“You” and “I”, “now” and “then”, and Some Related Thoughts

by

Paolo Santorio
Laurea, Università di Bologna, 2005

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

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2011

©Massachusetts Institute of Technology 2011. All rights reserved.

Author

Department of Linguistics and Philosophy
September 8, 2011

Certified by

Robert C. Stalnaker
Rockefeller Professor of Philosophy
Dissertation Supervisor

Accepted by

Alex Byrne
Professor of Philosophy
Chair, Departmental Committee of Graduate Studies
“You” and “I”, “now” and “then”, and Some Related Thoughts

by

Paolo Santorio

Submitted to the Department of Linguistics and Philosophy on September 8th, 2011
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

ABSTRACT

The overarching theme of this dissertation is the relationship between information and context—how context interacts with the contents of speech and thought. I pursue three issues within this broad theme.

Chapter 1 concerns indexicality. On classical accounts, indexicals like I, you, and now invariably refer to elements of the actual context of speech. This feature sets indexicals apart from definite descriptions like the best soccer team in Europe. I argue that the orthodoxy is wrong: all indexicals are subject to shift, exactly like descriptions. More precisely, indexicals are syntactically akin to logical variables and hence can occur both free and bound. After arguing for the new view, I show how it affects our general picture of context-dependence in semantics.

Chapter 2 concerns definite and indefinite descriptions, like the richest banker on Wall Street and a stockbroker. Descriptions are often treated as quantified phrases, sharing their structure with quantified phrases; hence they are taken to belong to a different semantic kind from referential expressions. Contrary to this view, I argue that all descriptions can involve a referential component in their meaning. This referential component is independent of the descriptive material and hence constitutes a separate dimension of the semantics of descriptions.

Chapter 3 investigates how subjects’ degrees of belief should evolve as their position in the world changes—for example, as time passes. I argue that we should recognize a new kind of cognitive operation, cognitive relocation. Relocation manipulates a special kind of information, de re information about time; moreover, it is a way of retaining rather than gaining information. The relocation account is able to reconcile intuitions about a well-known example, the Sleeping Beauty problem, with an intuitively plausible principle about context-dependent information, the Relevance-Limiting Thesis.

Thesis Supervisor: Robert C. Stalnaker
Title: Rockefeller Professor of Philosophy
Acknowledgments

This dissertation is, first and foremost, a product of the intellectual environment of MIT’s department of Linguistics and Philosophy. My greatest debt is to this community as a whole. My main hope for my career is that, wherever academia takes me, I find environments equally rich in ideas and passion.

Several people within this community deserve special mention; the natural place to start is my dissertation committee. Bob Stalnaker has been a generous, attentive, and extremely supportive supervisor. Aside from giving decisive feedback, Bob has proved invaluable for placing my work within a bigger picture. My interactions with Bob have shaped, more than anything else, my overall vision of what philosophy is about and how it should be done. My meetings with Irene Heim have been crucial for developing my general views on reference and indexicality. On a par with Bob, Irene has also shaped the kind of researcher I’ve become. I like to fancy myself as a hybrid of a philosopher and a semanticist; to the extent that the latter part actually applies, it’s thanks to Irene. Alex Byrne has given important feedback on everything I’ve written in the last three years. I thank him especially for always coming up with the odd question out—the one that would never get asked at MIT, but that, without Alex, would have put me in trouble in talks given to very different audiences. Steve Yablo has been a precious eleventh-hour addition. He kept apologizing that he had little to say about my work, and in retrospect that seems a good thing—for whenever he did have things to say, they turned out to be worrisome objections that prompted substantial rethinking.

Besides my committee, I am especially grateful to a number of other faculty members at MIT. Thanks to Agustín Rayo, who devoted hours and hours to reading and commenting on the many drafts I gave him. This dissertation would be significantly worse without his contribution. Thanks to Danny Fox, who first got me interested in semantics and kept dedicating time and attention to me whenever I asked. Thanks to Rae Langton, who proved to me that some not-so-far-off counterparts of me have happily decided to become historians of philosophy (specializing on Leibniz in worlds at closer range, and on Kant in some farther off ones).

Two friends and colleagues have had a major role in my work and in building my self-confidence as a researcher. Thanks to Dilip Ninan, with whom I’ve had very ex-
tended conversations about indexicality and de se attitudes, and who has given me invaluable feedback on most of this dissertation. I felt very privileged to have Dilip as an interlocutor. Thanks to Fabrizio Cariani, who has been a constant source of philosophical interactions, conversations, bantering, and (infallibly correct) academic gossip throughout these six years. Most importantly, Fabrizio gave excellent feedback in crucial, and sometimes emotionally very difficult, moments of my career.

I was lucky to be a part of the most boisterous, dorky, and international class that I’ve ever seen at MIT. Mahrad Almotahari, Tom Dougherty, Alejandro Pérez Carballo, Katia Vavova, Kenny Walden—you’ve made my first year and much more. I owe you most of my early training in philosophy. Alejandro, in particular, should be singled out for the (literally) hundreds of hours we spent discussing philosophy. I have also benefited a great amount from interactions with a number of graduate students at MIT in both philosophy and linguistics. I should mention, in particular, Lauren Ashwell, Rachael Briggs, Jennifer Carr, Luka Crnic, Ryan Doody, Ephraim Glick, Dan Greco, Dan Hagen, Brian Hedden, Ginger Hoffmann, Heather Logue, Sarah Moss, Chris Robichaud, Damien Rochford, Raj Singh, Paulina Sliwa, Tue Trinh, Igor Yanovich.

People I met in my academic travels also mattered a lot. First of all, thanks to Andy Egan for going way beyond the calls of duty and courtesy in discussing my thoughts with me, on several occasions, while we were both on the road. Then, thanks to the ANU for welcoming me as a visitor in the summer of 2009, and in particular to David Chalmers, Alan Hájek, Jo Lau, Leon Leontyev, Stephan Leuenberger, Gabe Rabin, Brian Rabern, Alex Skiles, Daniel Stoljar, Mike Titelbaum, Moises Vaca, Clas Weber. Thanks to the Arché research center in St Andrews for welcoming me twice as a visitor in the summers of 2010 and 2011, and in particular to Derek Ball, Yuri Cath, Steven Hall, Torfinn Huvenes, Dirk Kindermann, Julia Langkau, Andrea Onofri, Carlotta Pavese. Special thanks to Andreas Stokke, with whom I’ve had many productive exchanges. Thanks to a number of people that I cannot tie to a location: Corine Besson, Stavroula Glezakos, Orin Percus, David Plunkett, Philippe Schlenker, Yael Sharvit.

Thanks to my early teachers at UC Berkeley, who got me so enthralled about philosophy that I decided to switch fields and apply to PhD programs. Thanks, in particular, to Julien Deonna, John MacFarlane, Paolo Mancosu, Barry Stroud.

Thanks to all my friends and roommates in Cambridge and in far-away Bologna and Turin. You’re too many to list your names, but let me just say that I owe it to you, collectively, if I’ve both become a doctor and retained my sanity.

As usual, the most important acknowledgments come at the end. Thanks to my parents for their incredible support and their love, and for tolerating a somewhat difficult only child who’s chosen a very difficult career path. It is to them that I dedicate this dissertation. Thanks to Jennifer Carr, for her support, her care, her love. To her, I hope to dedicate many years to come.
## Contents

0 Introduction .......................................................... 11

1 Reference and Monstrosity ........................................... 15
   1.1 Overview .......................................................... 15
   1.2 Epistemic shift .................................................. 17
      1.2.1 Double indexing ........................................... 17
      1.2.2 Shifty conditionals ...................................... 18
      1.2.3 Three quick replies ..................................... 21
      1.2.4 Generalizations: other indexicals, other modals ......... 23
      1.2.5 Review ...................................................... 25
   1.3 Informational modals are monsters .............................. 26
      1.3.1 Preliminaries ............................................... 26
      1.3.2 Variables over counterparts ............................... 27
      1.3.3 Assignment shift ........................................... 29
      1.3.4 Reference without rigidity ................................. 32
   1.4 Monstrous semantic theory ..................................... 34
      1.4.1 Semantics, context, and content .......................... 34
      1.4.2 Re-anchoring indexicals .................................. 37
      1.4.3 Semantics without the context ............................. 39
   1.5 Extensions ........................................................ 41
      1.5.1 Names .......................................................... 41
      1.5.2 Gendered pronouns ......................................... 43
      1.5.3 Adverbs of quantification .................................. 45
      1.5.4 Binding of fake indexicals ................................. 47
   1.6 Outstanding issues ............................................... 49
   1.7 Conclusion ........................................................ 50

2 Denoting without Description .................................... 55
   2.1 Introduction ...................................................... 55
   2.2 Setup: descriptions and *de re/de dicto* phenomena .......... 57
CONTENTS

2.3 Referential opaque readings ........................................ 60
  2.3.1 The claim: a fourth reading .................................. 60
  2.3.2 The argument from gender features .......................... 63
  2.3.3 The argument from intersentential anaphora ................. 65

2.4 Two routes to an account ......................................... 69
  2.4.1 The DRT analysis: descriptions as variables ............... 70
  2.4.2 A conservative analysis: variables in the restrictor ........ 71

2.5 Outstanding issues ................................................ 74

3 Cognitive Relocation ............................................... 77
  3.1 Introduction ...................................................... 77
  3.2 De se attitudes and the Bayesian framework ..................... 79
    3.2.1 De se attitudes ............................................. 79
    3.2.2 The Bayesian framework and the problem of update ........ 80
    3.2.3 Preview: two routes to update .............................. 81
  3.3 Temporal relocation ............................................. 83
    3.3.1 De re information ......................................... 83
    3.3.2 An analogy: tracking movement ............................ 85
    3.3.3 Temporal information ...................................... 87
    3.3.4 Relocation .................................................. 90
  3.4 The formal framework .......................................... 93
    3.4.1 Sequenced worlds ......................................... 93
    3.4.2 Relocation as imaging .................................... 95
  3.5 The relevance of de se information .............................. 98
    3.5.1 The question of epistemic relevance ....................... 98
    3.5.2 The Sleeping Beauty problem .............................. 101
  3.6 Update and communication ..................................... 106
    3.6.1 The problem of content and the problem of calibration .... 106
    3.6.2 Comparison with Moss’s account .......................... 108
  3.7 Conclusion ..................................................... 109
STUDENT: Here is a counterexample: [produces a counterexample to a theory under discussion]. So there is an objection to the theory you gave.

DANNY FOX: This is not an objection. You don't have an objection until you have a theory that accounts for the previous data plus your counterexample. Without that, we just have a problem that we can put in the big pile of problems that we already have.

(from Danny Fox’s pragmatics class, some time in the Fall 2007)
This dissertation is a progress report on a number of related questions touching philosophy of language, philosophy of mind, and—to a smaller extent—epistemology. The overarching theme is the relationship between information and context—how context interacts with the contents of speech and thought. Within this theme, I have carved out three questions that I pursue in three chapters: linguistic indexicality and the interaction between epistemic modality and indexicals; the semantics of definite and indefinite descriptions and their capacity to refer; the evolution of self-locating attitudes through time. I’ve chosen them, in part, because I took them to be important open questions in the field; and, in part, simply because I happened to have original ideas about them. (The first of these factors is probably tied to the second more than I’d like to believe.)

The chapters are presented as fully independent papers. I string them together here, trying to show what relates them.

Chapter 1 concerns the semantics of indexicality. On the classical account dating back to Kaplan (1989), indexical expressions like I, you, and now invariably refer to elements of the actual context of speech and cannot be shifted by modal or tense operators. This feature sets indexicals apart from definite descriptions like the best soccer team in Europe, which can pick out different individuals when they occur embedded under tenses or modals. I argue that the orthodoxy is wrong: all indexicals are subject to shift, exactly like descriptions. More precisely, indexicals are syntactically akin to logical variables. They can be free, in which case they work on the model presented by Kaplan. But they can also be bound: this happens, in a systematic fashion, when they are in the scope of epistemic modals or attitude verbs. Consider:

(1) I might be in Chicago.

I claim that, in (1), I is a variable bound by might. It ranges over epistemic counterparts of the speaker, namely individuals who, for all the speaker knows, she might be. So (1)
CHAPTER 0. INTRODUCTION

says that there is a possibility where the speaker’s counterpart is in Chicago.

Allowing for indexical shift requires significant changes to our picture of reference and context-dependence; the second part of the chapter is devoted to illustrating them. In addition, I survey the prospects of generalizing the view to other expressions, including proper names and gendered pronouns like she and he.

Chapter 2 is about definite and indefinite descriptions, like the richest banker on Wall Street and a stockbroker. Descriptions are often treated as quantified phrases, sharing their structure with constructions like every banker in Manhattan: hence they are taken to belong to a different semantic kind from referential expressions. Contrary to this view, I argue that all descriptions can involve a referential component in their meaning. Crucially, this referential component is independent of the descriptive material (i.e., the predicate contained in the description). This independence is exhibited most clearly in attitude reports. In those contexts, it may be that the descriptive material is evaluated with respect to a subject's attitudes, and the description still refers to an actual world individual. For example, take:

(2)  Stacey wishes that the richest man on Wall Street would ask her out.

I claim that the richest man on Wall Street has a reading where (a) it is evaluated de dicto with respect the attitude verb, and (b) it refers to an actual world individual, which Stacey (falsely) takes to be the the richest man on Wall Street. I don't settle how this referential component works. I explore two hypotheses: on the first, descriptions just are essentially variables; on the second, they are still quantifiers but contain a variable in their covert restrictor. Hence, in either case, the mechanism by which descriptions refer is analogous to that exploited by deictic pronouns like she and he.

The arguments of the first two chapters are fully independent, but they both contribute to a unified picture of reference in language. This picture is summed up in the following slogan:

Variable Hypothesis. All reference to objects in natural language is performed by means of variables; all referential expressions occur systematically bound in the scope of operators with epistemic flavor, such as epistemic modals and attitude verbs.

Chapter 1 is a straightforward argument that the Variable Hypothesis holds for indexicals and shows how the argument can be extended to names and pronouns. Chapter 2 doesn't argue directly for an analogous conclusion about descriptions, but it provides the tools for treating them in conformity to the variable model.

Of course, the two chapters only provide a partial argument for the Variable Hypothesis. In the first place, more should be said about descriptions, in particular about
how standard \textit{de re} readings of them fit into this picture. Even more importantly, recent research has shown that variable binding behaves in complex and unexpected ways just under attitude verbs (see Sharvit (2011)). These issues require further developments of the semantics I give in chapter 1: at the moment, I can only leave these developments to further work.

Chapter 3 marks a shift in focus from language to thought. The question under investigation is how your degrees of belief should evolve as your position in the world changes—for example, quite simply, as time passes. We have developed theories of context-dependent attitudes, as well as a developed framework for modeling rational belief update; but it’s not clear how to bring the two together. I argue that we should recognize a new kind of cognitive operation, which I call ‘cognitive relocation’: the chapter is devoted to introducing this operation, explaining how it works, and exploring some philosophical consequences. The account is organized around two main ideas, which straddle philosophy of mind and epistemology. First, relocation manipulates a special kind of information, \textit{de re} information about time. This is information tied to the subject’s capacity of acquiring and retaining information via a causal route, for example via perception and memory. Second, while ordinary kinds of update are ways of gaining information, relocation is essentially a way of retaining information. By relocating effectively, you preserve your level of certainty in your \textit{de re} information about time. By relocating ineffectively, you become more uncertain about this information.

My account of relocation has a number of philosophical consequences. The most significant one is vindicating a version of a principle about the independence of \textit{de se} and \textit{de dicto} information, the Relevance-Limiting Thesis. The principle is very plausible, but it is widely taken to clash with intuitive analyses of a widely discussed example, the so-called Sleeping Beauty scenario. On the contrary, the relocation framework both vindicates the Relevance-Limiting Thesis and captures pre-theoretical intuitions about the Sleeping Beauty case.

The connection between the third chapter and the previous two lies mostly in drawing a contrast between the cases of language and thought. In the third chapter, I compare intrapersonal update and communication of context-dependent thought. There are obvious structural similarities between the two, but, crucially, there is no analog of relocation for the case of communication. If this is right, the cognitive processes involved in thought are strictly more complex than those involved in communication.
1. Reference and Monstrosity

My mommy always said there were no monsters—no real ones—but there are.
—from ‘Aliens’

I want to introduce a puzzle about indexicals. The puzzle shows that I, you, now, and the like can fail to pick out elements of the actual circumstances of speech: for example, I can fail to refer to the speaker. Rather, indexicals sometimes work as bound variables, despite an entrenched dogma that they can never be shifted or bound. This has some interesting philosophical consequences—among them, that it vindicates a broadly Fregean perspective on referential expressions, refuting the idea that indexicals are rigid designators. Or so I say.

1.1 Overview

It is a truism that some expressions in language depend on other expressions for their interpretation. Definite descriptions are paradigm examples of this dependence. For example, in

(1) Juventus was the best team in Europe.

the past tense shifts backward the time at which the rest of the clause is evaluated. Thus the description the best team in Europe picks out not the current best team in Europe, but the best team in Europe at some time before the time of utterance. It is less of a truism that other expressions don’t conform to this model. In fact, the discovery that some expressions behave differently from the best team in Europe was a turning point which shaped contemporary theories of content and reference. Proper names like Juventus or
Titus Flavius Vespasianus, as Kripke (1980) famously pointed out, are expressions of this kind. Indexicals like I, you, now, and that, which are my focus in this essay, are another example. Indexicals seem to be insensitive to the effects of those operators that are able to manipulate descriptions. Here is a telling example from Kaplan (1989):

(2) It is possible that in Pakistan, in five years, only those who are actually here now are envied.

Actually, here, and now in (2) obviously pick out the world, location, and time of utterance, despite the presence of expressions (call them ‘operators’) that normally shift the world, location, and time at which a clause is evaluated.

The phenomenon displayed by (2) lends support to a simple and elegant theory. Indexicals are directly referential: their semantic contents are simply their referents. This distinguishes them from descriptions like the best team in Europe, whose contents are functions from shiftable parameters (world, time, etc.) to truth-values. This semantic difference explains the differences in linguistic behavior. Tense and modal operators can affect the contents of descriptions because the latter are sensitive to the time or world of evaluation. But the same operators are inert with indexicals because their contents make no reference to shiftable parameters.

The claim that indexicals are directly referential is at the heart of contemporary orthodoxy about reference. The orthodox picture is wrong. In this paper I argue that all indexicals can depend for their interpretation on other expressions. More precisely, indexicals are akin to variables of first-order logic. They can be free: in this case they invariably pick out elements of the context, like the speaker, the addressee, or the time of utterance. But they can also be bound: this happens when indexicals occur in the scope of epistemic modals or attitude verbs. One typical example is might. I claim that I in

(3) I might be in Chicago.

is a variable bound by might. It ranges over epistemic counterparts of the speaker, namely individuals who, for all the speaker knows, she might be. So (3) says that there is a possibility in which the speaker’s counterpart in that possibility is in Chicago. Kaplan dubs operators which shift or bind indexicals ‘monsters’ and bans them from the semantics of natural language (or at least, from the semantics of natural languages like English). So one pithy way to put my claim is that epistemic modals and attitude verbs are all Kaplanian monsters.

The ban on monsters is a cornerstone of the received view of context, reference, and semantic theory. Lifting this ban requires major changes to this picture. In this paper I discuss two of them. First, indexicals are not rigid designators: they are se-
1.2. EPISTEMIC SHIFT

1.2.1 Double indexing

As a start, let me survey the view of context dependence that is nowadays standard. Modern work on context dependence begins with the observation that theories of meaning need to track contextual parameters, such as the world or the time of utterance, in two distinct ways. This observation dates back to Kamp (1971) and is implemented formally by using double indexing, i.e. by relativizing the interpretation of object language expressions to two kinds of parameter. On the one hand, indexicals directly invoke context in their lexical meaning. For example, in

(4) Juventus is the best team in Europe now.

*now* directly latches on to the time of utterance. This is captured by relativizing the interpretation function (i.e. the function which maps expressions of a language into
their meanings, usually denoted by the double brackets ‘[ ]’ to a context parameter. The meaning of indexicals directly mentions that parameter; for example:

\[ \text{[now]}^c = \text{the time of } c \]

On the other hand, the semantics keeps track of contextual coordinates separately, via an \( n \)-tuple called \textit{index of evaluation}. The index is needed because certain expressions are sensitive to the time and world at which they are evaluated. For example, the description \textit{the best team in Europe} picks out a different referent at different times and worlds. So the interpretation function is relativized to an index of evaluation besides a context. Index coordinates can be shifted by words, like tenses and modals, that are often referred to as ‘operators’. For example, in

\[(1) \quad \text{Juventus was the best team in Europe.}\]

the past tense shifts backwards the time at which the rest of the clause is evaluated:¹

\[
\begin{align*}
\text{[Juventus was the best team in Europe]}^t & = \text{true iff for some time } t' \text{ before } t, \\
\text{[Juventus be the best team in Europe]}^{t'} & = \text{true}
\end{align*}
\]

The crucial claim is that these two forms of interaction between meaning and context are independent and can’t affect each other. In particular, operators that shift index parameters can never shift indexicals, as example (2) suggests. This offers support to the directly referential view. If, as direct reference theory claims, the contents of indexicals incorporate no reference to times and worlds, then it is immediately predicted that no operators can have semantic influence on them.

So much for the standard view. Now I start building my case against it.

1.2.2 Shifty conditionals

Consider the following scenario:²

¹Modern accounts of tense in semantics actually lean towards a different hypothesis: tenses are explicit quantifiers which bind object-language variables ranging over times. (See, among many others, King (2003) and Kusumoto (2005).) The point is irrelevant for my purposes; I’m using tense just as a convenient illustration of the functioning of operators.

²The examples that follow are original, but the characters are borrowed from classical literature on indexicality and reference. The amnesiac Rudolf Lingens is Perry’s creation (1977) and Gustav Lauben (who also appears in Perry’s paper) can be traced back to Frege (1918/1967). It should be flagged that I’m not the first to notice the ‘shifty’ properties of indicative conditionals: cf. Jackson (1987), Weatherson (2001), Nolan (2003), and Williamson (2006). These philosophers mostly focus on the capacity of indicative conditionals to shift \textit{actually} or natural kind terms (though, interestingly, Nolan has a clear case of epistemic shift involving \textit{that} and briefly considers a view that is very close to my account). The crucial observation is that, as I’m going
1.2. EPISTEMIC SHIFT

Rudolf Lingens and Gustav Lauben are kidnapped. Lingens and Lauben are amnesics: each of them knows that he is one of the two kidnapped amnesics, but doesn’t know which. They will be subjected to the following experiment. First, they will be anesthetized. Then a coin will be tossed. If the outcome is tails, Lingens will be released in Main Library, Stanford, and Lauben will be killed. If the outcome is heads, Lauben will be released in Widener Library, Harvard, and Lingens will be killed. Lingens and Lauben are informed of the plan and the experiment is executed. Later, one of them wakes up in a library. He says:

(5) If the coin landed tails, I am in Main Library, Stanford.
(6) If the coin landed heads, I am in Widener Library, Harvard.

Here is my starting datum: utterances of (5) and (6) are perfectly felicitous. Indeed, they seem to capture exactly what the lost amnesiac knows about his own predicament. If he were asked “Where are you?”, a joint utterance of (5) and (6) would be a good and informative answer. To better appreciate the point, it’s useful to contrast (5) and (6) with the corresponding counterfactuals. Suppose that the amnesiac lost in the library, trying to recapitulate the causal connections between possible outcomes of the coin toss and awakenings in libraries, says:

(7) If the coin had landed tails, I would have been in Main Library, Stanford.
(8) If the coin had landed heads, I would have been in Widener Library, Harvard.

There is a stark contrast in the acceptability of the two pairs. (7) and (8) are not good things to say in the scenario. They sound false, or at least inappropriate: in any case, they are clearly worse than (5) and (6).

My next observation is that a Kaplanian account of indexicals predicts that at least one of (5) and (6) should be infelicitous, contrary to the data. So (5) and (6) give rise to a puzzle for Kaplan’s theory.

This is the problem, intuitively: Kaplan’s account tells us that all tokens of I invariably refer to the speaker. So the two tokens of I in (5) and (6) should pick out the same individual. But two different individuals are in the relevant libraries in the two relevant scenarios: if I invariably refers to the speaker, it cannot pick out both of them. So, no matter who the speaker is, one of the two sentences should come out false.

Here is a more formal analysis of the difficulty. For illustration, take a standard version of possible worlds semantics for conditionals. All conditionals involve universal quantification over sets of possible worlds. The domain of quantification is provided by to argue, all indexical reference works in a non-standard way in epistemic contexts.
the context and the if-clause is used to specify a restriction on that domain.³ Conditionals like (5) and (6) have epistemic flavor: they quantify over a set of worlds representing an epistemic state. There is controversy about how this set is determined: here I use a simplified account, where the relevant epistemic state is just that of the speaker.⁴ So I take the truth-conditions of a conditional with epistemic flavor to be, schematically:

$$[[if \ p, q]^c_i = \text{true if and only if for all } w' \text{ compatible with what the speaker of } c \text{ knows at } i \text{ such that } p \text{ is true in } w', q \text{ is true in } w']$$

Now, fix a context: suppose that Lingens is the speaker. Assuming that I invariably denotes the speaker, the truth-conditions of (6) are

$$[[6]^c_i = \text{true if and only if for all } w' \text{ compatible with Lingens' knowledge that are such that the coin landed heads in } w', \text{ Lingens is in Widener Library in } w']$$

But now, in all the heads-worlds compatible with Lingens’ knowledge, the amnesiac lost in Widener is Lauben and Lingens is dead. So (6) is predicted to be false. Conversely, in contexts where Lauben is the speaker there will be a wrong prediction about (5). Notice that the problem is not just that certain utterances get the wrong truth-values. No matter what the context is, one of (5) and (6) is predicted to be false. Then a joint utterance of (5) and (6) should sound contradictory. But clearly it doesn’t.

