is true and (15) is false is explained thus. (14) and (15) purport to be some description of the Superman stories and, as I have been arguing, when explaining some episode of the Superman stories, it is better to do so from the point of view of the inhabitants of Metropolis. Therefore, it is natural to interpret (14) and (15) as (18) and (19), respectively.

7.4.3 Ambiguity

The second argument in favor of my claim is based on the observation that, at least in some contexts, the intuition that (14) is true and (15) is false is not the only one evoked by those sentences. Personally, I would describe the experience as follows: when I am asked, out of the blue, about the truth value of (14) and (15), I would hesitate between saying that (14) is true and (15) is false (which is the intuition that Saul drew our attention to), and saying that (14) is true if, and only if, (15) is. On the one hand, I feel that (14) is right, appropriate, to describe some episode of the Superman stories, while (15) is not. This feeling makes me inclined to say that (14) is true and (15) is false. But, at the same time, I also have the opposite feeling:

Because Superman is Clark Kent, I feel that (14) is true if and only if (15) is.[4]

These intuitions are contradictory, but there is no reason why we cannot accomodate both, if it turns out that (14) and (15) are *ambiguous* in the relevant contexts. I will first examine an explanation for that ambiguity provided by Joseph Moore, and I will later present my own, based on the idea that (14) and (15) bear a tacit 'from the point of view of' prefix that can be completed in different ways.

Moore's Explanation

Joseph Moore observed that there are contexts in which (14–15) give rise to contradictory intuitions, and has provided an explanation of why this should be so, based on the idea that (14–15) are ambiguous[5]. First, he argues that the names 'Superman' and 'Clark Kent' are ambiguous between a reading on which they refer to the person, Superman, and another reading on which they refer to *aspects* of Superman. For our purposes, we can take an aspect of Superman to be a set of temporal parts of him. On this view, the *Clark Kent* aspect of Superman is the set of all those temporal parts of Superman in which he assumes his reporter identity; and the *Superman* aspect of Superman is the set of all those temporal parts of Superman on which he assumes his superhero identity.

Next, Moore distinguishes between what he calls *enlightened* and *unenlightened* contexts. The *unenlightened* context is defined as one in which speaker and audience falsely believe that Superman and Clark Kent are two distinct individuals[6]. The *enlightened* context is defined as one in which speaker and audience are aware of the fact that Superman is Clark Kent.

Moore argues that whether (14–15) are asserted in an enlightened or an enlightened context should make a difference to the interpretation of the proper names 'Superman' and 'Clark Kent'. Because in *enlightened* contexts speaker and audience are aware that Superman and Clark Kent are the same person, this forces the interpretation on which 'Superman' and 'Clark Kent' denote the *Superman* and *Clark Kent* aspects of Superman, respectively. On this reading, (14) is true and (15) false. In *unenlightened* contexts, speaker and audience are not aware of the identity of Superman with Clark Kent; hence there is no reason to think that

[4]Joseph Moore was probably the first one in pointing out that (14–15) can give rise to contradictory intuitions; Graeme Forbes concurred. For details, see Moore (1999b) and Forbes (1999).

[5]For details, see Moore (1999b), esp. pp. 93–97.

[6]See Moore (1999b), pp. 93–94.

'Superman' and 'Clark Kent' denote aspects, and they should be taken as denoting their referents instead—Superman himself. On this reading, (14) is true if, and only if, (15) is.

Moore's theory predicts that there will be two kinds of contexts in which we will hesitate between these two interpretations: First, those contexts in which the audience contains enlightened and unenlightened speakers alike; and second, those contexts in which the epistemic position of the audience is not known to us. Because each of the interpretations assigns different truth values to (14–15), it is natural that, in those contexts, we should hesitate concerning the truth value of (14–15)[7].

There of course are several important questions about Moore's account. For example, one should probably inquire further about his proposal that proper names are ambiguous between an individual and an aspectual reading. Nevertheless, the point I want to make is that, even if we grant Moore his theory of proper names, we can still show that his explanation of why we should have contradictory intuitions about (14–15) is not satisfactory.

The problem is that Moore's explanation does not account for the full range of contexts in which (14–15) evoke contradictory intuitions. For example, consider this. Suppose it is common knowledge between me and my friend that Superman is Clark Kent, and that my friend asks me about the truth value of (14–15). My friend says that no one else will know about the classification, that it is only for our own private amusement. If this were to happen, I feel I would still hesitate: Part of me would say that (14) is true and (15) is false, while another part of me would say that (14) is true if and only if (15) is. Nevertheless, Moore's theory predicts that in this context (14–15) should not evoke contradictory intuitions, since the epistemic state of the audience is perfectly clear.

Moore's explanation of why (14–15) evoke contradictory intuitions relies on the epistemic position of the audience for which (14–15) are asserted (or evaluated, as the case may be). But this does not seem to be the real explanation, as we still hesitate about (14–15), even in contexts in which the epistemic position of speaker and audience is perfectly clear and uniform.

[7] See Moore (1999b), p. 96

The Point-of-View Explanation

A better explanation is that (14–15) are ambiguous can be given by appealing to the claim that (14–15) bear a tacit 'from the point of view of' prefix. The proposal is this: (14) and (15) evoke contradictory intuitions because they can be interpreted either from our own point of view, or from the point of view of the inhabitants of Metropolis. Let me explain.

As I have argued above, the most natural way to interpret (14) and (15) is as part of an act of storytelling of the Superman stories. When they are taken in this way, they must be interpreted from the point of view of the inhabitants of Metropolis. On this interpretation, as we have seen, (14) is interpreted as (18), and is true; and (15) is interpreted as (19), and is false.

But there are circumstances in which we would not take (14) and (15) as part of an act of storytelling. Suppose that our friend Jones asks us to make a list of Superman's actions in the Superman stories, and that it is common knowledge between Jones and us that Superman is Clark Kent. When making that list, it seems clear that we would use sentences like (14) and (15), and moreover that (14) and (15) would describe *the same action*. The important point is that, in this context, once we have included (14) in the list of actions by Superman, we would not increase that list by adding (15) to it. This suggests that, for the purposes of carrying out Jones' request, we would be interpreting (14) and (15) as (23) and (24), respectively:

(23) From our own point of view, Clark Kent always arrived at the scene after one of Superman's daring rescues

(24) From our own point of view, Superman always arrived at the scene after one of Clark Kent's daring rescues

Needless to say, when (14–15) are interpreted as (23–24), they express the same proposition, which is what explains our intuition that (23) is true if, and only if, (24) is.

In the storytelling context and in the action-sorting context it is clear what point of view we should adopt in interpreting (14–15). In other contexts it will not be so clear. This is what happens in the scenario described at the end of the previous section: We are asked about the truth value of (14–15), but we do not know whether we should do our interpretation from the point of view of the inhabitants of Metropolis, or from our own point of view. In circumstances like this, we would naturally hesitate between the two

readings of (14–15), which would in turn cause us to hesitate about their truth value.

Thus (14–15) evoke contradictory intuitions because they are ambiguous, and the ambiguity arises because there is a tacit 'from the point of view of' prefix that can be completed in two different ways. Therefore, this is another reason in favor of our claim that, when we are inclined to say that (14) is true and (15) is false, we are interpreting them as (18) and (19).

7.4.4 Conclusion

Saul's challenge to contextualists has been defused. On our view, the explanation of how (14) and (15) can differ in truth value is perfectly consistent with contextualist premises. All contextualists claim that, within belief attribution contexts, coreferential proper names can make different contributions to the truth conditions, and the present view agrees with that claim. On the Belief Theory, proper names inside the scope of the 'from the point of view of' operator can also make different contributions to truth conditions. The explanation of the opacity in both cases is perfectly parallel.

Chapter 8

An Argument for Simulation: The Singular 'They'

Our Simulation Semantics proceeds on the assumption that, when speakers talk about the beliefs of another person, they sometimes engage in a simulation of the other person. I tried to defend this claim by appealing to the circumstances in which simulation was supposed to occur (namely, when there is a significant disagreement between the speaker and the audience, and the subject of the attribution). In this chapter I want to suggest that, in addition to this argument, there is some *linguistic evidence* that supports the Simulation Semantics.