There is an intuitive diagnosis of what happens in (5) and (6): I picks out not the actual speaker, but whatever individual is speaking in the circumstances singled out by the antecedent. In short, the referent of I seems to shift on the basis of the antecedent of the conditional. It’s convenient to have a name for the phenomenon displayed by the (5)–(6) pair. Drawing on this intuitive diagnosis—and without, for the moment, making any commitments to its correctness—let me call it ‘epistemic shift.’ My main concern in this paper is to explain epistemic shift and draw some general consequences for semantic theory.

³Syntactically, these truth-conditions are achieved by assuming the existence of a covert, i.e. unpronounced, modal quantifier taking scope over the whole conditional. The locus classicus for this theory is Kratzer (1981) and (1991). Kratzer builds her account on the classical work of Stalnaker (1968) and Lewis (1973); for a state-of-the-art version of this semantics, see von Fintel & Heim (2010).

⁴Here I’m skirting an ongoing debate in philosophy of language: different schools of thought (contextualism, relativism, expressivism) disagree about the exact mechanisms through which this parameter is determined. (For representative positions, see DeRose (1991), MacFarlane (2008), and Yalcin (2007).) But for the purposes of this puzzle the issue is irrelevant; the puzzle can be reproduced on any of these positions.
1.2. Three quick replies

Let me forestall three quick attempts at accounting for epistemic shift.

Non-possible worlds theories. The first is that the puzzle can be solved simply by switching to a non-possible worlds theory of conditionals. One obvious candidate is the material conditional analysis championed by Lewis (1976) and Jackson (1987). This analysis identifies the truth-conditions of indicative conditionals with the truth-conditions of corresponding material conditionals of first-order logic. Interestingly, on this account both (5) and (6) come out true.⁵ So it would seem that endorsing it is sufficient to defuse the puzzle.

This conclusion is too fast. As is well-known, the material conditional analysis declares true a large number of indicative conditionals that are infelicitous. So the semantics needs to be supplemented with an account of the assertability of these conditionals. I claim that the puzzle is going to resurface at this level. Jackson (1987), for example, claims that an indicative conditional is assertable just in case the speaker assigns a high credence to the consequent, conditional on the antecedent. So, in our scenario, (6) is assertable just in case the speaker has a high conditional credence that Lingens is in Widener, given a heads outcome. But the speaker assigns little or no credence to the proposition that Lingens is in Widener, hence (6) is declared to be unassertable, contrary to fact.

An analogous point holds for no-truth-value theories à la Edgington (1995), on which assertions of indicative conditionals are analyzed as speech acts of conditional assertions. These theories also use conditional probabilities to make predictions about assertability of conditionals. But again, (6) is assertable, even though on standard assumptions about indexicals it is predicted to have low or zero probability. More generally, it seems that the puzzle can be recreated on virtually any theory of epistemic conditionals, on the minimal assumptions that (a) the semantics of indexicals is Kaplanian, (b) both conditionals are true or at least appropriate, and (c) the speaker in the scenario has no inconsistent beliefs.

Haecceities. The second attempted solution is condensed in the following line of thought:

The puzzle can be solved by resorting to an ontology which includes non-qualitative individual essences or haecceities.⁶ Suppose that I invariably

---

⁵Suppose that Lingens is the speaker. Then the coin landed tails and the antecedent of (6) is false at the actual world. On the material conditional account, this is sufficient to make the whole conditional true. Similarly, mutatis mutandis, if the speaker is Lauben.

⁶Non-qualitative individual essences, or haecceities, are first introduced in connection with direct reference theory by Kaplan (1975). See also Lewis (1988), pages 220-248 for discussion.
refers to the individual with the haecceity of the actual speaker. And now, assume that the individual with the speaker's haecceity has Lingens's qualitative properties in tails-worlds and Lauben's qualitative properties in heads-worlds. Then both (5) and (6) come out true.

The basic idea is that \( I \) in (5) and (6) invariably refers to one individual, namely the actual speaker. However, that individual (singled out on the basis of his haecceity) has very different qualitative properties in the two relevant scenarios: in tails-scenarios he has Lingens's qualitative properties, in heads-scenarios Lauben's.

I grant that this strategy accommodates my original example (though at the cost of buying into a controversial metaphysical thesis). But the solution doesn't generalize: it's easy to produce examples that resist it. One way to do this is to use sentences involving two indexicals:

Things are as in the original amnesiac scenario, but for one tweak. In case of a tails outcome, Lingens will wake up in front of a mirror. In case of a heads outcome, Lauben will wake up in front of a glass window, with a perfect replica of himself (perhaps a twin, perhaps a clone) on the other side. And now suppose that the waking amnesiac points to the individual in front of him and says:

(9) If the coin landed tails, I am you.
(10) If the coin landed heads, I am not you.

Both conditionals are felicitous and, as before, a standard semantics misses the prediction. But here haecceities don't help. If the amnesiac is in fact looking at a mirror, then \( I \) and \( you \) both pick out the very same individual, hence only one haecceity is in play. But then (10) is false. If the amnesiac is in fact looking at his replica, then \( I \) and \( you \) pick out two distinct individuals, hence different haecceities are called into question. Hence (9) is false.

The haecceitistic route can provide at most a local fix to some of my examples, but not a general one. So I discard it. An account of epistemic shift calls for adjustments in the semantics of indexicals, rather than in our metaphysics.⁷

Vacuity. The third attempt flatly denies that one of the two conditionals is false on the standard semantics. Rather, both are true, even though one is only vacuously true.⁸

---

⁷Thanks to Agustín Rayo for pushing me to consider this objection.
⁸Thanks to an anonymous referee for both mentioning this objection and suggesting the response.
One can get this conclusion by claiming that, just by having the belief he would express by *I am here*, the surviving amnesiac can obtain a non-trivial belief about where he is. For example, just by believing the proposition he would express by *I am here*, Lingens could come to believe that Lingens is in Main Library. In this case, Lingens would also know that the coin landed tails, since only in that case he would make it alive to the library. But then (6) would be vacuously true, since its antecedent would be false in all of Lingens's knowledge-worlds.

This is a radical route to take, since the claim that Lingens knows the outcome of the coin toss is obviously implausible. But even granting this, there is a conclusive objection to this line. Occurrences of epistemic shift are not limited to conditionals. Notice that, rather than (5) and (6), the amnesiac could say simply:

(11) I might be in Main Library, Stanford.
(12) I might be in Widener Library, Harvard.

(11) and (12) are equally problematic for Kaplan's theory. By assumption, the speaker knows that only Lingens could be in Main Library, and only Lauben could be in Widener Library. But then a joint utterance of (11) and (12) is problematic on the assumption that *I* invariably refers to the speaker. The speaker should know that, no matter who he is, one between (11) and (12) is false.

1.2.4 Generalizations: other indexicals, other modals

It's time to chart some empirical ground. So far I have only talked about *I*; now I turn to investigating how far and along what dimensions the puzzle generalizes. As I'm going to point out, epistemic shift is a pervasive phenomenon that concerns all indexicals and all operators with epistemic flavor.

To start with, two quick observations. The first is that epistemic shift generalizes to all indexicals, including demonstratives. The amnesiac scenario is obviously contrived. But this is only because scenarios where speakers are unaware of their own identity are hard to come by. We can construct less far-fetched variants of the puzzle by switching to

---

⁹Moreover, the fact that the antecedent of (6) is known to be false raises questions about its felicity. Stalnaker (1975) points out that indicative conditionals whose antecedents are incompatible with what is commonly known by the speakers are infelicitous. But perhaps the supporter of vacuity can claim that in this case knowledge doesn't iterate: Lingens knows that the coin landed tails, but he doesn't know that he knows.

¹⁰Even though the data involving possibility modals is relevant, I will keep using (5) and (6) as my running examples. I have two reasons. First, conditionals make a clear case for my analysis, suggesting indexical shift as an intuitive diagnosis. Second, conditionals are not amenable to a pragmatic treatment of epistemic shift (see footnote 12). So they make a better case for a semantic analysis.
more prosaic varieties of ignorance. Here is one involving temporal indexicals. Suppose that you fell asleep at noon, and wake up without knowing whether you slept one or two hours. You say:

(13) If I slept one hour, it is now one.
(14) If I slept two hours, it is now two.

The problem, again: orthodox semantics wrongly predicts one of (13) and (14) to be false, no matter what time it is.

The second observation is that shifty readings are absent in conditionals with non-epistemic flavor. I already contrasted (5) and (6) with the corresponding counterfactuals. Now let me point out that the contrast extends beyond counterfactuals. Philosophers tend to group all indicative conditionals together. But it is standard in formal semantics to assume that at least some indicative conditionals involving will express the same kind of modality that is expressed by counterfactuals. Following this assumption, it is unsurprising that, in the amnesiac scenario, we can formulate will-conditionals that pattern with counterfactuals. Suppose that, after having been informed about the experiment but before undergoing it, one of the amnesiacs says:

(15) If the coin lands tails, I will be in Main Library, Stanford.
(16) If the coin lands heads, I will be in Widener Library, Harvard.

Like (7) and (8), (15) and (16) are not good utterances.

This establishes that epistemic shift concerns specifically epistemic conditionals. Can it generalize beyond conditionals? I already showed that it can be replicated under existential epistemic modals. It’s natural to suspect that it can take place under other verbs with epistemic flavor: for example, attitude verbs like believe. The issue is not straightforward.

On the one hand, I cannot produce data that directly establishes that epistemic shift obtains under believe. On the other, we are familiar since Frege (1892/1997) with the idea that referential expressions, including indexicals, are problematic in belief reports. Here is an illustration: suppose that Lingens mistakes a mirror image of himself for a different individual—perhaps Fidel Castro, who wears a beard similar to his. Then he can truly utter, addressing the individual he’s seeing:

(17) I believe that I am lost in a university library.

¹¹This analysis is substantiated by the morphological claim that will and would both realize a unique modal auxiliary, usually referred to as woll. (For discussion, see, among others, Abusch (1988), Condoravdi (2002), Werner (2006).) Notice that the view still leaves room for some will-conditionals to be epistemic.
I believe that you are not lost in a university library. Despite the fact that in the context I and you pick out the same individual. Using a standard label, indexicals display failure of substitutivity in attitude reports. Even though the specific form of the two puzzles is different, it is natural to suspect that both epistemic shift and failures of substitutivity are symptoms of a unique underlying difficulty concerning verbs and operators that have epistemic flavor.

I endorse the kinship between the two puzzles and look for a unitary account. My main motivation is that there is a conceptually and empirically uniform account that manages to explain both phenomena. Just the existence of this account provides evidence that the two are closely related. Moreover, there is an empirical connection between the two: in particular, it's possible to recreate a restricted kind of epistemic shift under attitude verbs. This connection emerges not with indexicals, but rather with other pronouns, like she and he: I present the relevant data in section 1.5.

These last remarks point out that epistemic shift is not confined to indexicals. In fact, it can be reproduced with virtually any kind of referential expression, including names and deictic pronouns. But taking up the task of rewriting the semantics for all these expressions would take me too far. For this reason, throughout the bulk of the paper I'm going to focus on indexicals. I will point out, in section 1.5, how my account can be used as a blueprint for building parallel semantics for other referential expressions.

1.2.5 Review

I have introduced epistemic shift by means of epistemic conditionals involving I. Then I have observed that the phenomenon generalizes to all indexicals. I have shown that epistemic shift is tied specifically to epistemic conditionals and disappears when we switch to other conditionals. Finally, I have suggested that epistemic shift is part of a broader phenomenon concerning verbs and operators that represent information states.

The fact that epistemic shift is systematically linked with a particular class of expressions is evidence that it is a semantic and not a pragmatic phenomenon. The natural hypothesis is that the semantics of might, must, believe and the like involves some undiscovered feature that is responsible for the shifty behavior of indexicals in their scope. This is the kind of account I pursue in the next sections.¹²

¹²Notice, in particular, that the problem can't be solved by using Stalnaker's (1978) pragmatic diagonalization strategy. This strategy consists in taking the proposition communicated by a sentence S in a context to be the so-called diagonal, i.e. (roughly) the proposition true at those possibilities p where the content expressed by S as uttered in p is true.

To see that this can't help, consider the diagonal proposition communicated by (6) in the amnesiacs scenario. Since we're considering indexicals, the diagonal is a function from contexts to truth-values (following Lewis (1980b), just take contexts to be centered worlds). Hence the
1.3 Informational modals are monsters

1.3.1 Preliminaries

Recall the intuitive diagnosis of the puzzle. In conditionals like (5) and (6), \( I \) picks out not the actual speaker, but whatever individual is speaking in the circumstances singled out by the antecedent. In short, the referent of \( I \) seems to shift on the basis of the antecedent of the conditional. The account I propose endorses and develops this intuition. To be sure, my claim will be that \( I \) in (5) and (6) is technically a bound variable, hence it’s not used to refer to a particular individual. But the intuitive diagnosis captures perfectly the main insight: the antecedents of (5) and (6) have semantic effects on \( I \). Similarly for other verbs with epistemic flavor, like \( \text{believe} \) and \( \text{might} \), and for all indexicals.

My presentation of epistemic shift has been fully framework-neutral. In particular, I made no assumptions about the semantics of conditionals, modality, or belief. To state an account of the phenomenon, I must abandon this neutrality. I help myself to two main assumptions. First, some kind of possible worlds account of epistemic conditionals is correct. Second, attitude verbs like \( \text{believe} \) should be analyzed, in the fashion of Hintikka (1962), as modal operators ranging over possible worlds compatible with the subject’s attitudes. I choose these assumptions simply because they capture the standard treatment of conditionals and attitude verbs in compositional semantics. But let me stress that, although I use them to state my account, they are not essential to it. My central claims—that indexicals work as variables, that they are systematically bound in epistemic environments, and that they range over epistemic counterparts of their referents—are still framework-neutral and can be combined with different accounts of conditionals and attitude verbs.

It’s useful to have a unique phrase to denote all the expressions I’m concerned with. I use ‘informational modals’ as a blanket label for the modals which report, describe, or express subjects’ attitudes. Typical cases of informational modals are \( \text{might} \), \( \text{believe} \), and epistemic \( \text{must} \); but also verbs like \( \text{suppose} \), \( \text{imagine} \), \( \text{desire} \), or \( \text{wish} \) will be in this diagonal of (6) is

\[
\lambda c. \lambda w'. [\lambda w. T (6)]^{c_{w'}} = \lambda c. \text{for all } w' \text{ compatible with the speaker’s knowledge in } w, \text{ that are such that the coin landed heads in } w', \text{ the speaker of } c \text{ is in Widener Library in } w'.
\]

Now, let the speaker of the context be Lingens (let the context be \( \langle w_T, t, \text{Lingens} \rangle \), where \( w_T \) is a tails-world). The centered proposition above is false in this context, since in no world compatible with Lingens’s knowledge does Lingens end up in Widener library.

So diagonalization fails. What goes wrong? Intuitively, the problem is that (as it happens in the standard Kratzer semantics) the antecedent of the conditional has no semantic effect on \( I \). But what we want is just to let the antecedent determine the context in which \( I \) is evaluated. This, of course, means taking the monstrous route.
1.3. INFORMATIONAL MODALS ARE MONSTERS

This section is devoted to giving my account of informational modals. I do this relatively informally and defer the heaviest technical details to the appendix.

1.3.2 Variables over counterparts

Epistemic shift and substitutivity puzzles arise in connection with modals that describe (or report, express, etc.) information states of various sorts, i.e. mental states which have representational content. The question I’ve been investigating is how indexical reference works in the scope of these modals. Epistemic shift and substitutivity puzzles show that the answer suggested by standard theories—that indexicals invariably contribute their referents to truth-conditions—is inadequate. What are the alternatives?

Here is a suggestion. When I occurs under an informational modal, it refers not to the actual speaker, but rather to representatives of the actual speaker in the relevant information state. Think of an information state as a set of possible worlds, i.e. the worlds that are compatible with the relevant attitude. Now, suppose that the subject of the information state has some attitude about the speaker of a certain context: for example, he thinks that the speaker has a flowing beard. Then in all the worlds compatible with his beliefs there will be a flowingly bearded person representing the speaker. Roughly, these representatives are individuals that, for all the subject knows, the speaker might be. I will elucidate the notion of a representative shortly: before that, let me illustrate the idea via an example.

Take again the library scenario and suppose that Lingens, who is in fact the man lost in the library, meets a second amnesiac, Herman. Lingens tells Herman about the coinflip scenario and Herman expresses sympathy for Lingens’s predicament. When they part, Lingens says:

(19) Herman believes that I am in a sad predicament.

Who does I in (19) refer to? Upon meeting Lingens, Herman has come to think of him in a number of ways: for example as the other amnesiac lost in the library or the survivor of the gruesome coinflip experiment. Of course, a number of metaphysically different individuals fit these properties. For example, it is compatible with Herman’s knowledge that Lingens is the survivor of the coinflip experiment, and also that Lauben is. I claim that these individuals act as representatives of Lingens in Herman’s information state. It is these individuals that the occurrence of I in (19) denotes.

Let me clarify two points. First, the talk of I denoting a multiplicity of individuals is obviously metaphorical. As I anticipated, the precise statement of the point is that I is semantically a bound variable. The individuals representing Lingens in the relevant epistemic state will constitute its range.
Second, the informal talk of representatives can be made precise by switching to talk of epistemic counterparts. In general, a relation of counterparthood is simply a relation of similarity (see Lewis (1968) and (1983)): \( x \) is a counterpart of \( y \) under a certain respect just in case \( x \) and \( y \) are sufficiently similar in that respect. More specifically, epistemic counterparthood is a three-place relation of similarity (\( x \) is an epistemic counterpart of \( y \) for a subject \( S \)) which captures a way a subject thinks of a certain object. \( y, z, \ldots \) are epistemic counterparts of \( x \) for \( S \) just in case (a) \( S \) has beliefs about \( x \) and (b) \( y, z, \ldots \) possess all the properties that \( S \) attributes to \( x \).¹³ Epistemic counterparts are generally used, within possible worlds theories of mental content, to 'locate' actual world individuals within belief worlds of subjects.

For illustration, consider again the Lingens-Herman example. Model the information in Herman's belief state as a set of possible worlds, namely the worlds that are compatible with what Herman believes. In all those worlds, Herman meets an individual in the library. In some of those worlds, that individual is Lingens; in others it is Lauben. But all of them have certain features in common: in particular, they all look a certain way, have a certain conversation with Herman, and so on. Those individuals are Lingens's epistemic counterparts in Herman's belief state.

In short, then, I suggest that we export the notion of an epistemic counterpart from possible worlds accounts of mental content to possible worlds semantics for epistemic operators. Bound indexicals range over epistemic counterparts of their referent in the actual context. For example, \( I \) ranges over epistemic counterparts of the actual speaker; \( you \) ranges over epistemic counterparts of the addressee; and so on.¹⁴

Let me clarify what commitments come with my proposal. I am committed to the claim that a notion of epistemic counterparthood is hard-wired in the semantics of epistemic operators. But I am not committed to any claims about mental representation and mental content. In particular, I am not committed to the idea that the possible worlds model provides the best, or even a good way of representing mental content. There is an obvious analogy, both formal and conceptual, between possible worlds accounts of attitudes and possible worlds semantics for attitude reports. I exploit this analogy, but I can do this without endorsing a possible worlds account of attitudes.

¹³Notice that Lewis (1983) gives a more specific statement of the notion of epistemic counterparthood. He identifies epistemic counterparthood with counterparthood by acquaintance, where the notion of acquaintance captures a kind of direct epistemic relationship between a subject and an object. To get his definition, we should substitute clause (a) of my definition with the clause: (a') \( S \) bears the acquaintance relationship \( R \) to \( x \) and believes that he is \( R \)-acquainted with \( x \). I prefer my definition because it doesn't commit me to a specific way of cashing out epistemic counterparthood.

¹⁴Interestingly, Hintikka (1969) makes something like this proposal for the case of names in pre-Kripkean times. His idea, though, seems to have fallen out of consideration by the time of Lewis's (1983).
1.3. INFORMATIONAL MODALS ARE MONSTERS

Similarly, I am not committed to a specific way of cashing out the notion of an epistemic counterpart. I have defined epistemic counterparthood by appealing to a notion of a subject having beliefs about a specific object. This notion is obviously, and conveniently, vague. Spelling out this notion requires taking sides in a number of debates about content and de re attitudes. For example, it requires choosing between a reductionist and a non-reductionist position about de re attitudes. (On the former, de re attitudes turn out to be a special kind of self-locating attitudes, while on the latter they don’t: see, among others, Lewis (1983) and Ninan (2008).) Obviously this debate might have an impact on the semantics of epistemic operators; but this impact doesn’t need to be assessed now. Here I state a general semantic framework that can be combined with different views about epistemic counterparthood.

Finally, I am not committed to a metaphysics that employs counterpart theory to model identity of individuals across possible worlds. I advocate the use of certain counterpart relations to model the range of some bound variables in the semantics. This is fully compatible with a background metaphysics where the same individuals exist at different worlds, as argued by Kripke (1980). The problem of trans-world identity concerns the nature of possibilities. The problem of determining the range of bound indexicals concerns how possibilities—once we’ve established what they are—should be used to model linguistic content. The two issues are to a large degree independent and I need not take a stance on the former here.

This gives a general introduction to the view. Now let me be more specific about the new semantics for epistemic operators.

1.3.3 Assignment shift

The suggestion that might, believe, and the like bind indexicals is grounded in a more general shift in the way of thinking of informational modals. On the standard modal view, informational modals are, in essence, quantifiers over possible worlds. On the view I’m advocating, they also encode in their meaning an apparatus that locates real-world individuals within the set of worlds quantified over. Thus on the new picture these modals manipulate a greater amount of information. The classical picture had them quantify over a set of worlds connected to the actual world via an accessibility relation; on the new picture, they quantify in addition over counterparts of actual individuals in each of the worlds in the set.

The new picture is implemented by letting epistemic operators manipulate the assignment parameter, which I construe here as a sequence of individuals.¹⁵ The general

¹⁵More often, the assignment is modeled as a function from syntactic indices of pronouns to individuals. The two formulations are equivalent (since one can just use the order in the sequence to recover the indices). I choose sequences merely to avoid cumbersomeness.
idea behind this is quite natural. When we are characterizing an information state, we cannot use referential expressions to pick out directly elements of the context. We need to 'locate' the relevant individuals in the information states. In a slogan, reference within epistemic contexts is always vicarious reference: it always passes through representatives. As a result, modals don’t act only on the world of evaluation, but rather have an effect on the whole referential apparatus of the language (at least, on the apparatus that handles reference to elements of the context). This apparatus is shifted in such a way that indexicals pick out representatives of the relevant individuals at each world.

Notice that ordinary quantifiers like every and some are also assignment shifters. But every and some shift the assignment in a piecemeal way, operating on one variable at a time. Epistemic operators force a shift of the whole assignment at once. In addition to this, they coordinate the shift of the assignment with the shift of the world parameter, in such a way that a world is always paired with the assignment involving epistemic counterparts of individuals at that world. Let me explain.

For illustration, I use a standard semantic framework, where the interpretation function is relativized to a context, a world, and an assignment (though, as will become clear in section 1.4, the context parameter is superfluous on the new semantics). On orthodox assumptions, the effect of might is a shift of the world parameter. This means that the clause in the scope of the modal is evaluated at a different world:

\[ [\text{might } \phi]^{c,w,g} = \text{for some accessible } w', [\phi]^{c,w',g} \]

On the new semantics, modals shift both the world and the assignment parameters:

\[ [\text{might } \phi]^{c,w,g} = \text{for some accessible world-assignment pair } \langle g', w' \rangle, [\phi]^{c,w',g'} \]

It should be clear how this lets modals have semantic effects on indexicals. On the new picture indexicals are variables. So their referent is determined by the assignment:

\[ [I]^{c,w,g} = [x_1]^{c,w,g} = g(1) \]

Modals shift the assignment, hence they have the potential for shifting indexicals. (This is obviously not enough to anchor indexicals to the context; I will say more in the next section.)

This establishes how informational modals function compositionally. But it doesn’t settle what the world-assignment pairs accessible from any given point are. This is a key issue, since without settling it we cannot determine truth-conditions.

It is at this point that relations of epistemic counterparthood come in. Consider first a toy case: pretend that the assignment only settles the value of one variable, namely I. In line with some accounts of belief reports in semantics (e.g. Heim (1994)), I assume that context supplies a counterpart function \( f_I \) which pairs each world in the information
1.3. INFORMATIONAL MODALS ARE MONSTERS

state with the counterpart of the speaker. Then we can specify a set of world-assignment pairs:

\[ \langle f_I(w_i), w_i \rangle \]
\[ \langle f_I(w_j), w_j \rangle \]
...

on which the object in the assignment is the counterpart of the speaker in the world.

And now drop the pretense that the assignment only handles one variable: the suggestion is that we can use a whole sequence of counterpart functions to generate assignments starting from worlds. Here is the idea: context supplies us with a sequence of counterpart functions, one for each pronoun in the language: \( f_I, f_{you}, \ldots, f_n \). This sequence tracks the way the subject thinks of each element of the context: \( f_I \) tracks the way she thinks of the speaker, \( f_{you} \) tracks the way she thinks of the hearer, and so on.¹⁷

We use these counterpart functions to ‘project’ assignments out of a world. Here are the sorts of world-assignment pairs we get by following this procedure:

\[ \langle \langle f_I(w_i), f_{you}(w_i), \ldots, f_n(w_i) \rangle, w_i \rangle \]
\[ \langle \langle f_I(w_j), f_{you}(w_j), \ldots, f_n(w_j) \rangle, w_j \rangle \]
...

Notice that this mechanism produces a simultaneous shift of the whole assignment at once. Every object in a shifted assignment is an epistemic counterpart of some object in the original assignment.

The net effect is that epistemic operators work as binders of all indexicals occurring in their scope. They shift the whole apparatus of contextual reference of the language at once. This massive effect explains why epistemic shift is replicated with virtually all indexicals in the language (and even when two indexicals appear in the same sentence, as in (9) and (10)).¹⁸

¹⁶For reasons having to do with the nature of epistemic counterpartship (see Lewis (1983)), these worlds will actually have to be centered worlds. For simplicity, I’m going to ignore this point throughout the paper.

¹⁷Notice that using functions of this kind in the semantics doesn’t require the (dramatically unrealistic) assumption that the subject genuinely has a way of thinking of all objects in the context. If the subject has no way of thinking of some element of the context, we just use ‘dummy’ functions that are everywhere undefined. Hence the context must only provide us with counterpart functions for the objects that the speaker actually has attitudes about.

¹⁸I discuss more extensively the new semantics for modals in the appendix and in Santorio (2010), which is a technical companion to this paper. Let me point out that my semantics shares a number of features with the semantics that Cumming (2008) gives for names. Despite the similarities, there are important conceptual and technical differences; I briefly discuss them in section 1.5.1.
Finally, let me consider how this machinery accounts for the puzzle. Consider again the problematic (5):

(5) If the coin landed tails, I am in Main Library, Stanford.

Make one plausible assumption: in the case of epistemic conditionals, the counterpart function $f_I$ that determines the range of $I$ picks out the possible individuals that the subject of the information state takes himself to be. I.e., in the case of (5), $f_I$ specifies the epistemic counterparts of the speaker, for the speaker himself.¹⁹ Then the truth-conditions that we get for (5) are:

$$[(5)]^{c,w,g} = \text{true iff for all worlds } w' \text{ compatible with the speaker's knowledge that are such that the coin landed tails in } w', \text{ the individual that the speaker takes himself to be in } w' \text{ is in Main Library, Stanford in } w'. $$

which is exactly right: the individual who's relevant for evaluating the conditional is not the actual speaker, but rather the person who's speaking in the circumstances singled out by the antecedent. Notice that the effect of treating indexicals as variables and letting them range over counterparts is that they contribute to truth-conditions a function, rather than just a referent. This is what solves the puzzle. The account also accommodates in a natural way substitutivity puzzles. Consider again the mirror scenario: Lingens, who mistakes a mirror image of himself for the likewise bearded Fidel Castro, utters

(17) I believe that I am lost in a university library.

(18) I believe that you are not lost in a university library.

Believe helps itself to different counterpart functions for different indexicals, even if these indexicals corefer. So $I$ in (17) and $you$ in (18) can be associated to different sets of counterparts, despite the fact that they happen to refer to the same individuals.

1.3.4 Reference without rigidity

Turning modals into monsters challenges the doctrine that indexicals are rigid. It's time to make this claim precise and explore some of its consequences.

¹⁹This assumption is independently plausible. $f_I$ must specify a way the subject of the epistemic state thinks of the actual referent of $I$. But now, by using the first-person pronoun the speaker is flagging that she is aware that she’s achieving self-reference. (Contrast my uttering My pants are on fire! with my pointing to the mirror and saying His pants are on fire! In the first case, though not in the second, you can infer, from the pronoun used, that I’m aware that I’m referring to myself.) So it’s plausible that, at least in normal cases, the counterpart relation used in an utterance of (5) defaults to one that captures the speaker’s de se way of thinking of himself.
1.3. INFORMATIONAL MODALS ARE MONSTERS

The rigidity doctrine has it that, relative to a choice of a context and an assignment, indexicals refer to the same individual with respect to all possible worlds.²⁰ Now, on the picture I suggest all semantic manipulation of indexicals passes just from the assignment. So rigidity is literally preserved on the new picture. Indexicals and free pronouns do pick out the same individual at all worlds, given an assignment; it’s just that, contrary to the orthodox picture, the assignment is sometimes shifted.

But this vindication of rigidity is vacuous. There is substantial failure of rigidity. Epistemic operators shift together, and in a coordinated way, both the world and the assignment parameter. In particular, the value of the assignment parameter is determined as a function of the value of the world parameter. The net effect is that indexicals are sensitive to the world parameter, and their semantic value does change as a result of shifts in the world of evaluation. This is clearly unexpected on the standard picture of indexicality. Contra the orthodox view, the referent of I can be different at different worlds. Correspondingly, the new picture takes a step in a Fregean direction at the level of truth-conditions. In epistemic contexts, the overall contribution of an indexical to truth-conditions is more informative than its referent. The effect of monstrous semantics is that indexicals are paired with (generally non-constant) functions from worlds to individuals.²¹

Let me emphasize that, despite the denial of rigidity, the new picture doesn’t retreat to descriptivism: it doesn’t assimilate the functioning of indexicals to the functioning of descriptions. In essence, the monstrous account separates claims about the compo-

---

²⁰I ignore the question of obstinacy, i.e. the question whether indexicals refer to the same individuals even in worlds where those individuals don’t exist (see Kaplan (1989), section IV). The question is irrelevant for my purposes.

²¹One might wonder whether this is really a major step. Isn’t the abandonment of rigidity already implicit in the use of counterpart theory? If our background metaphysics uses counterpart theory to model cross-world identity, indexicals inevitably refer to different individuals in different worlds. The reply is twofold. First, as I pointed out, a semantics which appeals to epistemic counterparthood is fully compatible with a non-counterpart-theoretic metaphysics. In this case, the denial of rigidity is clearly non-trivial. Indexicals embedded under informational modals pick out different individuals at different worlds, even though the same individuals exist at those worlds. Hence embedded indexicals work very differently from unembedded ones, which keep referring rigidly to the same individuals at all worlds. Second, consider the case in which the background metaphysics is counterpart-theoretic. We can still define a notion of rigidity: following Lewis (1988), a term is quasi-rigid (with respect to a context and an assignment) just in case in all worlds it refers to metaphysical counterparts of the individual it refers to in the actual world. Denying quasi-rigidity amounts to claiming that different kinds of counterparts relations (epistemic rather than metaphysical) are employed when indexicals occur under informational modals. This still seems an important point: among other things, it introduces a perspectival component in the semantics of indexicals, since epistemic counterpart relations are relative to a subject’s beliefs. But I’m happy to grant that ultimately one should find the denial of quasi-rigidity significant only insofar as one finds the switch from metaphysical to epistemic counterparthood significant.
sitional semantic values of indexicals from claims about the final truth-conditions of sentences involving indexicals. It agrees with direct reference theory about the former. The compositional semantic values of indexicals and other pronouns differs from the semantic value of descriptions. The two kinds of expressions work in different ways and (insofar as their lexical meanings are concerned) mention different kinds of parameters. For example, the meaning of I mentions the assignment, while the meaning of a description mentions the world of evaluation:

\[
[I]^{c,w,g} = g(1) \\
[\text{the speaker}]^{c,w,g} = \text{the person speaking at } w
\]

One consequence of this is that the monstrous account agrees with direct reference theory whenever indexicals are not embedded under epistemic operators. However, the monstrous account diverges from the orthodoxy in that it assumes that the semantics of natural language has the resources to manipulate the two parameters together, in a coordinated way. These extra resources, which are built into the semantics of epistemic operators, make it the case that, at the level of truth-conditions, indexicals can be paired with functions rather than just objects. This is exactly what a broadly Fregean approach would predict. As a result, even though their compositional semantic value is different, I and the speaker can provide the same truth-conditional contribution (this is arguably the case in (5) and (6)).

One lesson of the foregoing is that a semantic notion of cognitive significance can be combined with the basic ideas behind direct reference theory. The monstrous account exemplifies this combination. Independently of whether it is successful, it’s significant that a view of this kind is available.

1.4 Monstrous semantic theory

1.4.1 Semantics, context, and content

Section 1.3 presented a semantics that can predict and explain epistemic shift. But this doesn’t exhaust the task I have undertaken in this paper. Indexical reference is one of the cornerstones of a theory of context dependence and semantic content. Changing our theory of indexicals requires changes to the general architecture of this theory; and this, in turn, has potential repercussions for the role we assign to our semantic theory within a broader theory of cognition. In this section I turn to these issues. This will also allow me to complete my account of indexicals, explaining how I, you, now, and the like are anchored to the context of utterance when they occur free.
1.4. MONSTROUS SEMANTIC THEORY

I already introduced the idea that the semantics of natural language needs double indexing, i.e. it needs to keep track of contextual parameters in two different ways. As I said in section 1.2, orthodox views do this by relativizing the interpretation function both to a context and to an index of evaluation. This version of double indexing is normally combined with a general view of the interactions between context and meaning which is due to Kaplan (1989). Let me illustrate this view in some detail.

Kaplan’s key idea is that the distinction between the context parameter and the index corresponds to a distinction between two points of interaction between semantics and context. On the one hand, the semantics accesses contextual information at the lexical level, via the meanings of indexicals. Crucially, this kind of access takes place before compositional interactions. For example, in

(20) I am lost in a university library.

*I* first ‘grabs’ its referent, and then feeds it into the process that recursively computes the truth-conditions of the sentence. Hence we need access to contextual information (for example, we need to determine who the speaker is) to determine the input to compositional semantics. We need to look at the context in order to run compositional semantics.

On the other hand, context is also invoked at a different stage. Recall the functioning of operators like tenses and modals. They work by shifting the coordinates at which other expressions are evaluated. For example, in

(1) Juventus was the best team in Europe.

the past tense effects a kind of ‘backward shift’ of the time at which the clause *Juventus be the best team in Europe* is evaluated. Backward with respect to what? Normally, with respect to the time of utterance. The initial value of the time parameter is set to the time of utterance; it is from that time that the shift effected by operators proceeds. This process of initialization is performed at a different stage of a theory of meaning: following MacFarlane (2003b) and (2005), I call this stage ‘postsemantics.’ Postsemantics is usually placed after compositional semantics, and in any case is independent of it. We don’t need to initialize index parameters to carry on a compositional computation of truth-conditions.

Below is a diagram summarizing Kaplan’s picture. I use ‘semantic value’ to denote the kind of meaning that is handled by compositional semantics.
It’s worth emphasizing that this general view is not entailed by double indexing; there are alternative ways of implementing the latter. For example, Lewis (1980b) points out that we could use functions from contexts and indices to truth-values as the compositional unit. In that case, we would not need to access context at two separate stages. But Kaplan’s implementation is often taken as the standard both in the philosophy and the semantics literature.²²

Notice that this general view of context, meaning, and semantics is closely bound up with the prohibition against monsters. The mere setup, in fact, ensures that there cannot be operators that have semantic effects on indexicals. This is easy to see. Kaplan assigns indexicals a referent before they interact compositionally with other expressions. So indexicals can’t display the same kind of sensitivity to time, world, or other parameters as, say, descriptions do. The semantic values of descriptions are functions from parameters like times or worlds to referents; but the semantic values of indexicals are just referents: they involve no mention of shiftable parameters. Hence, once the context is fixed, the referents of indexicals are fixed once and for all.²³

Given that the standard setup stipulates away monsters, monstrous semantics requires rethinking some general aspects of the picture. I do this in two stages. First I focus on indexicals and show how they can be anchored to context on the new view.

²²An excellent illustration of this is given by the recent literature on truth-relativism: see, among others, MacFarlane (2003a) and (2008), Lasersohn (2005), Stephenson (2007). This literature aims at extending the standard apparatus to a non-orthodox picture: the benchmark view of context, content, and semantic parameters from which this extension proceeds is invariably that proposed by Kaplan in Demonstratives.

²³Indeed, the formal theory given in Demonstratives (section XVIII) doesn’t incorporate the prohibition against monsters. But this is just because that theory is not meant to capture in full the philosophical tenets of direct reference theory, as Kaplan himself points out.

²⁴Admittedly, even on the Kaplan picture there could be ways of generating monsters. This would require assuming the presence of special operators that can ‘reopen’ the slots filled in by indexicals: for example, an operator that, when fed the proposition expressed by ‘I am lost in a university library’, returns the property being lost in a university library. So the claim should be hedged as: on the Kaplan picture, the presence of monsters is a costly stipulation.
Then I zoom back out to the big picture, showing how I propose to modify the architecture of the system.

1.4.2 Re-anchoring indexicals

My basic proposal is very simple: indexicals are really just variables. In particular, they are variables with a special marker, which I represent as $[+c]$. For example, $I$ is syntactically represented as $x_{1[+c]}$, you as $x_{2[+c]}$, etc. Semantically, they work exactly on the model we use for other pronouns. A free pronoun like she denotes the object assigned to it by the assignment (for simplicity, I ignore gender):

$$[\text{she}]^g = [x_3]^g = g(5)$$

Similarly for $I$: its semantic value is just the object assigned to it by the assignment:

$$[I]^g = [x_{1[+c]}]^g = g(1[+c])$$

The challenge for this proposal, of course, is recapturing the data that originally motivated the standard account. After all, she can refer to different individuals within the same context, depending on speakers’ intentions or other contextual factors, while unembedded occurrences of $I$ are invariably anchored to the speaker. In some way, it is part of the meaning of $I$ that it refers to the speaker.

I suggest that the connection between indexicals and context is established not at the compositional stage, but rather at the postsemantic stage. In short, indexicals are anchored to elements of the actual context by the same mechanisms that fix the initial values of index parameters.

Let me get into some detail. Formally, the initialization of index parameters happens via a definition of truth at a context. In the Kaplan framework, this definition says that a sentence is true at a context just in case the semantic value it expresses at that context is true at the index coordinates of the context. For example, taking index coordinates to be worlds and times:

$$\varphi \text{ is true at } c \iff \text{the semantic value of } \varphi \text{ at } c \text{ is true at the world and the time of } c$$

The important observation is that this way of fixing index parameters captures facts that are specifically linguistic. To see this, consider again

(1) Juventus was the best team in Europe.

\footnote{See Kaplan (1989), p. 522 for an informal definition and p. 547 for a formal one.}
(1) conveys, as part of its meaning, that Juventus was the best team in Europe at some time before the utterance time. It has no reading on which it says that Juventus is the best team in Europe at (say) some time which precedes some time that the speaker intends to pick out.²⁶ This fact is captured by letting the initial coordinate of the index be set systematically at the time of the context itself, with no interferences from factors like salience or speakers’ intentions.

I claim that the anchoring of indexicals exploits an analogous mechanism. We let the values of indexicals be set to features of the context itself, like the speaker, the hearer, or the time of utterance, via the definition of truth at a context. There are several ways to do this. The simplest one adopts a minimal variant of Kaplan’s definition of truth at a context, formulated as follows:

φ is true at c iff the semantic value of φ is true at the world, the time, and the assignment of c

(Notice that this definition, unlike the previous one, uses a notion of semantic value that is not relativized to a context: more on this shortly.) In addition, we specify constraints on what can count as the assignment of the context. This specification will simply list what objects we assign to variables with special indices:

For any context c, an assignment g is the assignment of c only if

\[ g(1[+c]) = \text{the speaker of } c \]
\[ g(2[+c]) = \text{the addressee of } c \]
\[ \ldots \]
\[ g(5[+c]) = \text{the time of } c \]
\[ \ldots \]

Notice that, on this setup, the lexical entries of indexicals do not encode all the information that corresponds to their lexical meanings. Lexical entries capture only the compositional component meaning, and indexicals work compositionally as simple variables. This is not an inevitable feature of the monstrous picture. One alternative is to switch to a picture on which the meaning of indexicals involves two independent components. One gets used in the compositional computation in the usual way. The other remains inert at the compositional stage, and is then consulted by the definition of truth at a

²⁶Of course, setting aside non-standard uses like fictional statements. It’s interesting to notice that these cases are paralleled by non-standard uses of indexicals on which the latter do fail to pick out elements of the actual context, like (again) pretense or direct speech reports.
context to determine the assignment of the context. On this picture, the lexical entry for \( I \) would be split into two separate parts:

\[
[I]^{g} = [x_{I[+c]}]^{g} = g(1[+c]); \{g_{c}(1[+c])\} \text{is the speaker of} \ c
\]

The material in curly brackets is ignored by the compositional computation and is only consulted at the postsemantic level. So far as I can see, this kind of approach is substantially analogous to the one above for present purposes. So I won't take up the task of developing it formally here.$^{27}$

I want to stress one feature of the postsemantics strategy I have pursued. Despite requiring a detour through an extra definition, this strategy exploits mechanisms that are already present in our semantic machinery. I am not introducing new conceptual resources, but rather extending tools that we already have to the case of indexicals. This extension seems appropriate, as on the new picture index parameters and indexicals behave in exactly the same way: both start out anchored to the context and are shifted in the presence of appropriate operators.

### 1.4.3 Semantics without the context

How does the new treatment of indexicality modify the architecture of the semantics? The overall effect is an interesting simplification. The interaction between meaning and context happens at a single point, namely at the postsemantic level. No input from context is necessary to run the compositional semantics. So the context parameter becomes superfluous, at least for the compositional part of the theory. We can just run compositional semantics with the index and the assignment. The whole theory 'makes contact' with context only at the postsemantic level. This is a diagram representing the new picture:

---

$^{27}$For a fully developed semantic framework that implements a related idea, see Pott's (2005) extensive study of conventional implicature. As a side observation: notice that, if we went for this option, we would manage to reconstruct something analogous to Kaplan's character into the semantics of indexicals. The main difference would be that, on the new picture, character is only used after compositional interactions.
I’m not claiming that this is the only setup that can accommodate a monstrous semantics. On the contrary, the proliferation of alternatives to the Kaplan system suggests that there might be alternative ways of setting up the general framework that still yield the same predictions about truth-conditions. What I do claim, however, is that this setup is a straightforward and natural way of capturing the functioning of monstrous semantics. Similarly, Kaplan’s setup was a straightforward and natural way of capturing the workings of a non-monstrous semantics.

But why is this setup straightforward and natural? Because it puts on the same footing parameters that, on the new account, work fundamentally in the same way. The lesson of sections 1.2 and 1.3 is that the functioning of (say) the speaker and the addressee parameters is fully analogous to the functioning of index parameters. To repeat: all these parameters start out anchored to the coordinates of the context and are then shifted by appropriate operators. This setup accommodates the analogy, as it lets all of them be fixed by the postsemantic part of the theory. Formally, we might still be able to revert to a two-tier theory like Kaplan’s. But, so far as I can see, this would just be drawing distinctions where there are no significant differences.

The shift of framework is a significant point in itself. But it might also have an impact on questions connecting semantic theory with a general theory of cognition, in particular questions about the border between linguistic competence and general knowledge. Obviously these questions fall beyond the scope of this paper, but let me briefly hint at how the new semantics might be relevant for them.

Consider the question: what kind of knowledge is involved in speaking a language? It is broadly agreed that language use (construed in a broad sense, to include all forms of communicative behavior based on language) involves two different kinds of knowledge: on the one hand, linguistic knowledge, which constitutes a specific module of the mind; on the other, general world knowledge, which includes information concerning the context of speech. Syntactic knowledge is knowledge of the former kind; pragmatic knowledge is knowledge of the latter kind. Semantics is a disputed terrain. But Kaplan’s picture lends support to one claim: knowledge of the context of speech does mix with linguistic competence to generate meanings for sentences. This just because of indexicals. On the orthodox picture indexicals ‘grab’ their referents before compositional semantics. Hence we have to use contextual knowledge to settle the referents of indexicals before we go through a compositional computation. On the new picture, this is not the case. Indexicals are assigned referents at an independent stage of a theory of meaning; so compositional semantics can be run without looking at the context. This suggests that the new view of indexicality might lend new support to a modular view, on which compositional semantics has a high degree of autonomy from non-linguistic knowledge.

Of course, this is just a suggestive possibility raised by the new picture. A full as-
1.5 Extensions

Section 1.4 completes my account. Now I turn to investigating a few natural extensions.

1.5.1 Names

Consider the following tweak of the amnesiac scenario:²⁸

You and I are watching on a screen a man who's finding his way around a large university library. We decide to dub him 'Herman', and start using this name to talk about him while watching him. Then we are told about the coinflip scenario, and we are informed that the man on screen is the amnesiac who has survived. You say:

(21) If the coin landed tails, Herman is in Main Library, Stanford.
(22) If the coin landed heads, Herman is in Widener Library, Harvard.

Unsurprisingly, epistemic shift reappears. (21) and (22) are both true. But if names invariably refer to the object they pick out in actual circumstances, standard accounts miss this prediction. Again, the reason is that there is no single individual in the speaker’s epistemic alternatives that is in Main in case of tails and in Widener in case of heads. And again, an intuitive analysis suggests that Herman behaves in a shifty way: it picks out whatever individual we named Herman in the circumstances individuated by the antecedent.

It’s natural to think that we can account for this phenomenon by treating names as variables and letting them be bound by informational modals. This would not be a novelty: recently, Cumming (2008) has suggested just that names work as variables and that attitude verbs are able to bind them via shift of the assignment.

Two qualifications are in order. First, the counterpart relations associated to names do not, in general, work in the same way as those associated to indexicals. In this respect, Herman is a rather atypical example. Herman is introduced via an explicit stipulation to denote the man you and I are watching on screen (its functioning is similar to that of descriptive names; cf. Evans (1985a)). Hence, given our background knowledge, the counterpart relation that is most naturally associated to it singles out Lingens in tails-worlds and Lauben in heads-worlds. But now consider a more mundane name,

²⁸Thanks to Dilip Ninan for suggesting this version of the example.
like Lingens. Lingens is at the center of a complex practice of referring to an individual, Lingens, within a certain linguistic community. (It’s not important for present purposes whether this practice involves causal chains or other phenomena.) As a result, one counterpart relation that is made salient by the use of the name is the one picking out the individual named Lingens within the relevant community. This difference between the counterpart relations made salient by the use of Herman and Lingens explains the difference between (23) and (24):

(23) Herman might be Lauben.
(24) Lingens might be Lauben.

In the usual scenario, the speakers know that the individual they name Herman might turn out to be Lingens or Lauben. But it is common knowledge that Lingens and Lauben denote different people. This explains why (23) is intuitively true and (24) intuitively false.

Let me emphasize, though, that the use of one counterpart relation or the other is ultimately a matter of contextual salience. Different counterpart relations can be employed for different occurrences of the same name, depending on context. To see this, consider

(25) Jason believes that Lingens is not Lingens.

Jason has convinced himself that the lost amnesiac is Lauben. But he’s wrong. You know that all along he’s been tracking Lingens, and Lauben is now dead. In this scenario, you can truly utter (25). But of course, the two occurrences of Lingens in (25) must be paired with different counterpart relations, lest you want to attribute Jason a contradictory belief.

The second qualification concerns an important difference between Cumming’s and my account of assignment shift. Cumming claims that names occurring under informational modals may or may not be bound. By contrast, my account mandates binding of all referential expressions in the scope of these modals. This difference is significant both formally and conceptually: among other things, allowing for non-bound occurrences introduces a distinction between de re and de dicto readings of names in the scope of informational modals. On Cumming’s theory, the de re/de dicto distinction is exemplified just by the two occurrences of Lingens in (25): the first is de re (hence it picks out the actual Lingens), the second one de dicto (hence it picks out Lingens’s counterpart in Jason’s belief worlds).

A full argument for my position would require extensive discussion, but I can mention two points in its favor. First, the data of section 1.2 encourage the obligatory binding hypothesis. Shifty conditionals have only one reading: the shifty one. For example,
1.5. EXTENSIONS

consider again

(21) If the coin landed tails, Herman is in Main Library, Stanford.
(22) If the coin landed heads, Herman is in Widener Library, Harvard.

There is simply no reading on which one of the two is false (while still having epistemic flavor). This is fully expected on a view like mine, but not on a theory that allows for a de re/de dicto ambiguity.

The second point is that, so far as I can see, there is no good way to generalize Cumming’s theory to the case of indexicals. We can naturally make sense of a de re/de dicto distinction for occurrences of names. Take your favorite metasemantic account of the reference of names: i.e., an account of the factors that make it the case that a name refers to a certain object. This account, let’s say, claims that a name \( n \) refers to object \( o \) in virtue of relation \( R \) holding between them. (For example, \( R \) might consist in the existence of the causal chain connecting the two.) Then de re occurrences of \( n \), which are free variables, are assigned the individual that bears \( R \) to \( o \) in the actual world. E.g., a free occurrence of \( \text{Lingens} \)—like the first occurrence in (25)—refers to the individual linked to \( \text{Lingens} \) by \( R \) in the actual world. De dicto occurrences, which are bound variables, range over the individuals that bear relation \( R \) to the name in the worlds within the relevant information state. A bound occurrence of \( \text{Lingens} \)—like the second occurrence in (25)—refers, in each of the relevant worlds, to the individual linked to \( \text{Lingens} \) by \( R \) in that world. But it’s not clear at all how to give an analogous account for the case of indexicals. Neither features of the lexical meaning of indexicals nor features of their metasemantics seem to produce a credible story that covers both cases.²⁹ I have shown how, by contrast, an account based on obligatory binding can accommodate uniformly both names and indexicals.