The argument I will present relies on a feature of the use of the plural pronoun 'they.' One may think that the pronoun 'they,' if it refers at all, it has to refer to a plurality of objects. If this is right, then the speaker who says:

They are *F*,

for some predicate *F*, would become committed to the existence of a plurality of *F*s. However, this is not so. It is possible for a speaker of English to sincerely use the pronoun 'they' when she thinks that there is exactly one object to be referred by the use of the pronoun. For example:

(1) What about Hesperus and Phosphorus? *They* are the same planet.

(2) What about Clemens and Twain? *They* are the same writer.

It is clear that the person who answers the questions believes that Hesperus and Phosphorus are the same planet, and that Clemens and Twain are the same person. The presence of a plural pronoun is therefore curious, for the speaker uses a plural form, in spite of the fact that she believes that there is only one thing to be referred to. It is curious, but perfectly grammatical, as both (1) and (2) are perfect English[1].

These examples show that there are circumstances in which it is possible to use the pronoun 'they' sincerely, without thereby becoming committed to the existence of a plurality of objects to be referred to by using the pronoun. My hypothesis is that it is possible to use the pronoun 'they' in this way because in the contexts in which it is so used, there is an open question about whether there is exactly one object, or more than one object, to be referred to by the use of the pronoun. For example, in the little dialogues displayed in (1) and (2), the questions create a context in which there is an open question about whether Hesperus and Phosphorus are one or two planets, and about whether Clemens and Twain are the same writer or different ones. This pragmatic features of the pronoun would assimilate it to other well-known linguistic devices that can be used to refrain from committing oneself from attributing a property to an object when one is not sure about the object has the property or not. In Spanish, for example, it is possible to use words whose grammatical gender is masculine as a way of describing an object, without thereby expressing commitment about the real gender of the object so described. (I think that American English had such a rule once, but it has now become more difficult to pin down, thanks to the advent of political correctness.)

If this is right, then this feature of the use of the pronoun 'they' is explained by the following pragmatic rule:

> SINGULAR 'THEY': It is legitimate to sincerely use the pronoun 'they' when one believes that there is exactly one object to be referred to, as long as the use occurs in a context in which there is an open question about whether the pronoun refers to one or more objects

This rule seems to be further confirmed by the fact that, in the above examples, more uses of the pronoun 'they' are not acceptable:

[1] I suppose that I had known about this use of the pronoun 'they' since I learnt English, but I did not come to realize how *curious* it is until I read T.S. Champlin's discussion of it in Champlin (1993).

(3) What about Hesperus and Phosphorus? *They* are the same planet.

a. * *They are the ones that appear in the morning and in the evening.*

(4) What about Clemens and Twain? *They* are the same writer.

a. * *They are the ones who wrote 'Huckleberry Finn'*

The asterisk indicates that the sentences in (a) sound odd. The explanation for this phenomenon is that, by the time the (a) and (b) forms are uttered, the question about whether there is a plurality of objects to be referred to has already been answered in the negative, and for this reason it would not be acceptable for the speaker to go on using 'they.' If we continue our answer with the forms in (a), that seems to give the audience the impression that, after all, there is a plurality of objects to be referred to, which would seem contradictory on our part.

Let us now explore the interaction of the singular 'they' with belief attribution. Suppose, for example, that the following conversational exchange takes place between you and me:

(5) What about Jones, what does he believe about Hesperus and Phosphorus? He believes that *they* are the same planet.

(5) is perfectly acceptable, even if it is uttered in a context in which it has long been clear between you and me that Hesperus is Phosphorus. What is more, it is clear that we can answer the question in the way depicted in (5), without committing ourselves to the claim that there is a plurality of objects to be referred to by the use of the pronoun, and without committing Jones to that claim. However, on the face of it, (5) seems to violate the pragmatic rule spelled out in SINGULAR 'THEY'. What is going on? What legitimates us to use the pronoun 'they' in these circumstances?