1.5.2 Gendered pronouns

Another natural extension of the theory concerns deictic uses of pronouns like \( \text{she} \) and \( \text{he} \), in two respects. First, there are cases where \( \text{she} \) and \( \text{he} \) seem to pick out different referents depending on the antecedent of an epistemic conditional, as it happens for \( I \) in (5) and (6). The point should be familiar by now, and I leave it as an exercise to the reader to construct the relevant examples. Second, epistemic shift seems to affect

²⁹So far as I can see, the only natural suggestion would be to appeal to Kaplanian characters: for example, to let \( I \) denote whatever individual is speaking in the worlds in the relevant information state. This is essentially what happens in monstrous theories used to account for languages other than English: see e.g. Schlenker (1999), Schlenker (2003), Anand & Nevins (2004). These theories do in fact allow for both free and bound readings of indexicals under the monstrous verbs. But these theories are empirically inadequate for English.
also the functioning of gender in these pronouns. This point is significant because, as I anticipated in section 1.2, it brings to light a direct connection between epistemic conditionals and belief reports.

Yanovich (2010) has recently pointed out that the meanings of she and he involve an interesting indexical component. she and he invariably specify the actual gender of the individual they pick out (at least when that individual exists in actuality). To see this, consider these examples:

(26) If John were a woman, he would be much happier.
(27) [pointing to a woman] Jason believes that she is a man.
(28) Tom wishes he were a woman.

Take (26): the gender of the pronoun tracks the actual gender of the individual. The masculine he is used to pick out John, despite the fact that in the counterfactual worlds under consideration John is female. Similarly, mutatis mutandis, for (27) and (28). This suggests that she and he involve an indexical component: something in their meaning calls into question facts obtaining in the actual world. It’s useful to state explicitly how this is captured in lexical meanings. On standard views, she and he are analyzed as variables that have a presupposition specifying gender.³⁰ For example, she denotes the individual assigned to it by the assignment, and presupposes that that individual is female. Making room for the indexical element, the meaning of she can be specified as³¹

\[
\text{[she]}^{c,i,g} = [x_3]^{c,i,g} = g(5): \text{g(5) is female in the world of } c
\]

What matters for our purposes is that the pattern exhibited by (26)–(28) is disrupted, once more, in epistemic conditionals.³² Suppose that I’m talking about a child, Pat, whose gender I don’t know. I can say:

(29) If Pat is a girl, I’ll give her a toy bazooka.
(30) If Pat is a boy, I’ll give him a sewing kit.

³⁰The classical presuppositional account of gender is due to Cooper (1983). See the papers in Harbour et al. (2008) for recent literature on the topic.
³¹This, of course, is the lexical entry in orthodox frameworks that use a context parameter. In the framework I’m using, it can be reformulated by using a world variable \(w_\alpha\) that is invariably anchored to the context:

\[
\text{[she]}^{i,s} = [x_3]^{i,s} = g(5): \text{g(5) is female in } w_\alpha
\]

³²The data that follow are due to Yanovich, while the observation (in the next paragraph) that the phenomenon generalizes to first-person belief reports is mine.
Once more, we have a kind of epistemic shift. On standard accounts of indexicality, one of (29) and (30) will be infelicitous no matter what the context is, since its presupposition will not be satisfied. Yet the two conditionals are perfectly good sentences.

Here I won't focus on giving a full account of (29) and (30); it should be clear anyway that the monstrous route is a promising way to proceed. Rather, I want to point out how the phenomenon exhibited by (29) and (30) also infects some belief reports, in particular belief reports in the first person. Notice that, rather than (29) and (30), I could say

\[(31) \quad \text{I believe that either Pat is a girl and I'll give her a toy bazooka or Pat is a boy and I'll give him a sewing kit.}\]

(31) is a good sentence, even though the two pronouns in it cannot both be used appropriately if they are meant to specify the actual gender of the individual. So also in this case there is a kind of epistemic shift. Notice that the fact that (31) is in the first person, rather than the third, plays an important role. Suppose that the speaker, still ignorant about Pat's gender, is reporting Jason's beliefs rather than her own. According to my informants, (32) is better than (33) for this task:

\[(32) \quad \text{Jason believes that either Pat is a girl and he'll give her or him a toy bazooka, or Pat is a boy and he'll give her or him a sewing kit.}\]

\[(33) \quad ??\text{Jason believes that either Pat is a girl and he'll give her a toy bazooka, or Pat is a boy and he'll give him a sewing kit.}\]

The contrast between (32) and (33) is important. It shows that the distribution of feminine and masculine pronouns in sentences like (31) is semantically connected to the presence of believe (and cannot be explained, say, just by appealing to the properties of disjunction). It matters what worlds are quantified over and how they are related to the speaker's own beliefs: just a difference in the latter produces a difference in the acceptability of (31) and (33). So (31) exemplifies a genuine case of epistemic shift and vindicates the idea that we should pursue a unified shifty semantics for epistemic modals and attitude verbs.

1.5.3 Adverbs of quantification

Nunberg (1993) has pointed out that indexicals can have descriptive-sounding readings when occurring under adverbs of quantification. Suppose that the Pope utters

\[(34) \quad \text{I am usually Italian. (Recanati (2005), attributed to Nunberg)}\]
(34) is most naturally paraphrased as *the Pope is usually Italian*. Hence the truthconditional contribution of *I* seems analogous to that of a description. Similarly,

(35) Tomorrow is always the biggest party night of the year. (Nunberg (1993))

has a natural reading on which it says that, for every year, the day falling on a certain date is the biggest party night of that year. Again, *tomorrow* seems to somehow make a truth-conditional contribution analogous to that of a description.

A number of theorists have taken sentences like (34) and (35) as evidence for a kind of descriptivist semantics for indexicals. The most prominent account has been developed informally by Nunberg himself, and has been implemented compositionally (though only for the case of demonstratives) by Elbourne (2008). On Elbourne’s account indexicals are syntactically complex entities: crucially, they include a contextually specified descriptive component which takes an argument for situations. This allows them to interact compositionally with adverbs like *always* and *usually*, which, following a longstanding tradition in semantics (see for example Heim (1990) and von Fintel (1994)), Elbourne analyzes just as quantifiers over situations.

It’s natural to think that the semantics I’ve defended should somehow generalize to Nunberg cases. While there are no major technical obstacles, I’m reluctant to take this step at the present stage. The scale and the robustness of Nunberg-type phenomena are still ill-understood; so I’m not in the position to give a theory yielding reliable predictions. In the first place, while (34) and (35) are generally acceptable, they don’t sound as good and natural as the data about epistemic modalities that I’ve introduced in this paper. It’s not clear how this asymmetry should be accounted for. Moreover, the productivity of the phenomenon is quite limited; it’s easy to find examples where indexicals can’t have descriptive readings, even at the cost of making the whole utterance infelicitous. Suppose that, after having survived unscathed a fall from the third floor, I say:

(36) #I guess I was lucky. I usually die³³

The second clause in (36) is obviously infelicitous, even though its structure is analogous to (34) and it’s clear what the speaker intends to say. Before giving a theory, it would be important to establish generalizations about the availability of descriptive readings. This task would take me very far from my main argument and is better left to a different occasion.

Nevertheless, it’s worth pointing out that the general framework of sections 1.3 and 1.4 could easily be extended to Nunberg-type phenomena, if one wanted to do so. Following Elbourne, just assume that adverbs of quantification manipulate a modal pa-

³³Thanks to Irene Heim for the example.
rameter, be it a world or a situation. Then their functioning can be assimilated to that of informational modals: we let them shift the assignment and the modal parameter in a coordinated way. The outcome is that indexicals range over counterparts of their referent in the world of utterance, giving rise to descriptive truth-conditions. The resulting theory yields predictions that are basically analogous to Elbourne's theory (and hence the two theories share the overgeneration worries).

The monstrous framework I’ve presented is general enough to be applied beyond the epistemic domain. But specific constraints and specific aspects of the implementation (for example, the kind of counterpart relation in play under different modals) should be assessed on a case by case basis. Here I’m happy with providing (at best) a general blueprint for an account of indexical shift beyond the epistemic case: substantial work is needed to fill in the details.

1.5.4 Binding of fake indexicals

It’s useful to distinguish the phenomena I’ve been concerned with from some superficially similar issues. Since Heim (1994), it is acknowledged that there are seemingly bound readings of indexicals in certain quantified sentences. The classical example is:

(37) Only I did my homework.

(37) has a reading (the so-called sloppy reading) that can be roughly paraphrased as: nobody but me did her or his homework. But if my works as a genuine indexical, this reading cannot be derived compositionally. The only reading we get is the so-called strict one, on which (37) says that no other person than the speaker did the speaker’s homework. The standard solution (see Heim (2002) and von Stechow (2002) and (2003)) essentially consists in denying that the occurrence of my in (37) is a real indexical. Rather, it is a bound variable analogous in all relevant respects to a bound occurrence of he. Thus the logical form of (37) is:\footnote{\textsuperscript{34}}

\begin{equation}
(38) [\text{Only I}] \lambda t_i. t_i \text{ did } x_i \text{’s homework.}
\end{equation}

The bound variable $x_i$ is phonologically realized as a first-person pronoun because of morphological rules of agreement. Essentially, the idea is that the person, gender, and number of a bound pronoun must match the person, gender, and number of the phrase that binds it.\footnote{These rules can take the form of feature transmission or feature deletion (see Heim (2002) and von Stechow (2002) and (2003)). The difference is irrelevant from a semantic point of view: in both cases, the logical form involves pronouns with features that are uninterpreted semanti-"}
dexicals; the occurrence of something that looks like an indexical in the scope of the quantified phrase *Only I* is a mere morphological accident.

If this account is correct, the phenomenon displayed by (37) is completely distinct from the one I’ve been analyzing in this paper. I never questioned that the problematic bound pronouns I discussed are genuine indexicals. In fact, my account crucially exploits components of the meaning of *I* to determine the compositional contribution of its bound occurrences. (As I explain in the appendix, when *I* occurs bound, the binding modal checks that the counterpart relation associated to it is the one appropriate for the actual speaker. This process crucially relies on the fact that *I* is anchored to the speaker via the postsemantics.) By contrast, from a semantic point of view, *my* in (37) is simply not an indexical. So (37) is not a case of indexical binding at all.

One might worry that treating the two phenomena in altogether different ways misses a generalization. But this concern would be misplaced: there is good evidence for holding distinct accounts. First, there are obvious differences in the intuitive truth-conditions associated to the data. The natural paraphrase of

(5) If the coin landed tails, I am in Main Library, Stanford.

involves a descriptive phrase like *the person speaking* in the place of *I*. But, as I just pointed out, in the natural paraphrase of (37) *my* is substituted by a pronoun bound by the quantificational expression *only I*. It seems unlikely that a systematic semantic story could produce analogous truth-conditions for the two cases. Moreover, the shifty phenomena I have investigated seem to appear under a specific class of modal verbs. But the kind of phenomenon displayed by (37) doesn’t depend on the presence of a specific operator, not even *only*: it can reappear in virtually any situation in which a pronoun is bound. It’s hard to get telling data for the case of person features without using *only*. But the case can be made with number features:

(39) Few men brought their children. (Heim (2002))

*their* in (39) is a plural pronoun, but it can get a singular reading: the sentence can mean *The number of the x such that x brought his children is small*. It’s natural to think that (39) and (37) are instances of the same phenomenon; and indeed, in both cases the Heim-von Stechow analysis gets the right predictions.

In short, a theory of indexical shift should not be stretched to cover cases like (37) and (39). Both the intuitive data and the empirical distribution of the phenomenon look very different in the two cases. At least at the present stage, the prospects for a unified and explanatory theory are dim.
1.6 Outstanding issues

Let me mention a few questions that, for reasons of space, I can’t address in this paper.

**Kaplanian alternatives.** Throughout the paper, I have concentrated on giving my positive account of epistemic shift. But are there any prospects for giving an account that is more in line with the orthodoxy? Since Kaplan’s seminal work on *de re* attitude reports (1968), the semantics literature has tried to account for singular terms in the scope of attitude verbs with the tools he provided. However, all approaches that descend from Kaplan have to rely on important technical stipulations (as is often pointed out by their own proponents).³⁶ I suggest that monstrous semantics is just the way to solve these issues: indeed, it is just the account that Kaplan himself should have given (and in fact shares with it some important features, like the use of epistemic counterparts). I set up a comparison between an up-to-date version of Kaplan’s semantics for attitude reports and my account in Santorio (2010).

**Metaphysical modals.** I’ve focused on informational modals. But does the new account have an impact on metaphysical modals? One option would be to say ‘no’ and preserve traditional Kripkean semantics for metaphysical modality. But this seems undesirable. It would come at the cost of ignoring strong evidence that metaphysical and epistemic modals share their semantic architecture. After all, the very same words (*must, might, and could* are good examples) are often used to express both metaphysical and epistemic modalities. A second, more interesting option is to generalize monstrous semantics to all modals. On this view, all modals would shift the assignment function and let indexicals in their scope range over sets of counterparts of their ordinary referents. The difference in truth-conditions would be explained via a difference in the counterpart relations in play. Epistemic modals would use counterparts by acquaintance, metaphysical modals would use metaphysical counterparts. This view seems worth exploring.

**Other monsters.** A wave of recent work in semantics (among others Schlenker (2003) and (2010), Anand & Nevins (2004) and Anand (2006)) has shown evidence that languages other than English, for example Amharic and Zazaki, also contain indexical-shifting expressions (though see also von Stechow (2002) and (2003) for an attempt at an alternative, non-monstrous analysis of the data). These monsters are formally quite different from mine and are based on classical work in two-dimensional modal logic. Interestingly, though, the monsters of Amharic and Zazaki are again a kind of informational modal (mostly speech report verbs). I believe there is hope for a general characterization of monstrosity in natural language, on which Schlenker and Anand’s

³⁶For statements of this sort, see for example Anand (2006), page 26, and Schlenker (2010), page 18.
monsters turn out to be a special case of my monsters.

1.7 Conclusion

I have argued that, pace Kaplan, English turns out to be replete with monsters. I, you, now, and the like are systematically bound in the scope of modals like might and believe and epistemic conditionals. This teaches us interesting facts about the semantics of indexicals and modals, and suggests a shift in the architecture of our semantic theories. One lesson is that many (though not all) of the ideas behind direct reference semantics can coexist with a semantic notion of cognitive significance. Indexicals can contribute to truth-conditions something that is richer than their referents, even though they belong to an altogether different semantic kind from descriptions. A second lesson is that the architecture of our semantic theories should be redesigned: semantics makes contact with context not at the compositional level, but only at the postsemantic level.

I have started my paper with a new empirical puzzle and I have offered my account as a way of solving it. I don’t have an argument that monsters are the ultimate or the only way of solving the puzzle: arguments of this sort are hard to come by in empirically driven philosophy of language. But I hope that the interest of the proposal will not depend on whether it provides the ultimate or only account of epistemic shift. The mere fact that there is a monstrous alternative to standard views of indexicality and reference has, I think, an interest of its own. Contrary to common wisdom, there is an alternative way of doing semantics for indexicality. The main goal of this paper has been to convince you that this alternative is worth exploring.
Appendix: semantics

Setup. The basic functioning of the semantics, including the composition rules, is that given in Heim & Kratzer (1998). In line with classical accounts of the semantics for modality, I assume an intensional system, in which clauses are assigned truth-values relative to an index of evaluation. In line with standard semantics for de se attitudes, I assume that indices include at least a world, a time, and an individual parameter. (This allows attitude verbs to manipulate centered worlds.) For simplicity, I represent the index just with the schematic letter ‘i’. Because of the reasons pointed out in section 1.4, there is no context parameter. Thus clauses are assigned a truth-value relative to an index and an assignment:

\[ [\text{Fidel loves Ralph}]^{i,g} = 1 \text{ iff Fidel loves Ralph at } i, \text{ relative to assignment } g \]

Indexicals. As I state in the text, indexicals are variables with a special diacritic \([+c]\). Their semantics is analogous to the semantics of ordinary variables of the corresponding types. For example:

\[ \left[ 1 \right]^{i,g} = \left[ x_{1[+c]} \right]^{i,g} = g(1[+c]) \]

Composition rules. Ordinary clauses denote truth-values, while modals require a clausal argument of higher type. This generates systematic type mismatch. The mismatch is fixed via a new composition rule which enforces lambda-abstraction on the assignment function:

Monstrous Functional Application (MFA)

If \( \alpha \) is a branching node and \( \{ \beta, \gamma \} \) the set of its daughters, then for any index \( i \) and any assignment \( g \), if \([\beta]^{i,g}\) is a function whose domain contains \( \lambda g'. \lambda i'. [y]^{i',g'} \), then \([\alpha]^{i,g} = [\beta]^{i,g}(\lambda g'. \lambda i'. [y]^{i',g'}). \)

Modals. As is standard in semantics for the de se (see Chierchia (1989) and Anand (2006)), the modal base of epistemic modals and attitude verbs consists of a set of centered worlds. Differently from standard semantics for the de se, modals quantify over pairs of assignments and centered worlds. Before giving a lexical entry, it’s useful to define the notion of an assignment-centered world pair being compatible with a modal base and a sequence of counterpart functions.

(Notice two assumptions. (a) An assignment is a function from numerical indices to possible individuals. (b) Counterpart functions are matched to variables via their indices, so that for each variable there is a counterpart function with the same index.)
An assignment-centered world pair \( (g_k, i_k) \) is compatible with a modal base \( M \) and a sequence of counterpart functions \( (f_{n_1}, f_{n_2}, \ldots) \) iff

- \( i_k \in M \);
- for all indices \( n \), \( g_k(n) = f_n(i_k) \)

In short, we generate assignment-centered worlds pairs by 'plugging in' the centered worlds in the modal base into the sequence of counterpart functions.

Finally, here is the lexical entry for epistemic must (using ‘\( \alpha \)’ as a type for assignments and ‘\( \chi \)’ as a variable over functions from assignments to functions from indices to truth-values):

\[
\lambda g \cdot \lambda g' \cdot \lambda \psi \cdot \lambda \chi \cdot (f_1[+c]) (i) = (\text{if defined}) g(x_1[+c]), \\
(f_2[+c]) (i) = (\text{if defined}) g(x_2[+c]), \\
\ldots \\
\text{s.t. } \chi(g') (i') = 1, \psi(g') (i') = 1
\]

In essence, must does two things: first, it 'checks' that the counterpart functions provided by context are the right ones. For example, it checks that \( f_1 \), if defined, does in fact specify the way that the subject is acquainted with the speaker. Second, it quantifies over the relevant assignment-centered worlds pairs.

**Sample derivation.** Finally, I give a compositional derivation of 

(5) If the coin landed tails, I am in Main Library, Stanford.

(The derivation of (6) is analogous, *mutatis mutandis.*) I assume that the logical form of (5) is the following:

\[
\text{must} \ [\text{the coin landed tails}] \ [x_1[+c] \text{ is in Main Library, Stanford}] \ (M_{(i,(i,t))})
\]

Three qualifications. First, for simplicity, I treat the name Main Library, Stanford as a constant rather than a variable. Second, still for simplicity I ignore tense. Third, I use ‘s’ as a shorthand for \( (f_1, f_2, \ldots) \), which is a sequence of variables ranging over counterpart relations.

Here is the derivation:
1.7. CONCLUSION

\[ \text{MUST} \ [\text{the coin landed tails}] \ [x_{1[+c]} \text{ is in Main Library, Stanford}] \ (s) \ (M) \] \(i\varepsilon\) is true iff
(by Functional Application)

\[ \text{MUST} \ [\text{the coin landed tails}] \ [x_{1[+c]} \text{ is in Main Library, Stanford}] \] \(i\varepsilon\)(\(g(s)(g(M))\) iff
(by Monstrous Functional Application)

\[ \text{MUST} \ (\lambda g'. \ \lambda i'. [\text{the coin landed tails}] \ [x_{1[+c]} \text{ is in Main Library, Stanford}] \] \(i\varepsilon\)(\(g(s)(g(M))\) iff (by computing the denotations of the clauses)

\[ \text{MUST} \ (\lambda g'. \ \lambda i'. \text{the coin landed tails at } i') \ (\lambda g'. \ \lambda i'. g'(1[+c]) \text{ is in Main Library, Stanford at } i') \ (g(s)) \ (g(M)) \text{ iff (by the meaning of MUST, and keeping track only of the conditions imposed on the first-person index '1[+c]'})\]

\[ [\lambda x. \ \lambda y. \ \lambda (f_1, f_2, \ldots). \lambda M. \ (f_{1[+c]})(i) = \text{(if defined) } g(x_{1[+c]}), \ \text{and } \forall (g', i') \text{ compatible with } M(i) \text{ and } (f_1, f_2, \ldots) \ \text{and s.t. } \chi'(g')(i') = 1, \ \psi(g')(i') = 1] \ (\lambda g'. \ \lambda i'. \ \text{the coin landed tails at } i') \ (\lambda g'. \ \lambda i'. g'(1[+c]) \text{ is in Main Library, Stanford at } i') \ (g(s)) \ (g(M)) \text{ iff (by the definition of } \lambda\text{-notation)}\]

\[ [\lambda (f_1, f_2, \ldots). \lambda M. \ (f_{1[+c]})(i) = \text{(if defined) } g(x_{1[+c]}), \ \text{and } \forall (g', i') \text{ compatible with } M(i) \text{ and } (f_1, f_2, \ldots) \ \text{and s.t. the coin landed tails at } i', g'(1[+c]) \text{ is in Main Library, Stanford at } i') \ (g(s)) \ (g(M)) \text{ iff (by the definition of } \lambda\text{-notation again)}\]

\[ (g(f_{1[+c]}))(i) = \text{(if defined) } g(x_{1[+c]}), \ \text{and } \forall (g', i') \text{ compatible with } g(M(i)) \text{ and } (g(s)) \text{ such that the coin landed tails at } i', g'(1[+c]) \text{ is in Main Library, Stanford at } i' \text{ iff (fixing the modal base and simplifying)}\]

\[ (g(f_{1[+c]}))(i) = \text{the speaker and } \forall (g', i') \text{ compatible with what the speaker knows at } i \text{ and the sequence of counterpart relations } (g(s)) \text{ such that the coin landed tails at } i', g'(1[+c]) \text{ is in Main Library, Stanford at } i'.\]
2. Denoting without Description

2.1 Introduction

According to an influential line of thought, definite descriptions like *the richest man on Wall Street* belong to a different semantic kind from pronouns or proper names like *she* or *Bernard Madoff*. *She* and *Bernard Madoff* are syntactically simple expressions referring to individuals. Definite descriptions share their syntactic structure with quantified phrases—like, say, *every banker in Manhattan*—and are not referential expressions; their semantic value is not an individual, but rather a function of some kind. There is even stronger consensus that indefinite descriptions like *a stockbroker* should be grouped with quantifiers, rather than referential expressions. The indefinite article *a* is traditionally treated as equivalent to the existential quantifier in first-order logic: accordingly, sentences involving indefinite descriptions are given the truth-conditions of existentially quantified statements.

The quantifier analysis of descriptions dovetails with traditional views of propositional attitude reports. At least since Quine (1956), philosophers and linguists distinguish two readings of sentences like (1):

(1) Stacey intends to marry a stockbroker.

The first—the *de dicto* reading—is true if Stacey intends to marry some stockbroker or other, with no need of her having settled on a particular contender. The second—the *de re* reading—is true if Stacey has a marriage scheme concerning a particular individual who happens to be a stockbroker, whether or not Stacey is aware of her or his profession. It’s easy to see how the *de re/de dicto* ambiguity is explained by the quantifier analysis. The two readings are obtained by evaluating the descriptive material of *a stockbroker* with respect to Stacey’s attitudes, or with respect to the actual world. On the *de dicto* reading, Stacey’s intention is to marry some individual who is a stockbroker according to what she believes; on the *de re* reading, Stacey’s intention is directed to some person
who is a stockbroker in the actual world.

In this paper I argue that, contrary to the quantifier analysis, both definite and indefinite descriptions can have referential readings. Crucially, this referential link is wholly independent of the descriptive material (in essence, the predicate) that is part of the description. It may be that the descriptive material is evaluated with respect to a subject’s attitudes, and the description still refers to an actual world individual. Hence there are two unrelated dimensions to the meaning of descriptions: on the one hand, the descriptive condition specified by the predicate; on the other, the (possible) referential link to an object.

Of course, this view has repercussions for our overall semantics of attitudes. Indeed, my argument crucially exploits \textit{de re/de dicto} phenomena. My central claim is that attitude reports have a hybrid and so far unnoticed reading—what I call ‘referential opaque’ reading. Take again (1):

(1) \hspace{1cm} \text{Stacey intends to marry a stockbroker.}

On the referential opaque reading, the descriptive material \textit{stockbroker} may be evaluated with respect to Stacey’s attitudes, and at the same time the whole indefinite phrase \textit{a stockbroker} may display a semantic connection to an actual world individual. From here, the argument is simple: since this semantic connection doesn’t pass through the descriptive material, we must recognize that there is a separate mechanism that is responsible for it.