I believe that the Simulation Semantics defended in chapter 5 can contribute to explain the occurrence of the singular 'they' in (5), together with the pragmatic rule SINGULAR 'THEY' which appeared so plausible. In particular, we can use the Simulation Semantics to show that the context in which (5) is asserted is a context in which, after all, there is an open question about whether there is a plurality of objects to be referred to by using the pronoun 'they.'

To begin with, note that the context in which (5) takes place is a context in which there is an open question about whether Jones believes that Hesperus and Phosphorus are two different planets, or the same

one. Because there is this open question, the Simulation Semantics implies that there is some uncertainty about how to interpret the names 'Hesperus' and 'Phosphorus,' when they occur in the embedded sentence of an attribution of belief to Jones. If Jones believes that Hesperus and Phosphorus are different objects, then there will be a disagreement between Jones and ourselves concerning the number of planets, and it will be necessary to engage in simulation to describe his belief state. On this option, the names 'Hesperus' and 'Phosphorus' would refer to two different planets, and therefore there would be a plurality of objects to be referred to by using the pronoun 'they' in the *that*-clause of an attribution to Jones. On the other hand, if he believes that they are the same planet, then there will be no need for simulation, and both 'Hesperus' and 'Phosphorus' will refer to Venus. On this option, there will be only one object to be referred to by using the pronoun 'they' in the *that*-clause of an attribution to Jones.

The upshot is that, by the time the question in (5) is asked, there is an open question about whether there is a plurality of objects to be referred to by the use of the pronoun 'they' in the *that*-clause of a belief attribution to Jones. And these are precisely the circumstances in which, according to SINGULAR 'THEY', it would be legitimate to use the pronoun 'they,' even if we ourselves do not believe that there is a plurality of objects to be referred to by the use of the pronoun.

If the preceding is right, we should expect a pattern similar to the one in (3–4): The moment it is made clear that Jones takes Hesperus and Phosphorus to be the same, it is no longer acceptable to use the pronoun 'they' to attribute to Jones a belief about Hesperus and Phosphorus. Indeed, this is what we observe. Consider:

(6) What about Jones, what does he believe about Hesperus and Phosphorus? He believes that *they* are the same planet.

a. **He also believes that they are the planets that appear in the morning and in the evening*

The answer that Jones believes that Hesperus and Phosphorus are the same planet resolves the uncertainty about how to interpret the pronoun 'they,' and afterwards it becomes unadmissible to use it again; hence the oddity of (6a).

Since the Simulation Semantics helps us to understand why it is that we can legitimately and sincerely use the pronoun 'they' even when we do not believe that there is a plurality of objects to be referred to, the

existence of those uses of the pronoun is an argument in favor of the Simulation Semantics.

This argument can be attacked in along two major lines. One of them seeks to explain the acceptability of the uses of the pronoun 'they' by speakers who believe there is only one thing to be referred to by appealing to *grammatical features* of the contexts in which those pronouns are used. The other seeks to identify contexts in which the pronoun can be legitimately used by a speaker who believes that there is only one thing to be referred to, *but* there is no open question about whether there is only one thing to be referred to. I will discuss, in all, four different arguments, two of the first kind, and two of the second.

First, one may think, very naturally, that the use of the plural in examples like (1–6) is *forced* by some grammatical feature of the contexts in which the pronoun. For example, in (1) and (2), the pronoun 'they' has as antecedent the noun phrases 'Hesperus and Phosphorus' and 'Clemens and Twain,' which are gramatically plural (it is easy to see, for example, that it is impossible to use a singular verb after a conjunctive noun phrase like these ones). Couldn't it be that the reason why we use the pronoun is because we need the pronoun to agree in grammatical number with its grammatical antecedent? It does not seem so, at least for two reasons.