My main aim in the paper is establishing a factual claim and not giving a new analysis of descriptions. But I also survey two lines of thought for an account. On the first, descriptions are referential expressions exactly like pronouns or names: in particular, they are analyzed on the model of variables. On the second, descriptions are still quantificational expressions, but they involve a referential element. More specifically, they involve a covert free variable which works as a referential pronoun, picking out the intended referent. Both lines of thought, I argue, need substantial development to be turned into proper analyses.

Here is the structure of the paper. In section 3.2, I briefly present the quantifier view of descriptions and summarize the debate on \textit{de re} and \textit{de dicto} reports. In section 2.3, I give two arguments for the existence of referential opaque readings; both rely on observations about anaphora. Section 2.4 sketches the two potential routes to accounting for the new data; I close, in section 2.5, by pointing out how my arguments connect to two other issues in philosophy and semantics.

Before proceeding, a terminological note: in line with the linguistics literature, I will use ‘definite phrase’ and ‘indefinite phrase’ (or more simply, ‘definite’ and ‘indefinite’) to pick out, respectively definite and indefinite descriptions; I will also use the blanket label
2.2 Setup: descriptions and de re/de dicto phenomena

I start from a view on which both definite and indefinite phrases are generalized quantifiers (Barwise & Cooper (1981)). According to this view, the and a share their syntax and their semantic types with a large class of expressions, including, for example, every, some, most, many. All these expressions are functions taking as argument two predicates and returning a proposition; moreover, all of them are quantificational in nature, in the sense that they express quantitative relations obtaining between the two sets picked out by the predicates. In essence, this is just a linguistically enhanced theory of Russell's (1905) celebrated views on descriptions. The truth conditions of sentences containing both definite and indefinite descriptions, like

(2) The richest man on Wall street asked Stacey out.
(3) A stockbroker proposed to Stacey.

are, in both cases, existential (“There is a unique richest man on Wall Street, and he asked Stacey out” and “There is a stockbroker who proposed to Stacey”); the only difference consists in the uniqueness requirement associated to definite phrases.¹

What matters, for our purposes, is how the quantificational view of descriptions interact with the semantics of attitude reports. I first present the traditional intensional frameworks in which de re/de dicto phenomena were first analyzed. I then move to more contemporary frameworks.

Intensional semantic frameworks are natural descendants of intensional logics, like modal logic or temporal logic. All formal semantic frameworks make use of an interpretation function (denoted by the brackets ‘[ ]’), i.e. a function pairing expressions of the language with their denotations. The peculiarity of intensional frameworks (see e.g. Kaplan (1989)) is that this function is relativized to a number of index parameters: among these, a possible world. As a result, the assignment of truth-values to clauses is relativized to a world. For example, the denotation assigned to (4) is the one in (5):

(4) Jason is a stockbroker.
(5) $[[\text{Jason is a stockbroker}]]^w = \text{true iff Jason is a stockbroker in } w$

¹ Notice that the is a controversial case: several linguists who buy into generalized quantifier theory still want to treat the as an operator which, combined with a predicate, refers to an individual, in the style of Frege (1892/1997). (For example, see the treatment of definites in Heim & Kratzer (1998).) All my main points apply to a view of this sort; I stick to a uniform quantificational treatment just for expository convenience.
As a default, the value of the world parameter is just set to the world of utterance. But this value can be changed. On standard accounts, attitude verbs like believe or want work just by shifting the world coordinate of their complement clause.² The world that is relevant for evaluating this clause is not the world of utterance, but rather the set of worlds compatible with the relevant attitudes of the subject. For example, the denotation of (6) is (7):

(6) Stacey believes that Jason is a stockbroker.

(7) $[[\text{Stacey believes that } \lambda] \tilde{w} \tilde{w} = \text{true for all worlds } w' \text{ compatible with Stacey's beliefs in } w, \lambda] \tilde{w} = \text{true}$

As a result of this shift, all the linguistic material that occurs in the scope of an attitude verb is evaluated at the shifted world.

In this setup, the de re/de dicto distinction is easily explained via differences in the relative syntactic scope of descriptive phrases and attitude verbs. Consider again:

(1) Stacey intends to marry a stockbroker.

(1) might be syntactically disambiguated in two relevant ways. On the one hand, the indefinite phrase a stockbroker may take narrow scope with respect to the attitude verb:

Stacey intends [to marry [a stockbroker]]

On the other, and vice versa, it may take wide scope with respect to the verb:³

[a stockbroker] $\lambda_1$[Stacey intends [to marry $t_1$]]

These two syntactic possibilities give rise, respectively, to the classical de dicto and de re readings. If the indefinite has narrow scope with respect to the verb, then it is evaluated at the shifted world parameter introduced by the latter. This gives rise to the classical de dicto reading. Conversely, if the indefinite has wide scope, it is evaluated at the world of the context; then we get the de re reading.

To summarize, these are the two readings expected on the intensional account:

²The idea of treating attitude verbs as modal quantifiers is due to Hintikka; see his (1962) and (1969).

³In line with Heim & Kratzer (1998), I’m assuming background syntax and semantics on which determiner phrases like a stockbroker undergo syntactic movement leaving behind a trace coindexed with a lambda-abstractor. Also, I’m ignoring irrelevant complications in writing down the LFs. In particular, I’m ignoring (a) the fact that a stockbroker undergoes movement within the clause also in the first LF, by quantifier raising; and (b) the fact that the infinitival clause in (1) involves the covert subject PRO (see Chierchia (1989) for discussion of the latter point).
(a) *Classical de dicto.* Stacey intends to marry some stockbroker or other.

(b) *Classical de re.* There is a particular individual who is a stockbroker, and Stacey intends to marry that individual.

The problem for the intensional paradigm is that *de re* and *de dicto* phenomena in natural language are more complex than what is captured by the two classical readings. The easiest way to see this is that (1) has a further reading, as was pointed out by Fodor (1970). Consider the following scenario:

> Stacey, a resident of Wall Street, is fascinated by a group of men who gather regularly in a local bar and that she takes to be lawyers. She forms the plan of marrying one of them, though she doesn't settle which one. Unbeknownst to her, all the men in the group are actually stockbrokers.

In this scenario, (1) has a true reading. But this reading is not the classical *de dicto* or *de re* one: Stacey doesn't intend to marry some stockbroker or other and she doesn't have marriage plans concerning a particular individual. Rather, the true reading of (1) has, roughly, the following truth-conditions:

(c) *Narrow-scope de re:* Stacey intends to marry someone within a certain group, though not a particular person. Those people all happen to be stockbrokers.

It's easy to see that this third reading cannot be accounted for by the intensional theory. Suppose that we give the indefinite narrow scope with respect to the attitude verb. Then the indefinite must be evaluated at the shifted world-index. Hence we end up ascribing to Stacey the intention to marry someone she takes to be a stockbroker, which is wrong. And now, suppose that we give the indefinite wide scope. In this case, the intensional framework forces us to ascribe to Stacey the intention to marry a particular individual; but she doesn't have this intention.

The existence of the third reading, together with other puzzles, has produced a paradigm shift towards so-called extensional frameworks.⁴ In extensional frameworks,

---

⁴For a useful summary of the problems for the intensional theory, see chapter 1 Keshet (2008). Notice: while the extensional framework for representing modal talk has achieved the status of orthodoxy, it is not universally accepted. On the contrary, there is a live debate about alternatives in semantics. As has been recently pointed out, the variable-based theory has significant shortcomings, mostly due to overgeneration; see e.g. Percus (2000), Keshet (2008), Romoli & Sudo (2008). It’s clear that some amendments are needed, and some tentatively advocate a return to intensional systems (see for example Keshet (2008), chapter 5). Some theorists in the dynamic semantics camp have also produced alternatives: see e.g. Maier (2011). In any case, it’s not vital for my purposes that the variable-based theory is correct. Any existing account shares with it the basic assumption I’m attacking.
the interpretation function is not relativized to worlds. Rather, modal talk is taken to
be represented directly in the object language. Natural languages like English are taken
to contain covert variables ranging over worlds; hence the denotation of every clause is
directly a function from worlds into truth-values. For example, the structure of (4) is
given by (8) and its denotation by (9):

(4) Jason is a stockbroker.
(8) $\lambda_0. \text{Jason is a stockbroker } w_0$
(9) $[\lambda_0. \text{Jason is a stockbroker } w_0] = \lambda w'. \text{Jason is a stockbroker in } w'$

In the new framework, attitude verbs are formalized not as parameter-shifters, but rather,
quite simply, as quantifiers that are able to bind world-variables.

In the extensional framework, de dicto and de re readings are generated by letting
the variable associated to a descriptive phrase, respectively, bound or free. It’s easy to
see how this accounts for the third reading. We don’t need to give the indefinite narrow
scope to read it de re, as was the case in the intensional framework. We just let the world
variable associated to it not be bound by the attitude verb:

$\lambda_0 \text{Stacey intends } [\lambda_1 \text{to marry } w_1 \text{[a stockbroker } w_0]] w_0$

Let me emphasize an important point. Despite the shift from intensional to exten-
sional frameworks, the main conception of the de re/de dicto distinction has remained
exactly the same since Quine. Following Quine, let’s say that a description has an opaque
occurrence if it is evaluated at the index introduced by an attitude verb, and that it has
a transparent occurrence otherwise. On the standard conception, the de re/de dicto dis-
tinction essentially lines up with the transparent/opaque distinction. Opaque readings
are de dicto reading. Transparent readings coincide with one of the two de re readings.
Intensional and extensional frameworks differ on how these readings are achieved com-
positionally. This difference is significant, since it yields differences in empirical pre-
dictions; but the basic conceptual issue has remained untouched since Quine.

2.3 Referential opaque readings

2.3.1 The claim: a fourth reading

The standard view of de re/de dicto phenomena has an important consequence: all se-
mantic links between a descriptive phrase and an object are determined via its descrip-
tive content. For example, consider:
2.3. REFERENTIAL OPAQUE READINGS

(10) Stacey wishes that the richest man on Wall Street would ask her out.

(10) may or may not report a belief concerning a specific individual, i.e. the actual richest man on Wall Street. If it does, this must be in virtue of the definite phrase the richest man on Wall Street receiving a transparent interpretation. In this case, we use the actual world to evaluate the description; the denotation of the phrase is determined by finding the individual who satisfies the description. There is no other way for a description to denote: all connections between a description and an individual pass from via the descriptive material.

This is the claim I deny. I’m going to argue that descriptive phrases can involve a referential component that (a) links them semantically to an individual, and (b) is completely independent of descriptive content. In particular, this referential component may be present even when the descriptive phrase receives an opaque reading. To see what this reading amounts to, consider the following scenario:

While strolling around Wall Street, Stacey sees a wealthy-looking man accompanied by an elegant woman. Stacey’s friend, Eileen, tells her that the man is the wealthiest man on Wall Street and Stacey immediately wishes he would ask her out. But the man is actually a clerk living in Queens who enjoys pretending to be a rich banker in his free time.

I claim that, in this scenario, (10) has a true reading where the description the richest man on Wall Street involves a referential component picking out the Queens clerk. Of course, the description must also have an opaque occurrence, i.e. it must be evaluated at the world introduced by wish, since Stacey doesn’t have particular beliefs about the person who is actually the richest man on Wall Street. (That man, suppose, never leaves his house for fear of being kidnapped and Stacey has never seen him.) Hence the description the richest man on Wall Street in (10) refers to a particular individual, though this semantic link is completely independent of its descriptive content.

Another way to put my point is that attitude reports involving descriptive phrases have a fourth kind of reading, in addition to the three already spotted by contemporary theories. I call this reading ‘referential opaque’. Here are the truth-conditions of the referential opaque reading for (10):

(d) Referential opaque: Stacey has a wish concerning a certain man; the wish is that that man, whom she takes to be the richest man on Wall Street, would ask her out.

Proving that the fourth reading actually exists is hard. Notice that the scenario I described cannot distinguish the fourth reading from the more ordinary classical de dicto reading. In that scenario, also the classical de dicto report is true: Stacey does have the
wish that the richest man on Wall Street—read \textit{de dicto}—would ask her out. For all I’ve shown so far, it might be that there is no fourth reading and (10) is heard as true in virtue of its classical \textit{de dicto} reading.

The problem is not confined to (10), but generalizes. Classical \textit{de dicto} readings are, in general, strictly weaker than referential opaque readings. So one can’t prove the existence of the latter by producing a scenario which uniquely validates them. The argument must take a more circuitous route. This is exactly what I’m going to do. I’m going to give two arguments; both are based on a kind of anaphora. The first exploits special properties of gendered pronouns that are anaphoric on descriptive phrases. The second is based on intersentential anaphora. In both cases, the crucial point is that, to account for anaphora, we must assume that descriptive phrases do involve a referential component that is independent of their descriptive content.

Let me note two points before proceedings. First, throughout this section I only make one minimal assumptions about the referential element in descriptive phrases. This assumption is that the referential element is a free variable whose value is provided by context, analogously to pronouns like \textit{she} and \textit{he}. I don’t determine whether descriptions just \textit{are} variables or rather they are syntactically complex entities that \textit{contain} a variable. The reason is that I want to focus attention on the general claim—descriptions are referential expressions—rather than a specific analysis. I will give more details about possible analyses in section 2.4. Hence I limit myself to a schematic representation of the structure of sentences like (10), where indefinites are treated just as predicates, and definites as predicates prefaced by a uniqueness marker (‘\(\iota\)’). For example, (1) and (10) are represented, schematically, as

\[
\text{Stacey intends } [\text{to marry } x_1(\text{stockbroker})] \\
\text{Stacey wishes that } [\iota x_3(\text{richest man on Wall Street}) \land x_3 \text{ would ask her out}]
\]

Second, I should flag that I’m not the first to argue for the existence of a referential opaque reading. Fodor (1970) presented arguments for what is essentially the same claim, though couched in different terms. Fodor’s arguments have turned out to be inconclusive and a consensus has emerged that there is no referential opaque reading. Here I hope to reverse this opinion.\footnote{More precisely, Fodor argues that the traditional \textit{de re/de dicto} distinction confounds two axes of variation. Occurrences of descriptive phrases can be \textit{opaque} or \textit{transparent}, on the one hand, and \textit{specific} or \textit{unspecific}. The specific/unspecific distinction is meant to capture whether the belief reported is about a particular individual or not. These two axes of variations yield the four readings:}

\begin{tabular}{|c|c|}
\hline
\textbf{OPAQUE & UNSPECIFIC} & \textbf{OPAQUE & SPECIFIC} \\
\hline
\textbf{TRANSPARENT & UNSPECIFIC} & \textbf{TRANSPARENT & SPECIFIC} \\
\hline
\end{tabular}
2.3. REFERENTIAL OPAQUE READINGS

2.3.2 The argument from gender features

Traditional analyses take bindable pronouns like she and he to be semantically analogous to logical variables. Of course, though, she and he have richer meanings that variables: they also carry information about the gender of the objects they refer to. Following Cooper (1983), this information is usually encoded in the form of a presupposition:

\[ \text{[she]}^{w,g} = [x_5]^{w,g} = g(5): g(5) \text{ is female in } w \]

(Notice that, following standard accounts of variables, I’m relativizing the interpretation function to an assignment, i.e. a function mapping variables into objects.) Now, Yanovich (2010) has recently pointed out that this analysis is inadequate. The reason is that gender presuppositions have an indexical component in their meaning. She and he invariably specify the actual gender of their referent (provided that that referent exists in the actual world). To see this, consider

(11) Jason is convinced that my personal banker is a man. He also thinks she is a stockbroker.

(12) If Jason’s personal banker were a woman, he couldn’t get himself to trust him.

The anaphoric pronoun which picks out the speaker’s personal banker in (11) is she, since the person is a woman in the actual world. It doesn’t matter what gender Jason thinks my personal banker is. She is appropriate as long as its referent is female in the actual world. Similarly, mutatis mutandis, for him in (12), which involves a counterfactual modal. Hence the Cooper analysis should be revised. On a first attempt, we can take the new analysis to be:

\[ \text{[she]}^{w,c,g} = [x_5]^{w,c,g} = g(5): g(5) \text{ is female in } w_c \text{ if } g(5) \text{ exists at } w_c, \text{ and } g(5) \text{ is female in } w \text{ otherwise} \]

Fodor connects the two axes of variations to the two standard tests for opaque contexts, i.e. failure of substitutivity of coreferential terms and failure of existential generalization. Here I want to remain neutral on this point.

*For simplicity, I’m going to use an intensional system for representing modality throughout the paper. Nothing hangs on this choice.

*This analysis still doesn’t capture all the data pointed out by Yanovich: in particular, it is still problematic when there are multiple attitude verbs embedded under each other. (To my knowledge, no existing analysis manages to get all the data right.) But the problem seems orthogonal to my main concern here.

*Notice that the interpretation function is now indexed also to a context, according to Kaplan’s standard framework for indexicality (1989). Incidentally, this is the official view of the functioning of indexicals, but not the one I ultimately want to adopt. For that, see chapter 1 of this dissertation (and in particular section 5.2 for an alternative treatment of gender presuppositions).
In short, my strategy will consist in exploiting this property of gendered pronouns as a diagnostic. The indexicality of gender features allows me to distinguish between the classical de dicto reading and the referential opaque reading. Let me show how.

Consider the following scenario:

Walking along Wall Street at the end of the working day, Jason notices a distraught woman wearing an elegant man’s suit who complains to a friend “I lost all my money!”. The woman is a street artist and her bag has just been stolen. Because of the circumstances, Jason mistakes her for a male stockbroker who’s had a bad day on the job.

In this scenario, this report is appropriate:

(13) Jason thinks that a stockbroker lost all his money today.

Moreover, and crucially, this is also acceptable:

(14) Jason thinks that a stockbroker lost all her money today.

Admittedly, (14) is not as good a report as (13). But, I claim, it’s still acceptable, and in any case much better than a report where the gender feature is a mismatch in all respects. Consider a modified scenario where the person Jason sees is indeed a man. In this case (14) sounds plainly wrong. This stark contrast needs explanation.

Now, (13) is unsurprising from the point of view of the orthodox theory. The phrase a stockbroker must be evaluated at the world coordinate introduced by thinks, since the relevant person is not a stockbroker in the actual world. Given this, the only available reading on the orthodox theory is the classical de dicto one. Hence the report is not about any actual individual: rather, it ascribes to Jason a purely existential belief—something like “there is a stockbroker who lost all his money”. (Of course, Jason does have a belief about the particular person he saw, but this more specific belief is simply not what is reported by (13).) As a result, the actual gender of the person Jason saw is irrelevant; there is no connection between the report and that person. So it’s unsurprising that the pronoun used is masculine.

On the contrary, (14) is very surprising. As in (13), the indefinite a stockbroker must have an opaque reading. But this time the gender of the pronoun matches the gender of the actual individual Jason saw. This means that there is a semantic link between the phrase a stockbroker, which works as an antecedent for the pronoun, and the actual person Jason sees. The semantics somehow ‘sees’ that Jason has a belief about a real individual who happens to be a woman. Existing views of the de re/de dicto distinction cannot account for this fact.
2.3. REFERENTIAL OPAQUE READINGS

On these grounds, I claim that we should reject the orthodox account and admit the existence of a referential opaque reading. On this reading, the indefinite a stockbroker involves a referential component picking out the woman Jason actually saw. Thus the schematic representation of (14) is:

Jason thinks \[ x_5 \text{ (stockbroker)} \] and \[ x_5 \text{ lost all } x_5 \text{'s money today} \]

Clearly, an account of descriptive phrases that implements this idea has the potential to accommodate all the data. Given the scenario, the variable \( 'x_5' \) is assigned by context an individual who is female in the actual world. Since gender presuppositions are indexical, it is expected that the feature associated to the pronoun is feminine. Hence the acceptability of (14) is evidence for the claim that the occurrence of a stockbroker in it involves a referential element.

Let me clarify an important point. The foregoing does not constitute a full analysis of the data (and in particular, of (14)). Aside from giving more details on descriptive phrases, a full analysis would require (a) adopting a theory of attitude verbs, (b) giving a full account of gender presuppositions, and (c) showing how these three components interact. This is beyond the scope of this paper, especially since currently there is no developed theory of the functioning of gender presuppositions. Nevertheless, I take that the foregoing is an argument against the orthodox view. First, (14) shows that this view undergenerates. Second, and more importantly, the evidence shows unequivocally that there can be a semantic link between opaque descriptions and objects outside the domain of the subject’s attitudes. This is enough to make my point.

2.3.3 The argument from intersentential anaphora

My second argument comes from a domain that is traditionally important for the semantics of definite and indefinite phrases, namely anaphora. There is little or no work examining how anaphora interacts with attitude reports: I hope to show that we can make important discoveries by looking in this direction.

Consider the following scenario:

Jason sees a man wearing an elegant business suit jump off the edge of a tall building overlooking Wall Street. Jason assumes that the individual is a suicidal stockbroker. But the man is actually a professional stuntman who’s training for his next performance.

The following is true in this scenario:

(15) Jason believes that a stockbroker jumped off a building. But he was actually not a stockbroker, but a stuntman.
Here is my basic argument. The indefinite *a stockbroker* in (15) must be given an opaque reading, for the usual reasons. At the same time, the indefinite must incorporate reference to a specific individual. So (15) is schematically represented as:

Jason believes that \[ x_4(\text{stockbroker}) \text{ jumped off a building}. \] But \[ x_4 \] was actually not a stockbroker, but a stuntman.

The reasoning is elementary. If *a stockbroker* did not refer to the man Jason saw, it would make no sense for the speaker to go on and say that Jason is wrong about that individual’s profession. In other words, if there were no semantic link between *a stockbroker* and the relevant individual, the speaker couldn’t refer to the relevant individual via an anaphoric *he*.⁹ Hence, on the true reading of (15), the indefinite *a stockbroker* has a referential opaque reading.

The basic argument is very simple; the bulk of the work consists in defusing alternative strategies of accounting for (15). There are two such resistance strategies. The first claims that *he* in (15) is not an anaphoric pronoun, but rather a deictic one. The second resorts to a so-called e-type analysis of *he*, according to which it is a kind of covert description. Let me discuss them in turn.

**The salience strategy.** The first resistance strategy involves denying that *he* is a genuinely anaphoric pronoun. The report in (15) only has a (true) classical *de dicto* reading, the argument goes. At the same time, it raises to salience the fact that there is a specific individual that is causally responsible for Jason’s belief. Now, we know that pronouns like *he* can have deictic uses, i.e. they can be used to refer to salient individuals in the context (often with the accompaniment of pointing). The occurrence of *he* in (15) exemplifies just one of these uses.

The deictic strategy has some important and—in my view—decisive shortcomings, originally pointed out by Heim (1982) and (1990) (see also Kadmon (1987)). Heim’s point is that pronouns like *he* in (15) require linguistic antecedents of a very specific kind. Even when certain individuals are made very salient in the context, pronouns cannot pick up on them in absence of the appropriate linguistic antecedent. In short,

⁹To better see this, contrast the case in the text with the following variant:

While walking along Wall Street, Jason has a veridical-seeming hallucination of a stockbroker jumping off the edge of a building. At the same time, and by a striking coincidence, a stuntman dressed up as a stockbroker jumps off a building in a way such that, if Jason had been actually seeing, the scene in front of his eyes would have been exactly the same.

In this modified scenario, (15) becomes inappropriate. The reason, I claim, is that Jason is not in a position to have attitudes about the jumper. Of course, Jason still has a classical *de dicto* belief. But this is insufficient to license the anaphora.
pronouns like *he* in (15) require a formal link obtaining between them and their antecedent.

To see how this applies to our case, consider the following variants on (15):

Jason sees three men wearing business suits jump off a building; two are wearing a parachute. Jason assumes that they're all stockbrokers and that the one with no parachute must have died. But they are all stuntmen who manage to land safely.

(16) Jason believes that three men have jumped off a building and only one has died. He has actually survived.

(17) #Jason believes that three men have jumped off a building and only two have survived. He has actually survived too.

The opening clauses in (16) and (17) are truth-conditionally equivalent. Yet the anaphora in (16) is successful, while the one in (17) isn't. This shows that a pragmatic notion of salience that doesn't track syntactic facts cannot explain the functioning of anaphora. *He* in (16) and (17), as well as in (15), requires a linguistic antecedent.¹⁰

Let me point out that my rejection of the salience strategy doesn't entail that pragmatic factors are wholly irrelevant to the interpretation of (15). On the contrary, context plays a crucial role, since it determines what individual should be assigned to the free variable involved in *a stockbroker*. My point here is just that the link between *a stockbroker* and *he* must rely on syntactical and semantic scaffolding and can't be explained by salience alone.

**The e-type strategy.** The second resistance strategy involves claiming that *he* in (15) is a so-called e-type pronoun.¹¹ E-type accounts essentially treat anaphoric pronouns as covert definite phrases. Again, the strategy claims that (15) has a true classical *de dicto* reading. The speaker's correction is explained by the fact that *he* is a covert description involving reference to the individual causally responsible for Jason’s beliefs. In essence, the idea is that (15) is synonymous with something like:

(18) Jason believes that a stockbroker jumped off a building. But the person respon-

---

¹⁰Heim (1990) also notices that mention of the NP earlier in the sentence is not sufficient for successful anaphora; it is necessary that the NP be part of an indefinite or a definite noun phrase. This also concerns NPs embedded under attitude verbs; consider:

??Jason believes that the husband-wife relation requires trust. As a husband, he always trusts her.

It's perfectly clear what is meant, yet the occurrence of *her* is still infelicitous.

¹¹See, among many, Heim (1990), Neale (1990), and Elbourne.
sible for Jason's belief/the person Jason saw/etc. was actually not a stockbroker, but a stuntman.

The e-type strategy is currently a very live contender in debates about anaphora. But it lends no help in this case. In fact, just cases like (15), I claim, pose a major and so far undiscovered problem for e-type theories in general. Let me explain.

E-type theories are designed just to respond to the problem of the formal link I outlined above. Consider the pair:

(19) Jason's personal banker has a wife. She is a stockbroker.
(20) #Jason's personal banker is married. She is a stockbroker.

An account of anaphora must predict that the occurrence of she is felicitous in (19) but not in (20). Now, for the e-type theory, she is accounted for in the same way in (19) and (20): in both cases, it's a kind of covert description. Thus, if we imposed no constraints on what the covert description is in the two cases, we could get a felicitous interpretation for both (19) and (20). For example, we could get that in both cases the covert description is Jason's wife:

(21) Jason's personal banker has a wife. Jason's wife is a stockbroker.
(22) Jason's personal banker is married. Jason's wife is a stockbroker.

Hence, if we want to predict the asymmetry between (19) and (20), we must impose some constraints preventing (20) to be interpreted as (22). The basic idea shared by all e-type accounts is that the description must be reconstructed from the linguistic material of an antecedent descriptive phrase. For example, she in (19) must pick up the noun phrase from the indefinite a wife. Conversely, (20) is infelicitous because there is no appropriate descriptive phrase to work as the antecedent for she.

I’ve been speaking informally of pronouns ‘picking up’ linguistic material from their antecedent. The details of this process are actually quite complicated, and make an important difference to the final form of e-type theories. But these differences are irrelevant for my purposes. All e-type accounts have the same problem accounting for (15). The reason is simple: the linguistic material of the antecedent is just the wrong thing to use in building a covert definite phrase referring to the relevant object. All we could pick up from the antecedent of he is the noun phrase stockbroker. As a result, by using an e-type account we would make (15) synonymous with:

¹²Though consensus seems to be slowly peeling away from it. See, for example, Schlenker (2011).
¹³In particular, e-type approaches diverge on whether the ‘picking up’ process is semantic (as in Chierchia (1992)) or syntactic (as in Heim (1990) and Elbourne).
(23) Jason believes that a stockbroker jumped off a building. But the stockbroker was not a stockbroker, but a stuntman.

(23) is obviously nonsense. The problem, of course, is that the e-type pronoun picks up linguistic material that is evaluated at the world parameter introduced by ??, and uses it in a position where it must be evaluated at a different world parameter, that evaluated by the context. This creates trouble: he in (15) cannot be construed as the stockbroker or a similar description, simply because the relevant person isn’t a stockbroker. Hence just the fact that the indefinite phrase has an opaque reading prevents it from providing linguistic material to construct a suitable description.¹⁴

Hence e-type theories cannot yield an alternative account of (15), Indeed, data combining anaphora and attitude reports seems to raise a significant, and henceforth undiscovered problem for these theories.

So both resistance strategies fail. I conclude that examples like (15) provide further evidence for the claim that descriptive phrases have referential opaque uses. A stockbroker in (15) refers (or at least, incorporates reference) to the actual individual Jason has seen. This is what makes anaphora possible.

2.4 Two routes to an account

If the arguments of section 2.3 succeed, descriptive phrases like a stockbroker and the richest man on Wall Street can have referential opaque readings. This means that descriptive phrases may involve reference to an individual, on a par with referential expressions like names, indexicals, and ordinary pronouns. The referential element in their semantics is independent of their descriptive content: as a result, they may refer

¹⁴Is there really no way out for e-type theories? So far as I can see, to give an e-type account of (15), we would need to tweak the e-type theory in one of two ways. The first would be to go back to a more unconstrained theory. This seems to have disastrous consequences, reopening the problem of the formal link posed by pairs like (19) and (20). The second would be trying to reconstruct, in addition to the noun phrase, also the shifted world parameter at which a stockbroker is evaluated. On this route, we would need to have a variable that is bound by an attitude verb despite being outside its scope. I.e., we would have a structure of this sort:

$$\forall w'[ \ldots w']. (\text{stockbroker})_{w'} \ldots w_c$$

This is not unfeasible: it is essentially what dynamic accounts of anaphora (descending from Heim (1982) and Kamp (1981)) do. But, at least from a historical point of view, the rationale for developing e-type accounts of intersentential anaphora was just providing an alternative to dynamic accounts. So, if we wanted accepted dynamic binding of world variables, it's unclear what reason would be left for buying into an e-type account in the first place.
even when their descriptive material receives an opaque reading.

The foregoing requires significant changes to current accounts of descriptive phrases. Giving and defending a new account is beyond the aims of this paper. But, in this section, I survey two natural lines of inquiry for building such an account. This work will hopefully serve as a blueprint for building more developed theories.

2.4.1 The DRT analysis: descriptions as variables

The first line of inquiry descends from the treatment of descriptive phrases in Discourse Representation Theory (DRT), initially developed by Heim (1982) and Kamp (1981). On the DRT view, definite and indefinites alike just are, from a syntactic point of view, variables. The descriptive material adds a presupposition on the denotation of the variable. For example, a stockbroker is a free variable whose denotation is determined by an assignment, just like he; in addition, it presupposed that the individual denoted be a stockbroker. Formally:

\[ x_2(\text{stockbroker}) \]

Changing the functioning of descriptive phrases brings with it a number of other changes. This is very evident for the case of DRT, which rejects the orthodox conception of meaning as truth-conditions and moves to a picture where meanings are instructions to change the common stock of knowledge of the speakers. Several aspects of the meaning of descriptions—for example, the difference between definites and indefinites—are captured by DRT just by exploiting features of the new model. It's unclear how easily these parts of the descriptions-as-variables analysis could be extricated from the overall architecture of DRT. Perhaps treating descriptions as variables doesn't require going along with DRT the full way, but it anyway calls for non-negligible adjustments.¹⁵

Whether we go along full-throttle with DRT or not, just adopting a descriptions-as-variables analysis is not enough to solve the problem. This analysis systematically turns all descriptive phrases into variables. But the data I introduced suggests that some descriptive phrases have referential readings; at least some other ones have more traditional readings (for example, in classical de dicto attitude reports). One should show how the descriptions-as-variables analysis can yield readings of both kinds. Explaining how this can be done is not an easy task, and is anyway beyond the scope of what I want

¹⁵The distinction between definite and indefinite phrases is captured in DRT not via lexical meanings, but via a general condition stating that indefinite phrases open a 'new file' about an individual, while definite phrases are meant to add information to a file that is already open. (For a full illustration of the file metaphor, see Heim (1982) and Heim (1983).) It's not clear to me how this distinction can be reconstructed with completely traditional tools.
to do here.¹⁶ But let me just point to a natural direction of inquiry.

The suggestion is that the distinction between referential and non-referential readings should be captured via variable binding. Consider again (1):

(1) Stacey intends to marry a stockbroker.

On the referential reading, a stockbroker would work as a free variable, following the indications I gave above. On non-referential readings, by contrast, a stockbroker might be bound by the attitude verb intends. (For a general semantic account that allows attitude verbs to bind variables see chapter 1; similar ideas are used in DRT-style accounts of modals and adverbs of quantification.) This would allow us to capture a distinction in truth-conditions, as well as avoiding the problem of having to link a stockbroker to a particular individual on the classical de dicto reading.¹⁷

In summary, the descriptions-as-variables idea is potentially promising, but requires significant work. A proper development of the idea, as well as a proper assessment of it, are better left to further work.

2.4.2 A conservative analysis: variables in the restrictor

The second line of inquiry is more conservative. Following the quantifier-style analysis, indefinite and definite phrases alike are given quantificational force. Hence the basic form of the truth-conditions of, say

¹⁶So far as I can see, no solution can be extracted easily from existing versions of DRT. The only explicit account of de re/de dicto phenomena in DRT is given by Maier (2010). Maier’s account is rather complex, analyzing de re/de dicto phenomena as presuppositional phenomena and combining a basic DRT semantic with van der Sandt’s (1992) account of presupposition projection. But the main issue is that, on this framework, the variables introduced by opaque descriptive phrases have no connection to items outside the belief report. Hence Maier’s framework runs into exactly the same problem as the classical framework.

¹⁷There is an obvious prima facie tension between this idea and the account of attitudes that I defend in chapter 1. On that account, variables can have free occurrences within the scope of attitude verbs. Yet in chapter 1 I argue that individuation of objects within attitude states must always pass via representatives; semantically, this idea translates into the idea that all referential items in the scope of attitude verbs must always be bound. But the tension can be solved. The way to do this is to assume that both referential and non-referential occurrences of description are bound when occurring under attitude verbs, only in different ways. In particular, the binding of referential descriptions will follow the account defended in chapter 1: these descriptions will range over counterparts by acquaintance of the relevant objects; attitude verbs will ‘check’ that they are actually linked to an actual world object. By contrast, nothing like this will happen for the case of non-referential descriptions. This will produce the wanted truth-conditional difference between the two readings. Of course, the compositional implementation of these mechanisms is non-trivial; this is one of the many technical issues that a descriptions-as-variables analysis should address.
A stockbroker fled with all my money.

are still existential (“there was a stockbroker who fled with all my money”). The idea is that we can capture the referential readings of descriptions by inserting a referential component somewhere in their structure. As I’m going to show, this hybrid theory is very easy to develop: indeed, it basically falls out of a standard quantifier-style theory of descriptions, plus some minimal assumptions.

The natural place to locate a referential element in the structure of a description is the covert restrictor. There is wide agreement, in philosophy and semantics, that all quantified phrases have a covert argument which restricts the domain of quantification (see, among many, von Fintel (1994), Stanley & Gendler Szabó (2000)). Suppose that Stacey, while talking about the Wall Street population, says:

Every stockbroker is filthy rich.

Arguably, (25) has a reading on which the quantified phrase every stockbroker doesn’t range over all stockbrokers in the world, but just over stockbrokers working on Wall Street. (In fact it looks like this reading is the one that would be most naturally intended in the context.) This reading is normally predicted by assuming that quantified phrases like every stockbroker have a covert argument for domains of quantification. As usual in these cases, the covert argument is represented syntactically as a variable whose value is provided by context.¹⁸ So the structure of (25) is:

Every [stockbroker D] [is filthy rich]

(For simplicity, I assume that ‘D’ is just a variable ranging over sets of individuals. For discussion of this point, see Stanley & Gendler Szabó (2000).)¹⁹

Now, we want to build a variable in descriptive phrases. One natural thought is that the variable could be contained just in the covert restrictor. Suppose that the covert restrictor doesn’t come under the form of a variable ranging over sets, but rather in the form of a complex \( \langle f(x) \rangle \), where ‘x’ is a variable over individuals, and ‘f’ a variable ranging over functions from individuals into sets. (This, indeed, is exactly the analysis of covert restrictor that is proposed, on independent grounds, in Stanley & Gendler Szabó (2000).) On this analysis, the structure of (24) and (25) will be, respectively

¹⁸Notice: I’m assuming, with Stanley (2000), that anything that has an effect on the truth-conditions of a sentence must be traced to syntactic items. This assumption is far from uncontroversial, at least in philosophy. I make it simply because I’m fully convinced by the arguments in its support; quantifier domain restriction seems just to provide a strong case in its favor, as Stanley & Gendler Szabó (2000) show.

¹⁹I’m also assuming that the covert restrictor and the overt predicate stockbroker would be able to combine via a Predicate Modification-style rule (see Heim & Kratzer (1998)).
2.4. TWO ROUTES TO AN ACCOUNT

A [stockbroker \(f(x)\)] [fled with all my money]

Every [stockbroker \(f(x)\)] [is filthy rich]

How does this work? In cases like (25), the changes are minimal. The only difference is that we pick our set in a more roundabout way, by first fixing on a salient object in the context, and then recovering a set of individuals as a function of that object. But now, suppose that I utter (24) intending to pick out a specific individual—say, Bernard Madoff. In that case, I assume (a) that the individual variable \(x\) is sent into the relevant individual, namely Bernard Madoff, and (b) that the function variable \(f\) is sent into a function taking individuals into their singletons. As a result, the truth-conditions of (24) are the following:

\[
[a \ [\text{stockbroker } f(x_1)] \ [\text{fled with all my money}]]_{c,w,g} = \text{true if and only if }
\exists x \text{ s.t. } x \in \{\text{Bernard Madoff}\} \text{ and } x \text{ is a stockbroker and } x \text{ fled with all the speaker's money.}
\]

As is evident, the truth-conditions involve existential quantification. But this quantification is truth-conditionally idle: the domain is restricted to a unique individual, and moreover this individual is picked out directly by the variable in the covert restrictor. Hence the reading is truth-conditionally equivalent to that of a sentence involving just a referential expression, with no quantifier.

Notice that, interestingly, the formal resources that this account uses are exactly the ones provided by one of the standard theories of covert restrictors. I have just used the account in Stanley & Gendler Szabó (2000). The only extra assumptions concern the fixing of the covert variables. These assumptions are, first, that the 'f' variable can take as value the function taking individuals into their singletons, and second, that the assignment of the individual variable can be determined by the referential intentions of the speaker. Both these assumptions seem quite natural. Hence the conservative account is theoretically cost-free; indeed, it looks like we should expect to get referential readings of descriptions, given the way the standard semantics for covert restrictors works.

Despite these advantages, I don't want to endorse the restrictor analysis here. This for two reasons. First, the analysis I sketched has implications for all quantified phrases in natural language, and not only for definites and indefinites. All quantifiers in natural language have covert restrictors. Hence, if the restrictor analysis is correct, we should expect to get referential readings for other quantified phrases than definites and indefinites. More empirical investigation is needed to establish whether this is true.²⁰ Second,

---

²⁰For an interesting argument that all quantifiers in natural language behave in a fundamentally analogous way with respect to anaphora, see Nouwen (2003). It should be noted that Nouwen's analysis takes place in a DRT framework.
I mentioned already that the data of section 2.3 pose a substantial problem for e-type theories. Now, the main alternative to e-type-theoretic account of anaphora are just accounts of pronouns based on a DRT analysis. So it might be that eventually we are forced to go with a descriptions-as-variables analysis anyway. Also in this case, more work is needed to settle the issue.

2.5 Outstanding issues

Let me mention two related issues that are worth pursuing elsewhere.

**Donnellan’s attributive/referential distinction.** In a suggestive paper, Donnellan (1966) claimed that definite descriptions can have two readings. One of them is attributive, i.e. aims at singling out an individual on the basis of a descriptive condition; the other is referential. Importantly, according to Donnellan, utterances involving referential (but not attributive) uses of descriptions may be true even if the description doesn’t apply to the relevant object. If my claim is correct, Donnellan’s distinction might have a semantic vindication. In particular, if the covert restrictor approach is right, we could explain how definites (on a par with other quantified expressions in language) have both referential or attributive readings. Notice that, even though this approach would ground the attributive/referential distinction in the semantics, it would avoid postulating any sort of structural ambiguity. The difference between attributive and referential readings would just amount to a difference in the covert material provided by context.²¹

**Impact on a general account of de re/de dicto phenomena.** As I said in section 3.2, extensional accounts of modal talk have achieved something like the status of orthodoxy. But these accounts suffer from overgeneration, as has been pointed out, for example, by Percus (2000), Keshet (2008), and Romoli & Sudo (2008). As a result, some theorists are promoting a switch back to refined versions of the intensional theory (for two attempts in this direction, see Maier (2010) and Keshet (2011)). Now, opaque referential readings are clearly unexpected on the old-style intensional theory; accounting for opaque referential readings seems an important test case for the new breed of intensional accounts.

Finally, let me close with a brief summary. I have argued that definite and indefinite descriptive phrases may have a referential component in their meaning. This referential

---

²¹Aside from explaining Donnellan’s observations, the availability of this solution would also be important for the methodological debate—started with Kripke (1977)—about the merits of semantic and pragmatic accounts of Donnellan’s ambiguity.
component is independent of the descriptive material; hence there are basically two unrelated dimensions to the meaning of descriptions. The argument has passed through the semantics of attitude reports. I argued that descriptions occurring in attitude reports have referential opaque readings: they can refer to an individual, and at the same time their descriptive material can be read opaquely. If this is right, descriptive phrases are more like names and, in particular, pronouns than is commonly thought. Exactly how much they are like pronouns remains up for grabs.
3. Cognitive Relocation

3.1 Introduction

Imagine an omniscient god who lives in a fully deterministic world. At any time, the god has complete information about the past, present, and future, including information about who she is and what time it is. No news can come to her: she’s not in a position to learn anything. But since the god lives in time, something in her belief state must keep changing as time goes by. Suppose that today, August 31st 2011, she believes it’s August 31st 2011: if she has to stay omniscient, tomorrow she must believe that it’s September 1st 2011. She must undergo a cognitive change to keep her beliefs up to date. What is this change and what rational constraints apply to it?

The case of the god is an instance of a general problem concerning agents who occupy a position in time. As this position changes, agents must modify their beliefs to keep track of it. Providing a precise model of this kind of belief change has proved surprisingly problematic. In philosophy of mind, we have a developed framework for representing \textit{de se} attitudes, i.e. attitudes about one’s own position in the world: for example, the beliefs that you are fond of rutabagas or that today someone will eat rutabagas. In epistemology, we have a probabilistic framework for modeling rational belief update, the so-called Bayesian framework. But it’s not clear how to bring the two together. The Bayesian framework is designed to handle \textit{de dicto} beliefs, i.e. objective, third-personal beliefs. But \textit{de se} attitudes seem to evolve in ways that are incompatible with standard Bayesian theory: the god’s update is just an example of this.

This paper aims at providing a model of content and update that bridges this gap. The main claim is that the Bayesian framework is not unfit to handle \textit{de se} content: rather, it just needs supplementation. Rational change in \textit{de dicto} belief is normally due to learning new information. By contrast, rational change in \textit{de se} beliefs is sometimes due to learning new information and sometimes due to a different process, which I call ‘cognitive relocation’ (or, more simply, ‘relocation’). Cognitive relocation is just the
CHAPTER 3. COGNITIVE RELOCATION

operation whereby agents keep track of their moving position in time and space.

The core of the paper is devoted to giving an account of relocation; but here I can give a quick preview. The account is organized around two main ideas, which straddle philosophy of mind and epistemology. The first is about content. There is a kind of information that takes center stage in a theory of temporal update: this is \textit{de re} information about time. In short, \textit{de re} information is tied to the subject’s capacity of acquiring and retaining information via a causal route, for example via perception and memory. An example of a \textit{de re} temporal attitude is your belief that at \textit{this time}, individuated on the basis of your current perceptions, you’re reading a philosophy paper. This notion of \textit{de re} information shares features with both \textit{de se} and \textit{de dicto} information. On the one hand, it is obviously tied to a subject’s occupying a certain position in the world. On the other, it’s unproblematic to see how this information can be retained through time.

Second, relocation is essentially a way of retaining rather than gaining information. By relocating effectively, a subject preserves her level of certainty about \textit{de re} temporal information. If you keep perfect track of time, you’ll always be certain about how far away \textit{this moment} is from \textit{that moment} you have memories of. Conversely, by relocating ineffectively, a subject becomes more uncertain about this kind of information. By losing track of time, you become uncertain about how far apart different times you individuate \textit{de re} are—say, how far \textit{this new time} is from \textit{that time} you have memories of. So relocation differs from more ordinary update procedures. The latter systematically reduce your uncertainty, taking you from a state of uncertainty to a state of certainty; in the worst case scenario, you learn nothing, and your credal state stays unchanged. Relocation aims at preserving your current epistemic ground, preventing you from spreading your credences more than they are currently spread. Technically, this kind of update is captured by an update operation that generalizes Lewis’s imaging procedure (1976). This confirms that relocation is a very different process from ordinary learning, since the latter is ordinarily modeled by conditionalization, which has been shown to be incompatible with imaging.

The theory has interesting philosophical consequences. Among these, it vindicates an intuitive principle about change the relationship between \textit{de se} and \textit{de dicto} information, that Titelbaum (2008) calls ‘Relevance-Limiting Thesis’: or better, it vindicates a suitably disambiguated version of it. The principle says that a subject who only updates her beliefs in response to a change in her position in the world should not modify any of her attitudes about \textit{de dicto} information. In essence, the principle states that mere changes in location are irrelevant to a subject’s \textit{de dicto} attitudes. The principle is plausible, but it has been thought to be incompatible with some intuitive verdicts about a much discussed example, the so-called Sleeping Beauty problem. In the final part of the paper I show how, contrary to common assumptions, my framework can both vindicate the principle and yield the intuitive verdicts about the Sleeping Beauty case.
This is the structure of the paper. Section 3.2 contains a brief recap of the Bayesian framework and of the challenge posed by self-locating information. In section 3.3, I state informally my account, which is then implemented formally in section 3.4. Section 3.5 illustrates how my theory brings to bear on a question that has been the subject of much recent debate, i.e. the question of the epistemic relevance of \textit{de se} information. Finally, section 3.6 investigates the relationship between relocation and communication of \textit{de se} beliefs, paving the ground for a comparison between my account and another recent theory.

## 3.2 \textit{De se} attitudes and the Bayesian framework

### 3.2.1 \textit{De se} attitudes

\textit{De dicto} beliefs are beliefs about what the world is like; \textit{de se} or self-locating beliefs are beliefs concerning your position in the world. The master argument for drawing the distinction is that \textit{de se} attitudes play an essential role in the explanation of action.¹ Borrowing an example from Perry (2006): suppose that you and Perry are sitting on the opposite ends of a table and that there is a large cake between you. The two of you agree on all \textit{de dicto} beliefs and desires that are relevant to the situation. You both believe that you made the cake, and you both want that Perry, rather than you, eat the cake. Nevertheless, the two of you should act very differently to fulfill your goals. You should push the cake towards Perry and shout “Eat it!”; Perry should start digging into the cake with his fork. The difference between you and Perry is not captured by any difference in your mental map of the world. Rather, it is a difference in where you locate yourselves on the map. Perry has the self-locating belief that he would express as \textit{I am the one we both want to eat the cake}. You don’t have that belief.

The \textit{de se}/\textit{de dicto} distinction is reflected in formal models of attitudes. Traditional models represent the information contained in a belief as a set of possible worlds, i.e. the worlds compatible with that belief. Following Lewis (1979), we can extend them to \textit{de se} attitudes by taking the basic objects of belief to be sets of \textit{centered worlds}, i.e. triples of possible worlds, times and individuals.² The possible worlds component models the \textit{de dicto} part of a subject’s beliefs; the times and individuals component model the \textit{de se} part. A subject’s overall belief state at a time is modeled as the set of centered worlds that are compatible with what she believes—call it her \textit{doxastic set}. A subject believes

---

¹Arguments of this sort are found in Perry (1977) and Perry (1979); see also the discussion in Perry (2006), Stalnaker (2006) and Stalnaker (2008).

²Exegetical note: for Lewis centered worlds are pairs of worlds and individual stages rather than triples of worlds, times and individuals. Arguments for taking the latter option are given by Egan (2004).
a set of centered worlds (a centered proposition) just in case her doxastic set is a subset of that proposition. For example, Jason believes at $t$ that he will land on the moon the next day just in case

$$\text{for all centered worlds } \langle w, t, x \rangle \text{ in Jason's doxastic state, } \langle w, t, x \rangle \in$$

$$\{\langle w', t', x' \rangle : x' \text{ will land on the moon on the day after } t' \text{ in } w'\}$$

Centered worlds that are compatible with what $S$ believes at $t$ are called $S$'s doxastic alternatives at $t$. For convenience, I will call the times that are members of $S$'s doxastic alternatives at $t$ and $w$ her temporal alternatives, and the individuals that are members of her doxastic alternatives at $t$ and $w$ her personal alternatives.

3.2.2 The Bayesian framework and the problem of update

According to the Bayesian picture, an agent's overall belief state can be modeled as a probability distribution over sets of possible worlds: the probability assigned to a world represents the agent's degree of belief, or credence, in that world. Since propositions can be characterized as sets of possible worlds, the credence an agent assigns to a proposition $P$ will be simply the sum of the credences she assigns to the members of $P$. For simplicity, throughout this paper I will assume that agents' credences are defined over a finite number of worlds.

In this framework, the axioms of the probability calculus work as basic constraints of rationality for an agent's credence distribution at a given time. But they are silent about how an agent's credences should evolve over time: in particular, they don't say how an agent's credences should be modified in the light of new evidence. The rule that specifies how update should proceed is conditionalization, defined as follows:

$$Cr_{E}(A) = Cr(A | E) = \frac{Cr(A \land E)}{Cr(E)}$$

(‘$Cr_{E}$’ denotes the agent's new credence function after the update on her evidence $E$, and $A$ is an arbitrary proposition.) Intuitively, conditionalization can be understood as a two-step process. First, once you learn $E$ you erase all your credence in non-$E$ worlds. Second, you take the credences in the remaining worlds and you magnify their values in a way that the new values sum to 1 again.

Notice that, once an agent assigns credence 1 to a proposition, it is impossible to lower it via conditionalization. Conditionalizing amounts to erasing your credences in some worlds and renormalizing the remaining ones. If all worlds with positive credence validate a proposition $P$, then all worlds remaining after update by conditionalization, no matter what you learn, will still validate $P$. Exactly for this reason, self-locating information is problematic for the Bayesian framework. Suppose that at noon you’re absolutely confident that it's not 3 pm. As time passes, you should gradually gain credence...
in that proposition. But this is impossible if you update by standard conditionalization, since conditionalization doesn't allow you to lower your credence in propositions to which you assign credence 1.