The first reason is that this proposal would not be able to explain why the second 'they' in (3) and (4) is not acceptable:

(3) What about Hesperus and Phosphorus? *They* are the same planet.

a. * *They are the ones that appear in the morning and in the evening.*

(4) What about Clemens and Twain? *They* are the same writer.

a. * *They are the ones who wrote 'Huckleberry Finn'*

One would expect the first and the second 'they' in (3) and (4) to have as antecedent the same phrase ('Hesperus and Phosphorus' and 'Clemens and Twain,' respectively), but then it would not be clear why the use of the plural pronoun in the numbered forms is acceptable, while the use of the pronoun in the (a) forms is not.

One might try to counter this argument by pointing out that the second 'they' in (3) and (4) does not have the same antecedent as the first. Indeed, after the first 'they,' there is the gramatically singular phrases

'the same planet' and 'the same writer' which, the argument goes, would serve as antecedent for subsequent pronominal uses. Since the phrase is gramatically singular, that makes the use of plural pronouns unacceptable.

This reasoning is interesting, but notice that we can reformulate the examples so that there is no gramatically singular antecedent. Consider, for example:

(7) What about Hesperus and Phosphorus? *They* are identical.

a. * *They are the ones that appear in the morning and in the evening.*

(8) What about Clemens and Twain? *They* are identical.

a. **They are the ones that wrote 'Huckleberry Finn'*

In (7) and (8), there is no gramatically singular nominal phrase that could be used as the grammatical antecedent of the second 'they;' there is just the adjectival phrase 'are identical,' which cannot serve as antecedent for a pronoun.

The second reason is that we can give some examples similar to (1) and (2) in which there is no gramatically plural phrase to serve as the antecedent of the pronoun. Consider, for example, the following discourse:

(9) *A:* In past days we have observed a planet that we have called 'Phosphorus,' and that is visible in the morning sky, hovering close to the rising Sun. I wonder whether it is identical to 'Hesperus,' a planet that we discovered long ago, which is visible in the evening sky, hovering close to the setting Sun.

B: I will tell you: *They* are the same!

In this example, there is no gramatically plural antecedent to serve as the antecedent of *B*'s use of 'they.' Nevertheless, the exchange seems perfectly natural, and certainly does not commit *B* to the existence of a plurality of objects to be referred to by the use of the pronoun 'they.' The grammatical explanation for the use of 'they' cannot work in this case, but our pragmatic explanation does, since it is clear that, in this context, there is an open question about whether Hesperus and Phosphorus are the same planet.

Second, one may seek to explain the acceptability of the plural pronoun by appealing to other grammatical features of the context in which it appears. In particular, in the examples we have been reviewing, it appears followed by a verb followed by the locution 'the same.' Could it be that the expression 'the same' forces the use of the plural pronoun? Let us take a look at this possibility.

It could be argued that there are two different phrases 'the same,' paralleling the two readings of 'is' recognized in classical literature on the subject. On one reading, the phrase 'the same' can be used to predicate the relation of identity. This is what happens, for example, when we say:

(10) Hesperus and Phosphorus are the same

The utterer of (10) does not want to be understood as saying that there are two different objects, but just one. But there is another reading on which the phrase 'the same' is used to predicate identity in some respect or other. This is the reading observed, for example, in sentences like:

(11) Your car and mine are the same

The utterer of (11) does not want to be understood as saying that there is just one car which you and I have; rather, what she means is that there are two cars that are identical in some relevant respects, and that you have one and I have the other.

One could then claim that the *identity* reading of 'the same' always demands a verb in the plural, and that as a result, the subject of the verb (if there is one) has to be in the plural too. Therefore, in sentences like (1) and (2), we are simply forced to use the plural pronoun.

This line of argument has two sorts of difficulties. In the first place, if it is true that the expression 'the same' has two different readings, it has to be explained what those readings could be, and I simply do not see how to do that. In the second place, and more decisively, there are phrases containing the phrase 'the same' in object position, in which the verb is in the singular, and yet in which, if 'the same' is ambiguous, it is surely the identity reading the one which is intended. For example, consider the following sentence:

(12) The car parked in front of the house is the same that we saw yesterday

The utterer of (12) does not want to be understood as saying that the car parked in front of the house today

is different from the car parked in front of the house yesterday; she wants to say that a single car was parked there both times.