On a first take, the problem of de se update seems to be an instance of the familiar problem for Bayesian frameworks. Suppose that a coin is tossed an infinite number of times and that the outcome of each coin toss is independent of the others. It seems plausible that, for any infinite heads-tails sequence $S$, the credence you should assign to the coin landing in accordance to $S$ is zero; yet you can learn that a particular sequence is the one that is actually instantiated by the coin. Hence you should be able to learn something to which you previously assign probability zero.³

But there is an important difference between the classical problem of learning propositions with probability zero and the problem of de se update. Given an arbitrary sequence $S$, the probability you assign to the coin landing in accordance to $S$ is zero; but, crucially, it's still an open epistemic possibility for you that the coin lands in accordance to $S$. But things are different with de se update. At noon, you are completely certain that it's not 3 pm. That possibility is incompatible with what you believe at that time. Yet you're somehow able to reopen that possibility in due time. My account of relocation accepts this intuitive difference: the problems are distinct. The problem of learning propositions with probability zero is a technical problem tied to the setup of Bayesian machinery. The problem of de se update has more philosophical roots and requires recognizing a very different kind of belief change. Let me illustrate this idea.

### 3.2.3 Preview: two routes to update

The major assumption that I make in this paper is the following. The problem of intertemporal update emerges because two different kinds of cognitive processes are conflated. One is genuine learning: this is the process where a subject moves from a state of uncertainty to a state of certainty with regard to a proposition by discarding information that was previously compatible with her beliefs. This is the process that is normally and appropriately modeled by conditionalization. The other is relocation, i.e. the process whereby subjects readjust their credences in response to changes of position in the world. These process are very different in nature and must be modeled in different ways.

This is not the orthodox way of framing the problem of de se update. Most often, the problem is characterized as stemming from de se information by itself: on this picture, standard update rules like conditionalization simply fail for de se information. But this is a mistake. It's important to notice that conditionalization is perfectly appropriate to model update with de se information for a whole range of cases. Unsurprisingly, these

³For a classical discussion of this problem, see (among many others) McGee (1994).
are intuitively cases of learning: again, cases where certainty is gained by ruling out open possibilities. Here is one:

Locked up in a dark cell, you know that it’s either noon or midnight: for each of the two possibilities, you have .5 credence that it is that time. You know a lot about the objective circumstances: in fact, you know that both at noon and at midnight you’ll be wondering what time it is, locked up in your cell. All you lack is the relevant self-locating information—the information that now it is noon or midnight. Taken by pity, the prison guard sends you a message on your (clockless) i-Phone, telling you it is noon. (For simplicity, assume that your receiving the message and your updating take place instantaneously.)

Conditionalization is perfectly suited to model your update in this case. You erase your credence in the midnight temporal alternatives and renormalize accordingly, becoming certain that it’s noon.

What is problematic is the kind of update that the omniscient god must perform. This case does not involve a subject passing from a state of uncertainty to a state of certainty with regard to a certain proposition. At every moment, the god is perfectly omniscient with regard to any information she might have. The point is that something has to change if she has to preserve her omniscience. This kind of case just doesn’t fit the mold of a Bayesian theory of update. Of course, the god’s case is just an extreme case, but there are more mundane examples: ordinary believers also live in time, hence they must constantly undergo an analogous kind of update. This kind of update just isn’t captured by standard techniques. Let me point out that I’m not the first to make this claim. Arntzenius (2003) discusses an example similar to the god’s case, claiming that it it shows that credences of rational agents “can develop in ways that are as yet unrecognized, and indeed are not allowed according to standard Bayesian lore” (p. 366). This paper aims just at accounting for this new kind of update.

Before turning to this account, let me notice two points. First, actual agents relocate only along the temporal dimension. Throughout their lives, believers ordinarily change their location in time. But they don’t change their identity (perhaps aside from gory scenarios involving fission and fusion). So my account of relocation will deal exclusively with the temporal case. Nothing in the abstract structure prevents it from being extended also to the personal case; but it’s not clear what kind of phenomenon would mandate a cognitive operation of this kind. So I will focus on time, and will leave investigation of further applications of the framework to future work.

Second, taking this line is not the only way to tackle the problem of lowering credences in propositions with probability 1. Another option is (following an unorthodox
3.3. TEMPORAL RELOCATION

but numerous Bayesian contingent⁴) to deny that rational agents can assign credence 1 to any proposition aside from logical truths, ditch standard conditionalization, and switch to an update procedure that validates this assumption (for example, Jeffrey conditionalization). This solves the immediate problem of explaining how agents can lower their credences in propositions about time, since those propositions never get assigned credence 1. Of course, though, the unorthodox Bayesian still has the burden of showing that she can make correct predictions in general. I choose my solution because I find it more promising, both philosophically and technically. The driving intuition is that, in cases like the god’s example, the probability 1 problem is just a symptom of a more general conceptual difficulty with treating all instances of update on a par. I don’t have direct arguments for this choice. But I hope to convince you that, if we make it, we get an elegant and robust theory that yields surprising predictions.

3.3 Temporal relocation

3.3.1 De re information

The omniscient god I described in the introduction does perform a kind of update: hence the centered worlds information in her doxastic state will keep changing. But, if she really is omniscient and her omniscience is preserved through time, this update doesn’t involve the discovery of any new information. Obviously her belief state changes in some way, yet the god doesn’t get any news. If at noon on August 31st 2011 she knows that it’s noon on August 31st 2011, she also knows that in two hours it will be 2 pm, that in one day it will be September 1st, and that in sixty years it will be 2071. If she keeps perfect track of time as time passes, the information she has is just preserved through time. Of course, the center of the unique centered world in her belief state keeps changing; but each change is perfectly expected.

The basic idea behind my account of temporal relocation vindicates the intuition that the god doesn’t get any news. There is a kind of temporal knowledge that, if a subject keeps track of time, remains stable as time passes. This knowledge involves a particular kind of information: de re information about temporal distances. I model keeping track of time as preserving one’s level of certainty in this information, and losing track of time as becoming more uncertain about it. Hence relocation differs in a basic way from ordinary learning. Learning essentially involves an increase in your certainties; relocation aims at holding to your current epistemic ground.⁵

⁴See, among many, Jeffrey (1983b) and (1983a), and Diaconis & Zabell (1982).
⁵To my knowledge, the idea that tracking time requires the ability to preserve a certain kind of content has been discussed previously only by Gareth Evans: see his discussion of demon-
The notion of *de re* information is central to my account. What is this information? Consider the omniscient god again. Despite her omniscience, her overall cognitive state is in some ways asymmetric with respect to the past and the future. The god—who, remember, is located and lives in time—has first-personal experiences and memories about past events and about objects she has perceived in the past. But she can’t have first-personal experiences and memories about the future. To be sure, she might be able to imagine future events with maximal vividness, as if she were watching a movie on a high-definition screen. But this doesn’t amount to perceiving and having memories of those events. Perception and memory require a causal link that is missing for the case of future objects and events.

For my purposes, *de re* information about an object or an event is just information that is based on cognitive processes requiring a causal link, like perception or memory. For illustration, suppose that at some point the god has witnessed, say, a momentous cosmic explosion. As a result of her perceiving and remembering the explosion, she is able to entertain *de re* information about it: for example, the information that *that explosion* [individuated on the basis of her visual memories] was extremely bright. I claim that, as her stock of experiences and memories increases, the god is able to entertain new information of this sort. The idea that subjects are able to entertain new *de re* beliefs as they gather more perceptual information about the world is going to play an important role in my account.⁶

The notion of *de re* information so characterized is an interesting hybrid of subjective and objective. On the one hand, its existence is clearly tied to a subject’s having a perspective of some sort. By stipulating that the god lives in time, I’m assuming that she has at least a temporal (if not a personal) perspective: at any point, some objects and events are in the past with respect to her, while some others are in the future. Just this perspectival asymmetry grounds the asymmetry in her cognitive state. On the other hand, the possibility of entertaining *de re* information is not tied to the subject’s particular position in time. As long as her memories of the explosion are preserved, the god can entertain the information that *that explosion* was bright. Hence *de re* information enjoys a kind of stability that makes it similar to *de dicto* information.

⁶Notice that I’m claiming that the information gathered via perception and memory enables the subject to identify a certain object *de re*. But I don’t want to make commitments about the nature of the information involved in perception and memory *per se*. All I say is consistent with the idea that this information is not *de re*.
There is an apparent tension between the claim that the god’s stock of *de re* information changes with time and the assumption that the god is omniscient. As time goes by, the god can entertain information that wasn’t available to her before. But then is it true that she can’t learn anything? This question would deserve extended discussion and I don’t have the space to do this here. What matters for me is pointing out at least a sense of ‘learning’ and ‘omniscience’ in which the two claims are not in tension. On the picture that I propose, the god’s acquiring new *de re* beliefs doesn’t amount to her ruling out some metaphysical possibilities that she previously was not in a position of ruling out.⁷ In this sense at least, the god’s new knowledge doesn’t amount to knowledge of new facts. Rather, in rough outline, by acquiring new perceptual information the god acquires new conceptual resources; these conceptual resources allow her a different point of access to the same metaphysical possibilities.

Before proceeding, let me make two qualifications. First, my characterization of *de re* information echoes, in several respects, characterizations of the notion of acquaintance. Acquaintance is a kind of direct epistemic relationship between subject and object that is based on a kind of causal link. The boundaries of acquaintance are blurry, but the clear paradigm cases are just perception and memory.⁸ I’m not interested in settling whether my notion of *de re* information can be perfectly captured via acquaintance. But, for convenience, I will help myself to the notion of acquaintance in my discussion of the *de re*. Second, I characterized a particular notion of *de re* information and am committed to using it as a part of my account. But the basic structure of the account (and, in particular, the formal machinery) could be combined with different notions of the *de re*. Hence what follows should be of interest to you even if you disagree with my conception of *de re* attitudes.

### 3.3.2 An analogy: tracking movement

I begin with an analogy with tracking distances. Suppose you’re traveling on a train between Boston and New York. For simplicity, suppose also that you know all the relevant objective facts about your train ride: the aspect of the landscape, the exact moments of arrival at each station, and so on. In metaphorical terms, you have in your head a fully specific third-personal map of your itinerary. During the trip, you can keep track of the distance covered by your train: you can follow its movement in a way that allows you to know where the train is on your third-personal map. It’s not easy to say exactly how

---

⁷Hence there is a contrast between this picture of *de re* belief and the picture of the *de re* grounded in haecceitic metaphysics. (For an example of the relevant picture of belief, see Stalnaker (2008); for its metaphysical grounds see, among many, Adams (1981).)

⁸The notion of acquaintance is introduced by Russell in his celebrated Russell (1917). Russell imposed very restrictive conditions on acquaintance though. For more up-to-date accounts, see Lewis (1979) and (1983) and Pryor (2004).
CHAPTER 3. COGNITIVE RELOCATION

this tracking capacity works, but this is not important. What matters is, rather, what kind of knowledge it produces.

Here is my suggestion: your distance tracking capacity delivers both de se and de re information. The former part is obvious. Tracking distances involves knowing the distance between your origin point and the point where you are. But there is more. If at time $t_0$ you’ve been tracking the movement of your train, you’re also able to locate this place, i.e. the place you’re currently perceiving, on a third-personal map. This is not de se information, but rather de re information about a point in space.

These two kinds of information are genuinely different. The quickest way to show this exploits the connection to action that characterizes de se beliefs. Suppose that, when you realize you’re 90 miles from Boston, you decide that in exactly 50 more miles you’ll walk up to the restaurant car and buy a beer. You might retain the information that when the train was at that place [individuated on the basis of your memories] you decided to get a beer, but still lose track of distances, and hence become unable to get up and get a beer at the right moment. In this case you retain de re information about the relevant place. At the same time, you’re unable to locate that same place in egocentric coordinates, i.e. in a way that exploits your current position: this is showed just by the fact that some of your dispositions to act get undermined. But then de se and de re information have different preservation conditions, hence they are distinct.⁹

Nevertheless, there is a close connection between the two. Barring skeptical scenarios, at $t_0$ you have the information that you are at distance $d$ from Boston just in case you also have the information that that place [i.e. the place you’re currently perceiving, call it $p_0$] is at distance $d$ from Boston. I capture this by saying that your de re information about $p_0$ is aligned with your de se information at $t_0$. I give a precise characterization of alignment in section 3.4. But the basic idea is simple: a subject’s de re information about a place $p_r$ is aligned with her de se information just in case she’s certain of the distance between $p_r$ (individuated de re) and her current location (individuated de se). Intuitively, there is alignment just in case a subject is able to individuate $p_r$ in egocentric terms: i.e., just in case she can tell how far away $p_r$ is from her current location.¹⁰

Notice that de se and de re information can stay aligned through time. Suppose you keep track of the movement of the train. After 10 miles the de re information that $p_0$ is

⁹Thanks to Dilip Ninan for discussion here.

¹⁰The claim that de se and de re information are connected has been made other times in the literature: cf. for example Stalnaker (2008) and Moss (2011). But there is an important difference between my account and Stalnaker’s and Moss’s. Both those accounts claim that de se information amounts to a kind of de re information. (For Stalnaker’s case, this is just a stipulation built in the model, while Moss defends the claim via an elaborate analogy with descriptive names.) I am not making a similar move here. I accept (or at least, I leave it open) that de se and de re information differ in kind. I claim that there is an epistemic connection between them, so that by coming to know one you will inevitably come to know the other.
3.3. TEMPORAL RELOCATION

90 miles from Boston is aligned with your \textit{de se} information. In fact, you can individuate $p_0$ in egocentric terms: you know that it's 10 miles away from the place where you are. But this connection can also be lost. Suppose that you fall asleep and lose track of where you are. As you wake up, you still retain the information that $p_0$ is 90 miles from Boston (assuming that you can still identify $p_0$, for example by recalling your visual memories of it). But there is no piece of \textit{de se} information that you can univocally link to this piece of \textit{de re} information. In this case, I speak of \textit{misalignment}.

In summary: tracking distances involves manipulating both \textit{de se} and \textit{de re} information about places. The two kinds of information are learned together and have the same epistemic ground. Nevertheless, they are distinct, as is shown by the possibility of misalignment. Now let me show how these ideas can be transposed to time.

3.3.3 Temporal information

Analogously to distances, tracking time involves manipulating both \textit{de se} and \textit{de re} information. Suppose you have dinner at 8 pm on August 31st, 2011. If you keep track of time, after two hours you will have the \textit{de se} information that two hours have passed since 8 pm on August 31st 2011 and the \textit{de re} information that \textit{this moment} is two hours after 8 pm on August 31st 2011. These two pieces of information come together: you cannot get one without the other. They stay aligned as long as you keep track of time and they misalign if you lose track of time. Suppose that, after dinner, you fall asleep on your couch: as you wake up, you might still retain the information that \textit{that time} was the time at which you had dinner, and moreover that it was 8 pm on August 31st 2011. But you're no longer able to link that information to \textit{de se} coordinates.

Let me say more about \textit{de re} knowledge of times. I characterized \textit{de re} attitudes as involving a causal component. This requirement might seem problematic for the case of times, since times are causally inert. Yet we do seem to routinely have \textit{de re} information about time: indeed, it seems that you can always focus on a particular time (perhaps by saying to yourself ‘now!’) and that this is generally sufficient for you to have \textit{de re} attitudes about it.\footnote{To better see this, notice that, on a par with \textit{de re} information about events and objects, we seem to be able to retain \textit{de re} knowledge about time by retaining information in memory. For example, some hours after focusing your attention on a particular time, you might wonder what time \textit{that time} was.} Hence I assume that, in order for you to have \textit{de re} attitudes about time, it’s sufficient that you be acquainted with some event occurring at that time. This allows you to individuate the relevant time on the basis of information involved in perception or memory, as the time at which \textit{those events} [individuated on the basis of perception or memory] have taken place.

One constraint follows from this: you cannot have \textit{de re} information of times at
which you haven’t been located. Hence you don’t normally have de re information future times, as well as times that precede your birth. The reason, of course, is that you cannot perceive or have memories about events happening in the future or before your birth. One consequence of this is that, as time goes by, you’re going to acquire de re information about new times. This fact is going to play a big role in my account.

Now let me illustrate how how the account works. I examine in detail the example in the following scenario:

*The Prisoner.* On Monday, you are imprisoned and left by yourself in an isolation cell. For two days, you manage to keep track of how much time has elapsed since your imprisonment: as a result, on Wednesday, you are certain that two days have passed.

For clarity, I’m going to use demonstrative phrases with consecutive numerical indices to pick out your de re ways of identifying days. In this scenario, you identify Monday, the day of your imprisonment, as that\(_1\) day; Tuesday, the next day you’re acquainted with, as that\(_2\) day; and Wednesday, the day you’re acquainted with after that, as that\(_3\) day. Given that you’ve been tracking time until Wednesday, at that point you know both the de re proposition that that\(_3\) day is two days later than that\(_1\) day and the de se proposition that that\(_1\) day was two days ago. Given that you’re certain about the distance between the present time and that\(_1\) day, your de re knowledge about Monday is aligned with your de se temporal knowledge.

And now, consider the following extension of the example:

*The Groggy Prisoner.* You go to sleep on Wednesday night. Upon waking up in your cell and feeling unusually disoriented and groggy, you wonder how much time has elapsed since you went to sleep. Perhaps it’s been only the night, perhaps a night and a full day.

Your losing track of time brings about de se uncertainty: you’re not sure whether today is one or two days after that\(_3\) day. But, as in the movement case, it also produces de re uncertainty. Upon waking up, you gain new de re beliefs about that\(_4\) day (and, as usual, this de re information is aligned with your de se information). But you’re uncertain how far that\(_4\) day is from that\(_3\) day: perhaps one day, perhaps two.

In summary, upon waking up your belief state changes in two important ways. First, you acquire de re beliefs about a new day. Second, you acquire a new kind of de re uncertainty. It’s useful to represent this change via diagrams. Insofar as your temporal information is concerned, there is only one doxastic alternative compatible with what you believe on Wednesday:
(I use labels of the form ‘\(d_i\)’ as shorthand for the demonstrative phrase of the form \(\text{that}_i \text{ day}\).) The arrow represents \(de \ se\) information: it shows that you take \(\text{that}_3 \text{ day}\) as the day you’re located at. And these are the two alternatives compatible with your beliefs as you wake up:

In both cases the arrow points on \(\text{that}_4 \text{ day}\), showing that your beliefs about the latter are aligned with your \(de \ se\) temporal beliefs. But there is no alignment for your other \(de \ re\) beliefs about time: for example, you’re now uncertain whether \(d_3\) was one or two days ago. In addition to the \(de \ re–de \ se\) link being broken, you have acquired some \(de \ re\) uncertainty. For example, you’re uncertain whether \(\text{that}_4 \text{ day}\) is one or two days after
that day. As the diagram shows, this uncertainty essentially concerns the structure of the de re time sequence. You’re ignorant about the distance between certain times, individuated de re.

For clarity, let me emphasize that this picture doesn’t come with new ontological burdens. The dots appearing in the diagrams don’t model a new category of times infesting our ontology besides ordinary times. Rather, they represent ways of individuating times. I explain in the next section how these intuitive notions should be captured formally. For the moment, I just want to highlight that de re attitudes about times don’t require a new kind of times in our ontology (much in the same way as de se attitudes and centered content don’t demand for new kinds of times and individuals).

I also want to stress two points that will be important later on. First, de re uncertainty is uncertainty about some ‘objective’ information, not essentially tied to your position within the world. As time goes by, you will keep being uncertain about the very same propositions. This is significant, since it shows that we can acquire a kind of third-personal, non-self-locating uncertainty just via relocation (or better, by failing to relocate effectively). Second, even though you acquire a kind of uncertainty, you fully retain your previous attitudes towards de re propositions. Your uncertainty concerns de re propositions about that day: i.e., propositions that become available to you because you’ve moved to a new position in time.

3.3.4 Relocation

The foregoing shows what information is involved in tracking time. But, so far, I’ve said nothing about the process that determines the evolution of subjects’ belief states. There are three main questions to address: first, what evidence works as input to this process; second, how we get from this evidence to an updated belief state; and third, how this update process interacts with ordinary learning.

The answer to the first and the third question is straightforward. As for the former: your time-tracking capacity constantly provides you with the self-locating information that a certain amount of time has passed from a previous moment. If you’re a perfect time-tracker, this capacity will invariably deliver perfectly precise information: at each moment, you will know exactly how much time has gone by. If you’re like most other agents, the information you get will fall within a more or less wide range. Notice that the deliverings of your time-tracking capacity will be susceptible to adjustment on the basis of other information you might have: time might pass more slowly if you’re bored, or tired, or you’ve taken drugs. So the evidence you use for updating is what your time tracking capacity tells you, conditional on specific information you might have about the deliverings of that capacity under your circumstances.¹²

¹²For simplicity, I will ignore this complication in what follows. Incidentally, notice that
As for the interaction between relocation and genuine learning, the answer is simple. You must first relocate, and then update on the genuine evidence that you’ve learned. If you did otherwise, your resulting beliefs would be dramatically off the mark. Suppose that, a few minutes after your awakening in prison, the guard proudly shows you his new iPad. If you updated on the self-locating information the guard shows me an iPad now before relocating, you would be led astray. After relocating, you would find yourself believing that the guard showed you his iPad back then, at the moment of your awakening. So it seems clear that relocation should precede all update. Your keeping track of where you are in the world is a precondition for your being able to learn accurate information from the environment.

The remaining question concerns the process of relocation proper. How do we get from the evidence that some time has passed to an updated belief state? I answer in two steps. First, I consider the idealized case in which you’re a perfect time-tracker. Then, I generalize to the more mundane case in which you only get an approximate sense of the passing of time.

Borrowing a metaphor from the metaphysics of time, think of the self-locating component of your temporal beliefs as a kind of spotlight that is being shined on a point in a sequence of objectively individuated times. When you get evidence that some time has passed, you move the spotlight forward along that sequence. Take the prison case again. There you start out with the spotlight shining on Monday, the day of your imprisonment:

As you get the information that some time has passed, the spotlight moves forward. After a day, it will be on Tuesday:

Out of metaphor, the claim is that your credences get translated forward along the temporal dimension. You start out with a credence distribution \( C_r \) over possibilities. (I’ll someone might object to the use of the term ‘evidence’ to denote the information that works as input to cognitive relocation. Evidence is something we learn from; here I am claiming precisely that relocation is not a kind of learning. I use ‘evidence’ just for simplicity: if you have qualms with it, feel free to use a different label.)
say more soon about the possibilities in question.) Once you learn that a certain amount of time \( i \) has passed, your credence in each possibility \( p \) is shifted to the minimally different possibility \( p' \) whose \textit{de se} temporal coordinate is shifted by an amount \( i \). The result is a new credence distribution \( Cr' \) that differs from \( Cr \) in two ways. First, the credences you assign to self-locating propositions differ, since the agent believes that his position in time has changed. Second, you’re able to ‘see’ new \textit{de re} propositions (in the case at hand, propositions about \textit{that} \( 2 \) day) and have credences in them.

Now assume you’re an imperfect time-tracker. In this case, your evidence is weaker: you know that some amount of time has passed, but you’re not sure how much. Here I consider a toy case where you’re undecided between just two possibilities; the generalization to more complex (finite) cases is elementary.\(^{13}\) Take once more the prison scenario and consider what happens as you wake up on \textit{that} \( 4 \) day, i.e. after you lose track of time. You’re unsure whether one or two days have passed since you went to sleep: for simplicity, suppose that you assign .5 credence to each of the two alternatives. In this case, there isn’t a unique point to move the spotlight to. The way to proceed, then, is to shift your credences piecemeal: for each possibility in your belief state, half of its credence is moved one day forward, and the remaining half is moved two days forward. Let me explain.

As an example of imperfect time-tracker, take the prisoner who wakes up groggy and confused after going to sleep on Wednesday night. Her new belief state includes two different doxastic alternatives. One of them results from shifting the spotlight by one day, the other from shifting it by two days:

\[
\begin{align*}
\text{d}_1 & \quad \text{d}_2 & \quad \text{d}_3 & \quad \text{d}_4 \\
\text{d}_1 & \quad \text{d}_2 & \quad \text{d}_3 & \quad \text{d}_4
\end{align*}
\]

The prisoner’s credences undergo the same shift as in the perfect tracking case, only piecemeal. Take an arbitrary possibility \( p \) in her belief state. She should move half of its credence to the minimally different possibility \( p' \) whose \textit{de se} temporal coordinate is shifted one day forward; she should move the remaining half to the minimally different possibility \( p'' \) whose \textit{de se} temporal coordinate is shifted two days forward.

The general idea should be clear. The credence you assign to each possibility within your doxastic state is shifted forward along the temporal dimension. The import of the shift, as well as the amount of credence shifted to different destinations, is determined by the evidence you get from your time-tracking capacity. This is how you relocate.

\(^{13}\)Problem: since times form a dense ordering, ordinary cases of relocation actually do involve an infinite amount of possibilities in your evidence. Distributing credences over an infinite sets of propositions introduces complications that are orthogonal to my main point here.
I have illustrated the process by talking about possibilities. It's useful to point out what effects this has on the resulting credences in propositions.

- First, as is obvious, credences in *de se* propositions are going to be overhauled. If you start assigning credence 1 to the proposition that it's Wednesday and then get the evidence that a day or more has passed, your credence will be fully moved over to other propositions.