We could even arrange the example so that there is no relative clause after 'same,' to complete the parallel with (1) and (2):

(13) *A:* I though I saw that car yesterday, parked in front of the house.
B: Yes, it's the same

Examples like this show that, if there is a reading of 'the same' on which it means identity, as opposed to something else, that reading does not require the use of plural verb.

The next line of argument tries to figure out counterexamples to our pragmatic rule SINGULAR 'THEY'. The third argument relies on the following example:

(14) *A:* Today I have been studying two new planets: Hesperus and Phosphorus.
B: But they are the same!

One might think that, in this context, there is no open question about whether Hesperus and Phosphorus are the same: *A* thinks that they are not, and *B* thinks that they are. For this reason, there is no open question about whether the pronoun would refer to one thing or more. But since, in this case, *B*'s response seems perfectly acceptable, cases like this seem a counterexample to SINGULAR 'THEY'.

However, it is important to realize that this example is special: There is complete disagreement between speaker and hearer over the reference that an utterance of 'they' would have, were it to be uttered: The speaker thinks it would refer to only one thing, while the audience thinks it would refer to two things. How is communication even possible when there is this sort of disagreement? As we explained in §6.4.2, in this case speaker and audience communicate by revising their belief state so that both the possibility that Hesperus is Phosphorus, and Hesperus is not Phosphorus, is compatible with it. Therefore, by the time (14) is uttered, speaker and audience are, in effect, in a belief state which is compatible with both the hypotheses that Hesperus is Phosphorus and that Hesperus is not Phosphorus. And if this is right, then the context

in which (14) is uttered is a context in which there is an open question about whether the pronoun would refer to one thing or several. Therefore, on close inspection, this does not appear to be a counterexample to SINGULAR 'THEY'.

The fourth and final argument draws our attention to examples like the following:

(15) Let me talk about the ancient astronomers. They thought that Hesperus and Phosphorus were different planets. But as everyone knows, *they* are the same.

(16) Nobody doubts, of Hesperus and Phosphorus, that *they* are the same.

Intuitively, (15) and (16) are asserted in a context in which it is common knowledge to everyone involved that Hesperus and Phosphorus are the same planet, and in which, therefore, there is no open question about whether a use of the pronoun 'they' in that context would refer to more than one object.

There nevertheless is something special about (15) and (16), and it is that they are assertions that do not communicate anything new to the audience: By assumption, the context in which (15) and (16) are asserted is a context in which everyone *knows* that Hesperus is Phosphorus, and moreoever there is the expectation (and probably the common knowledge too) that everyone will know that too. Why bother then saying that everyone knows it, or that nobody doubts it?

Here is a sketch of a tentative answer. The idea is that, in many occasions in which someone says something that everyone knows, there is a pretense in place to the effect that the audience does not know what the speaker is saying. And there is a good reason for this pretense. In ordinary conversation, we abhor repetition: It is wasteful of time, and makes conversation non-efficient. But in other settings, we simply have to state what may be obvious to everyone. Many of those settings are contexts in which we have a pedagogical or academical purpose in mind: We want to convince our audience of some conclusion, or draw their attention to some inference that they had previously missed, and to do so we have reiterate some of the things that we and the audience know. The function of the pretense is to reconcile this need to state the obvious with the conversational directive to avoid stating what is common knowledge.

I am not sure that this sketch is enough to explain why we sometimes state propositions that are common knowledge, but at least it provides a plausible explanation for it. From our point of view, the virtue of

this account is that, if it is right, then (15) and (16) are not counterexamples to SINGULAR 'THEY', since, on this explanation, (15) and (16) are asserted in a context in which there is a pretense that it is not known that Hesperus and Phosphorus are the same planet.

My conclusion is that the best explanation for why we can legitimately use the plural pronoun 'they' when we know there is only one thing to be referred to is that there is a pragmatic rule to the effect that the pronoun is acceptable, as long as it is used in a context in which there is an open question about whether the pronoun would refer to one or more things. And if this is granted, then this use of the pronoun is yet another argument in favor of the Simulation Semantics defended in chapter 5.

Bibliography

Bezuidenhout, A. (1996) "Pragmatics and Singular Reference," *Mind and Language* **11**.

Boghossian, P. (1989) "Content and Self-Knowledge," *Philosophical Topics* **17**, 5–26.

Boghossian, P. (n.d.) "What the Externalist Can Know A Priori," http://www.nyu.edu/gsas/dept/philo/faculty/boghossian/papers/Externalist.html.

Braun, D. (1991) "Proper Names, Cognitive Contents, and Beliefs," *Philosophical Studies* 289–305.

Braun, D. (1993) "Empty Names," *Nous* **27**, 449–469.

Braun, D. (1998) "Understanding Belief Reports," *The Philosophical Review* **107**, 555–595.

Burge, T. (1973) "Reference and Proper Names," *The Journal of Philosophy* **70**, 425–439.

Burge, T. (1979) "Individualism and The Mental," *Midwest Studies in Philosophy* **4**, 73–121.

Burge, T. (1988) "Individualism and Self-Knowledge," *The Journal of Philosophy* **85**, 649–663.

Champlin, T. (1993) "A Curious Plural," *Philosophy* **68**, 435–455.

Cole, P., ed. (1978) *Syntax and Semantics 9: Pragmatics*, New York Academic.

Crimmins, M. (1992) *Talk about belief*, MIT Press.

Crimmins, M. (1998) "Hesperus and Phosphorus: Sense, Pretense and Reference," *The Philosophical Review* **107**, 1–47.

Crimmins, M., and J. Perry (1989) "The Prince and the Phone Booth," *The Journal of Philosophy* **86**, 685–711.

Davidson, D. (1968) "On Saying That," in Davidson (1984), 93–108.

Davidson, D. (1980) *Essays on Actions and Events*, Oxford University Press.

Davidson, D. (1984) *Inquiries into Truth and Interpretation*, Oxford University Press.

Davis, S., ed. (1991) *Pragmatics. A reader*, Oxford University Press.

Devitt, M. (1981) *Designation*, Columbia University Press.

Donnellan, K. (1970) "Proper Names and Identifying Descriptions," *Synthese* **21**, 335–358.

Donnellan, K. (1974) "Speaking of Nothing," *The Philosophical Review* **83**, 3–31.

Evans, G. (1982a) "The Causal Theory of Reference," in *Collected Papers*, Oxford University Press.

Evans, G. (1982b) *The Varieties of Reference*, Oxford University Press.

Fitch, G. (1993) "Non Denoting," *Philosophical Perspectives* 461–486, 461–486.

Forbes, G. (1999) "Enlightened Semantics for Simple Sentences," *Analysis* **59**, 86–91.

Frege, G. (1892) "On Sense and Reference," in P. Geach and M. Black, eds., *Translations from the philosophical writings of Gottlob Frege*, Basil Blackwell.

Frege, G. (1918) "The Thought," in Salmon and Soames (1988).

French, P., T. Uehling, and H. Wettstein, eds. (1979) *Contemporary Perspectives in the Philosophy of Language*, University of Minnesota Press.

French, P., T. Uehling, and H. Wettstein, eds. (1986) *Midwest Studies in Philosophy IX: Studies in Essentialism*, University of Minnesota Press.

Gärdenfors, P. (1988) *Knowledge in Flux*, The MIT Press.

Goldman, A. (1978) "Perceptual Objects," in Pappas and Swain (1978), 271–296.

Grimm, R., and D. Merrill, eds. (1986) *Contents of Thought*, University of Arizona Press.

Kanger, S., and S. Oehman (eds.) (1980) *Philosophy and Grammar*, D. Reidel Publishing Company.

Kripke, S. (1976) "A puzzle about belief," in Salmon and Soames (1988), 102–148. Originally published in Margalit (ed.): *Meaning and Use*. Dordrecht, D. Reidel, 1976.

Kripke, S. (1980) *Naming and Necessity*, Harvard University Press.

Larson, R., and G. Segal (1995) *Knowledge of meaning: An introduction to semantic theory*, The MIT Press.

Lau, Y. (1994) *Belief in Semantics and Philosophy*, Doctoral dissertation, MIT.

Lewis, D. (1974) "Radical Interpretation," in Lewis (1983), 108–121.

Lewis, D. (1978) "Truth in Fiction," in Lewis (1983).

Lewis, D. (1980) "Index, Context, and Content," in Kanger and Oehman (eds.) (1980), 101–118.

Lewis, D. (1983) *Philosophical Papers*, volume 1, Oxford University Press.

Moore, J. (1999a) "Misdisquotation and Substitutivity: When Not to Infer Belief from Assent," *Mind* **108**, 335–365.

Moore, J. (1999b) "Saving substitutivity in simple sentences," *Analysis* **59**, 91–105.

Munitz, M., and P. Unger, eds. (1974) *Semantics and Philosophy*, New York University Press.

Neale, S. (1990) *Descriptions*, The MIT Press.

Pappas, G., and M. Swain, eds. (1978) *Essays in knowledge and justification*, Cornell University Press.

Putnam, H. (1975) "The meaning of 'meaning'," in *Collected papers*, volume 2, Cambridge University Press.

Richard, M. (1983) "Direct Reference and Ascriptions of Belief," in Salmon and Soames (1988), 169–196. Originally published in *Journal of Philosophical Logic*, **12**: 425–452.

Russell, B. (1904) "On Denoting," in Russell (1956), 41–56.

Russell, B. (1956) *Logic and Knowledge*, George Allen and Unwin.

Salmon, N. (1986) *Frege's Puzzle*, MIT Press.

Salmon, N., and S. Soames, eds. (1988) *Propositions and Attitudes*, Oxford University Press.

Saul, J. (1997) "Substitution and simple sentences," *Analysis* **57**, 102–108.

Saul, J. (1998) "The Pragmatics of Attitude Ascriptions," *Philosophical Studies* **92**, 363–389.

Saul, J. (1999) "The Road to Hell: Intentions and Propositional Attitude Ascriptions," *Mind and Language* **14**, 356–375.

Schiffer, S. (1987a) "The 'Fido'-Fido theory of belief," in J. Tomberlin, ed., *Philosophical Perspectives*, volume 1, Ridgeview, 455–480.

Schiffer, S. (1987b) *Remnants of Meaning*, The MIT Press.

Schiffer, S. (1992) "Belief ascription," *The Journal of Philosophy* **89**, 499–521.

Schiffer, S. (1995) "Descriptions, Indexicals, and Belief Reports: Some Dilemmas (But Not the Ones You Expect)," *Mind* 104, 107–131.

Soames, S. (1987a) "Direct reference, propositional attitudes and semantic content," in Salmon and Soames (1988), 197–239. Originally published in *Philosophical Topics*, **15**:47–87; 1987.

Soames, S. (1987b) "Substitutivity," in J. Thomson, ed., *On Being and Saying*, The MIT Press, 99–132.

Stalnaker, R. (1974) "Pragmatic Presuppositions," in Munitz and Unger (1974), 197–213.

Stalnaker, R. (1978) "Assertion," in Cole (1978), 315–332. Also reprinted in Davis (1991), pp. 278–289.

Stalnaker, R. (1984) *Inquiry*, The MIT Press.

Stalnaker, R. (1986a) "Belief Attribution and context," in Grimm and Merrill (1986), 140–156.

Stalnaker, R. (1986b) "Counterparts and Identity," in French et al. (1986), 121–140.

Stalnaker, R. (1987) "Semantics for belief," *Philosophical Topics* **15**, 177–190.

Walton, K. (1973) "Pictures and Make-Believe," *The Journal of Philosophy* **82**, 283–319.

Walton, K. (1993) "Metaphor and Prop Oriented Make Believe," *European Journal of Philosophy* **1**, 39–56.

Yagisawa, T. (1997) "Salmon Trapping," *Philosophy and Phenomenological Research* **57**, 351–370.