- Second, imperfect time-trackers will spread their credences over several new *de re* propositions, i.e. propositions they couldn't entertain before relocating. For example, the groggy prisoner ends up with .5 credence in the proposition that *that* 4 day is one day after *that* 3 day, and .5 credence in the proposition that *that* 4 day is two days after *that* 3 day.

- But, third, credences in other *de re* propositions are unaffected. For example, the groggy prisoner's credences about the distance between *that* 1 day and *that* 3 day remain untouched by the update (i.e., he keeps being certain that the they are two days apart).

These points will be important in my account of the epistemic significance of relocation.

### 3.4 The formal framework

The aim of this section is to formalize the informal account of relocation that I sketched in section 3.3. I first introduce a general framework for modeling content. Then I propose an update rule that captures the functioning of relocation.

#### 3.4.1 Sequenced worlds

There is controversy about how to model *de re* attitudes in a possible worlds framework. Descriptivists (for example, Lewis (1979) and (1983)) analyze *de re* content in terms of *de dicto* and *de se* information. Non-descriptivists (for example, Ninan (2008)) claim that *de re* content should rather be taken as primitive. This difference is significant, since it produces differences in our final model of content. Descriptivists can stick to centered worlds, while non-descriptivists must use richer entities. Nothing in my account of relocation commits me to either of these views. But I choose to specify my model in terms of a non-descriptivist framework. The reason is that the non-descriptivist framework exhibits explicitly what *de re* beliefs a subject has, thus allowing for a clearer illustration.
of the theory.¹⁴

Following Ninan (2008), I start from Lewis’s centered worlds framework and add to it an extra component which models de re attitudes. Lewis models a subject’s belief state as a set of triples of worlds, times, and individuals which are doxastic alternatives of the world, the time, and the individual the subject takes herself to be located at. I use sets of quadruples: the added element is a temporal sequence, i.e. a sequence of times that are alternatives of times the subject has de re beliefs about. So the elements of a subject’s belief state have the form:

\[ \langle w_i, t_j, x_k, \langle t_1, t_m, \ldots, t_j \rangle \rangle \]

Notice two things. First, the length of the temporal sequence in a subject’s doxastic alternatives is determined by the number of times she has de re beliefs about. Second, in virtue of the epistemic connection between de se and de re beliefs about time, in each world the last element of the temporal sequence is invariably identical to the de se temporal alternative. This captures the fact that a subject always believes that what she picks out as this time and what she picks out as now coincide.

For illustration, take again your belief state in the prison scenario. You know you entered the prison on Monday; when you lose track of time, you’re unsure whether three or four days have gone by. For simplicity, assume you have no further uncertainty and that the only times you have de re beliefs about are the days you’ve spent in prison. Then there are two sequenced worlds in your belief state:

\[ \langle w_\text{\#1}, \text{you, Thursday, } \langle \text{Monday, Tuesday, Wednesday, Thursday} \rangle \rangle \]
\[ \langle w_\text{\#2}, \text{you, Friday, } \langle \text{Monday, Tuesday, Wednesday, Friday} \rangle \rangle \]

Finally, with the formal apparatus in place, I can define a precise notion of alignment:

\[ S’s \text{ de re beliefs about the } i\text{-th point in his de re sequence are aligned with her de se beliefs just in case, for all sequenced worlds } \langle w_i, t_j, x_k, \langle t_1, t_m, \ldots, t_i, \ldots, t_j \rangle \rangle \text{ in } S’s \text{ belief state, the time interval } |t_j-t_i| \text{ is constant.} \]

It should be easy to check that this definition yields the result that, in the groggy prisoner case, your de re beliefs about the present day are aligned with your de se beliefs, but no beliefs about other days are.

¹⁴There is also another advantage to this choice: the sequenced world framework can be easily reinterpreted as a reductionist framework, provided that we add an extra condition. This condition is that no belief state can contain two sequenced worlds \( \langle w_i, t_j, x_k, \langle t_1, t_m, \ldots, t_i, \ldots, t_j \rangle \rangle \) and \( \langle w_i, t_l, x_j, \langle \ldots, t_k, \ldots \rangle \rangle \), i.e. sequenced worlds that agree in the centered world coordinate but disagree in the sequence coordinate. (Basically, the idea is that for each belief state we should be able to specify a function from a centered world to the object in the \( i \)-th position in the sequence.)
3.4. THE FORMAL FRAMEWORK

3.4.2 Relocation as imaging

The foregoing determines the basic elements of the model. Now I turn to how credences are distributed and shifted over these elements.

Here is an overview of the proposal. Relocation exploits a version of imaging, an updating procedure which was first defined by Lewis (1976). Conditionalization, recall, can be thought as a two-step process: first, you erase credences in possibilities that are incompatible with your new evidence; then, you renormalize the remaining credences so that they sum up to 1. Imaging works in an altogether different way. No credences are ever erased. Rather, update by imaging essentially works by shifting around existing credences. We first define a closeness relation between the worlds in our algebra. This relation is such that, for each world \( w \) and proposition \( p \), there is a \( p \)-world \( w' \) (or, a set of \( p \)-worlds \( W' \)) that is (are) closest to \( w \). When we image on a proposition \( p \), we transfer the credence in each world to the closest \( p \)-world or worlds. The updated credence distribution \( Cr' \) obtained in this way is called the image of \( Cr \) on \( p \). In particular, I use a version of imaging that I call ‘weighted imaging’ and that allows for multiple similarity relations to be considered at the same time.

Let me flesh out the details. The first step is defining the relevant closeness relationship. My suggestion is very simple: for the purposes of temporal relocation, closeness is just given by temporal ordering. Take an arbitrary sequenced world

\[
\langle w, t_n, x, \langle t_1, t_2, t_3, \ldots t_n \rangle \rangle
\]

The closest world verifying the proposition that an amount of time \( i \) has passed is

\[
\langle w, t_n + i, x, \langle t_1, t_2, t_3, \ldots t_n, t_n + i \rangle \rangle
\]

In other words, the closeness relationship takes each sequenced world into a sequenced world that has (a) a time coordinate shifted forward and (b) an extra member in the \( de re \) time sequence. The world parameter, the individual parameters, and all other members of the temporal sequence are unaffected.

This settles immediately how you should update in cases of perfect time-tracking. Let \( w_p \) be the closest centered world to \( w \) which verifies \( p \) (where \( p \) is a proposition specifying that a certain amount of time \( i \) has passed since the subject last performed an update). Given a credence distribution \( Cr \), your new credence \( Cr' \) in a sequenced world \( w' \) upon learning \( p \) should be:

\[
Cr'(w') = \sum_w Cr(w) \times \begin{cases} 
1 & \text{if } w_p \text{ is } w' \\
0 & \text{otherwise}
\end{cases}
\]

Here is an algorithm summarizing the functioning of (1):
For each world \( w' \) in your algebra, determine whether that world is the closest \( p \)-world to some other world.

- If not, assign it credence 0.

- If yes, take the sum of your previous credence in all worlds \( w \) such that \( w' \) is the \( p \)-closest world to them, and set that as your new credence in \( w' \).

This procedure, which is just the one defined by Lewis, can’t be applied straightforwardly to cases of imperfect time-tracking. The problem is the assumption that, for each world, there is a unique closest world which verifies your evidence. If you’re an imperfect time-tracker, there will be no such closest world.

The solution consists in adopting a generalized imaging procedure, which I call ‘weighted imaging’. Weighted imaging is a generalization of imaging. In essence, rather than telling you to shift all your credence in a world \( w \) to a new world \( w' \), it tells you to shift your credence in \( w \) piecemeal to a set of worlds \( w_1, \ldots, w_n \), according to given probabilistic weights. Weighted imaging is a kind of general imaging, a family of update procedures introduced and characterized in detail by Gärdenfors (1982) and (1988).

One limit of the current proposal is that I only define weighted imaging for the finite case, i.e. for the case in which the evidence you get from your time-tracking capacity singles out a finite number of options. I.e., I assume that the evidence you use to relocate will be a finitary disjunction of the form

- An exact amount of time \( i_1 \) has passed since you last relocated,
- or an exact amount of time \( i_2 \) has passed since you last relocated,
- or \( \ldots \),
- or an exact amount of time \( i_n \) has passed since you last relocated.

This is obviously an idealization, since time forms a dense ordering. But it will do to illustrate the gist of the proposal.

Start by considering how to model the evidence you get from your capacity of tracking time. Suppose that, as you wake up, your tracking capacity suggests a .5 credence in the proposition that one day has passed since you last fell asleep, and a .5 credence in the proposition that two days have passed. Thus your evidence can be represented as a set of two mutually exclusive and exhaustive propositions (call them ‘ONE DAY’ and ‘TWO DAYS’) and by probabilistic weights that you attach to each of them. We can represent this as a set of proposition-probability pairs; for simplicity, call it a weighted partition:

\[
\{ \{ \text{ONE DAY}, .5 \}, \{ \text{TWO DAYS}, .5 \} \}
\]

I assume that the your evidence takes the form of a weighted partition. Hence the input to the update procedure is something of the form:
3.4. THE FORMAL FRAMEWORK

\{ \langle E_1, p_1 \rangle, \langle E_2, p_2 \rangle, \ldots, \langle E_n, p_n \rangle \}

At this point, it’s easy to see how to generalize standard imaging to weighted imaging. Take a sequenced world \( w \): we shift a fraction \( p_n \) of the previous credence in \( w \) to the \( E_n \)-closest world to \( w \), for all \( \langle E_n, p_n \rangle \) in the evidence set. Here is a formal statement of the rule:

\[
Cr'(w') = \sum_w Cr(w) \times \begin{cases} 
p_1 & \text{if } w_{E_1} \text{ is } w' \\
0 & \text{otherwise} \\
end{cases} + \ldots + \sum_w Cr(w) \times \begin{cases} 
p_n & \text{if } w_{E_n} \text{ is } w' \\
0 & \text{otherwise} \\
end{cases}
\]

Here is the rule restated in algorithm form:

- For each world \( w' \) in your algebra, determine whether that world is the closest \( E_n \)-world to some other world, for some \( n \).
- If not, assign it credence 0.
- If yes, for each world \( w \) such that \( w' \) is the \( E_n \)-closest world to it, move a fraction \( p_n \) of your credence in \( w \) to \( w' \).
- The probability of \( w' \) is the sum of all the credences that have been moved to it in this way.

In essence, while lewisian imaging employed a unique closeness ordering, weighted imaging makes use of a weighted mixture of orderings. Each world shifts determinate fractions of its credence onto each of the worlds to which it is closest, according to a certain ordering.

This is the basic update rule which models temporal relocation. As I mentioned in section 3.3, this rule is meant to be combined with, rather than substitute, standard conditionalization (which I take to be defined on sequenced worlds). Rational agents first update by relocation, thus shifting their epistemic state to what they take to be their new coordinates. At this point, they are ready to take in any news they might get in their new position, which they do by conditionalizing.

In closing, let me make a few remarks about the evidence required by weighted imaging. One worry is that weighted partitions are too fine-grained and too complex. The time-tracking capacity of actual agents only yields very coarse information and cannot be taken as providing a partition of options with precise probabilities attached to them. The reply, quite simply, is that weighted partitions are obviously an idealization, on a par with idealization standardly in place in various versions of the Bayesian apparatus. There are obvious similarities with the case of Jeffrey conditionalization, which requires direct learning of probabilities from experience.¹⁵

¹⁵For discussions of this point, see Christensen (1992).
More seriously, one might worry that requiring such a complex notion of evidence will leave some important epistemological questions unanswered. Weighted imaging simply assumes that subjects assign some credence to different hypotheses about how much time has passed. But this says nothing about a substantial question: are there any rational constraints on assigning credences to these hypotheses? And if so, what are they? I agree that this is an important limitation. However, it is a limitation that we should fully expect. My account aims at explaining a specific phenomenon: intertemporal update. How the inputs to cognitive relocation are obtained, and what constraints hold for them, are questions outside my ken. My concern here is how your mental state should change, once you have the relevant evidence; this question can be answered without settling questions concerning the inputs to this procedure.

3.5 The relevance of de se information

3.5.1 The question of epistemic relevance

Sections 3.3 and 3.4 conclude my account of relocation. At this point, I turn to exploring some philosophical consequences.

Much recent debate has focused on the epistemic relevance of de se information. Suppose that a subject gets evidence that only rules out de se possibilities—i.e., evidence that can only tell her with certainty something about where she is located within the world, and not about what the world is like. Does this ever warrant a change in her credences in de dicto propositions? In other words: can acquiring purely de se information ever have an impact on one’s credences towards de dicto propositions?

As is often pointed out, it’s easy to elicit intuitions favoring a ‘no’ answer. Consider the following example, due to Titelbaum (2008):

Imagine a situation in which an agent is certain of every experience she’s going to have between two times. She’s watching a movie, say, that she’s seen a hundred times and has thoroughly memorized; she’s at home alone on her well-worn couch in the dark; the phone is unplugged; et cetera. It’s not quite right to say that the agent learns nothing as the movie progresses—after all, she learns that some time has passed. But all she learns is self-locating information; she has no basis for inferring new non-self-locating conclusions. (2008, page 556)

In this case, it’s natural to think that the agent should not change any of her credences in de dicto propositions. All she gets is information about her change of position in
3.5. THE RELEVANCE OF DE SE INFORMATION

the world. But (on a par with the omniscient god\textsuperscript{16}) this change is fully expected. The subject gets no new information. Hence she has no reason to alter her de dicto degrees of belief.

The movie-watcher example has been taken to support a very general principle about the interaction of de se and de dicto information. Here is one formulation of this principle, due again to Titelbaum:

\textit{Relevance-Limiting Thesis (RLT):} it is never rational for an agent who learns only self-locating information to respond by altering a non-self-locating degree of belief.

Much of the debate about the relevance of de se information has focused on (RLT) or equivalent principles. Now, is (RLT) true or false in the light of my account? Neither; the question is confused. If I’m right, (RLT) is infected by the conflation of two different cognitive processes, namely relocation and genuine learning. This conflation is harmful because, once we get to disambiguate (RLT), one of the resulting claims is true and the other false. Let me explain.

The source of ambiguity in (RLT) is the expression 'learning self-locating evidence'. If we take it to denote relocation, (RLT) is turned into the following:

\textit{Relevance-Limiting Thesis\textsubscript{R} (RLT\textsubscript{R}):} it is never rational for an agent who updates her belief state exclusively via cognitive relocation to alter a non-self-locating degree of belief.

It should be easy to see that RLT\textsubscript{R} is true (with a qualification) on the account I’ve given. Non-self-locating degrees of belief, recall, are degrees of belief in de dicto and de re propositions. Now, in section 3.3 I stressed that relocation has no impact on (a) the subject’s credences in de dicto propositions, and (b) her credences in the de re propositions she can entertain before the update. To be sure, relocation \textit{does} have an impact on the subject’s overall de re credences: by moving to a new time, one gets to entertain beliefs and assign credences to new de re propositions. Just in this way she can gain a kind of uncertainty. But this impact concerns only propositions in which the subject had no previously defined credences. Her old credences in de dicto and de re propositions remain unchanged.

This is a welcome consequence, as it yields a number of intuitive verdicts. For example, it does predict that Titelbaum’s movie-watcher should not alter her de dicto credences, since the only cognitive change she undergoes is just relocation. More in general, it vindicates the intuitive idea that you shouldn’t change your credences about the

\textsuperscript{16}Indeed, there is a difference: by stipulation the god is a perfect time-tracker, while Titelbaum’s movie-watcher might still lose track of time. (For example, if the movie contains qualitatively identical sequences at different points.) But this is irrelevant for current purposes.
world just because you change your position within it. (Among other things, this intuition has played an important role in discussions of the so-called Sleeping Beauty case; more on this shortly.)

Things are different for the other disambiguation of \((\text{RLT})\). In this case, we take the expression ‘learning self-locating evidence’ to denote genuine learning, i.e. the kind of update on evidence that is normally modeled via conditionalization. The resulting principle is:

**Relevance-Limiting Thesis** \(_\text{(RLT)}\): it is never rational for an agent who genuinely learns only self-locating information to alter a non-self-locating degree of belief.

Nothing in my account entails \((\text{RLT})\) or its denial. But it’s very easy to see that it’s false, independently of one’s specific story about \(\text{de se}\) update. Here is an example that should make the point:

*The Dali moustache.* A malicious genie will partially erase your memory on Monday and Tuesday at 3:23 pm: on both occasions, you’ll forget anything that happened after 11:59 pm on Sunday and will be left wondering whether it’s Monday or Tuesday. To make the prank more interesting, he has decided to flip a fair coin. If tails comes up, a Salvador Dali-style moustache will appear on your face both days as the memory erasure takes place. If heads comes up, the moustache will appear only on Monday. As your memory gets erased and you start wondering what time it is, you catch a glance of yourself in the mirror. To your dismay, an upward-pointing moustache now lies on your upper lip.

I take the following to be uncontroversial: discovering that you have a moustache involves genuinely learning information and hence it can be effectively modeled by conditionalization. But then getting the self-locating information that you now have a moustache will lower your credence in heads and correspondingly raise your credence in tails. This is intuitive: discovering that you now have a moustache should provide some confirmation for the hypothesis that the coin landed tails, simply because a similar discovery will happen twice as often in tails-worlds.\(^\text{17}\)

This provides a general answer to the question of relevance. At this point, I turn to discussing a specific example that has been the subject of much debate, the Sleeping Beauty problem. As I hope to show, my framework allows for an elegant solution to the problem, which manages to vindicate the main intuitions behind both positions in the debate.

\(^{17}\)Dorr (2002) gives an example that is structurally analogous. (Dorr’s example, in its turn, is a variant on the Sleeping Beauty puzzle.) I take it that at least the great majority of participants in the debate agree that Dorr’s case is uncontroversial too.
3.5. THE RELEVANCE OF DE SE INFORMATION

3.5.2 The Sleeping Beauty problem

Here is a concise summary of the Sleeping Beauty scenario, as it is introduced by Elga (2000)¹⁸:

Sleeping Beauty. Sleeping Beauty is told that she will undergo an epistemology experiment. Scientists will put her to sleep on Sunday night. Then they will wake her up either only on Monday or on both Monday and Tuesday, depending on the toss of a fair coin: if heads, she will be woken up just on Monday; if tails, both days. If the coin has landed tails, on Monday night scientists will administer Beauty a drug that will erase all memories of her awakening, so that she will have no recollections of it upon waking up again on Tuesday.

The controversial question is: as she wakes up on Monday, what credence should Beauty assign to the de dicto proposition that the coin landed heads?

With few exceptions, participants in the debate are split between halfers (the rational credence for Beauty to have in the coin landing heads is 1/2) and thirders (the rational credence for Beauty to have in the coin landing heads is 1/3).¹⁹ For current purposes, however, I frame the debate in slightly different terms. I distinguish between a relevantist and an antirelevantist position. Here is how I characterize relevantism:

Relevantism: the cognitive changes that Beauty undergoes between Sunday night and Monday morning are epistemically relevant to her credences in (at least some) non-self-locating propositions.

Antirelevantism is the denial of relevantism. Antirelevantism is a natural generalization of the halfer position, since on standard assumptions Beauty starts out with .5 credence in heads. Conversely, relevantism is a generalization of the thirder position. My framework is not sufficient to fix a precise numerical value for Beauty’s credence (since, without further assumptions, it doesn’t say what credences Beauty should assign to the propositions in her weighted partition when she relocates). But it is sufficient to decide between relevantism and antirelevantism. The verdict is that relevantism is true, even though some important aspects of the antirelevantist position are also vindicated. One consequence is that (given standard assumptions about Beauty’s starting credence in heads) the halfer position also fails. Let me explain how.

A first observation about the Sleeping Beauty scenario is that Beauty loses track of time between Sunday and Monday. Hence both her de se and de re uncertainty increases.

¹⁸The ancestry of the Sleeping Beauty problem traces back to the game theory literature: see Piccione & Rubinstein (1997).
CHAPTER 3. COGNITIVE RELOCATION

On Sunday night, all her relevant *de re* beliefs about time are fully aligned with her *de se* beliefs. Since on Sunday she has full *de se* information, the alternatives within her doxastic state only differ in the world component (one is a tails-world, the other a heads-world).²⁰

\[ \langle w_H, \text{Sunday, SB, } \{ \text{Friday, Saturday, Sunday} \rangle \]  
\[ \langle w_T, \text{Sunday, SB, } \{ \text{Friday, Saturday, Sunday} \rangle \]  

(For simplicity, I assume that Beauty has *de re* beliefs only about the current day and the two previous days.) Between Sunday and Monday, Beauty loses track of time. To be sure, there is a twist with respect to ordinary cases where a subject loses track of time. Beauty's losing track of time is not due to a failure of her time-tracking capacity: for all is said in the description of the scenario, this capacity might be working perfectly. The problem is, rather, that all evidence Beauty gathers via that capacity is made unreliable by the possibility of memory erasure. There are several possible predicaments where Beauty gets the evidence she does; yet in some of them it's Monday, in others Tuesday. Hence (as virtually everyone in the debate agrees) Beauty should lend some credence both to the proposition that one day has passed since she fell asleep and to the proposition that two days have passed.²¹

Since Beauty can't relocate effectively, she acquires *de re* uncertainty. The mechanics and the result of this process should be clear by now. After relocating, her information about time will involve two main alternatives. The day Beauty identifies demonstratively as *this day* may be one or two days apart from the day she identifies (as *that day*) on the basis of her last waking memories:

²⁰Some complications are tied to the fact that, when Beauty is awake on Sunday night, the coin toss still has to take place. Thus (unless we accept determinism) \( w_H \) and \( w_T \) should not be thought of as maximal ways the world is at the time of Beauty's thinking, but rather as (roughly) maximal world histories (see, among many, Thomason & Gupta (1980)). These complications seem irrelevant to my argument.

²¹To my knowledge, Hawley (2011) is the only one to make the heroic attempt to deny this. Notice that, even though philosophers agree that Beauty should lend *some* credence to each proposition, there is wide disagreement about how much credence she should have in each.
(I use ‘F’, ‘SA’, ‘SU’, and ‘NEW’ as shorthand for Beauty’s demonstrative ways of picking out, respectively, Friday, Saturday, Sunday, and the day of her awakening.) This kind of uncertainty will be combined with Beauty’s uncertainty about the outcome of the coin toss. Each of the two sequenced worlds previously compatible with her beliefs will shift part of its credence to a sequenced world where one day has passed since Sunday, and the remaining part to a sequenced world where two days have passed. As a result, after relocation there are four alternatives compatible with Beauty’s belief state:

\[
\begin{align*}
(\text{H}, \text{Monday}, \text{SB},\langle \text{Friday, Saturday, Sunday, Monday}\rangle) \\
(\text{H}, \text{Tuesday}, \text{SB},\langle \text{Friday, Saturday, Sunday, Tuesday}\rangle) \\
(\text{T}, \text{Monday}, \text{SB},\langle \text{Friday, Saturday, Sunday, Monday}\rangle) \\
(\text{T}, \text{Tuesday}, \text{SB},\langle \text{Friday, Saturday, Sunday, Tuesday}\rangle)
\end{align*}
\]

These alternatives represent the way Beauty thinks of her predicament after she has relocated but before she has updated on any new evidence. All that has changed with respect to the Sunday belief state is that Beauty believes that one or two days have passed since Sunday.

A further change takes place after relocation. At this point Beauty updates on what she learns upon waking up. What does she learn? In short, that she is not located at \( (\text{H}, \text{Tuesday}, \text{SB},\langle \text{Friday, Saturday, Sunday, Tuesday}\rangle) \). At that sequenced world Beauty is not awake: since she knows that she is in fact awake, she can confidently rule it out.²²

²²Ignoring skeptical possibilities where she can’t be sure that she’s awake. This case is not usually discussed in the Sleeping Beauty literature, though it introduces interesting (and challenging) complications.
As a result, at the end of her overall update, Beauty only has three doxastic alternatives left:

\[
\begin{align*}
(w_H, \text{Monday, SB, } \{\text{Friday, Saturday, Sunday, Monday}\}) \\
(w_T, \text{Monday, SB, } \{\text{Friday, Saturday, Sunday, Monday}\}) \\
(w_T, \text{Tuesday, SB, } \{\text{Friday, Saturday, Sunday, Tuesday}\})
\end{align*}
\]

In short, Sleeping Beauty’s overall update process involves first an increase and then a decrease in her uncertainty about non-self-locating propositions. By relocating, Beauty acquires uncertainty, hence the alternatives compatible with her belief state pass from two to four. By successively conditionalizing on her evidence, Beauty rules out one of these four alternatives.

The foregoing is not enough to give a full solution to the Sleeping Beauty problem, since I’ve said nothing about what Beauty’s eventual credence in heads should be. (I discuss the point further, though still not conclusively, in the appendix.) But, first, it is enough to establish relevantism. Whatever her final credence in heads, Beauty undergoes cognitive operations that have an impact on it. At least in this sense, her update process is relevant to her credences in \textit{de dicto} propositions. Moreover, it rules out halving. Assume that, on Sunday, Beauty starts out with credence .5 in heads. Then we can show that, if the propositions that one day has passed and that two days have passed
3.5. THE RELEVANCE OF DE SE INFORMATION

both receive non-zero credence, the probability of heads is smaller than .5. Finally, the account naturally shows how the changes in Beauty’s belief state can lead to a change in her de dicto credences. Antirelevantists often insist that it’s hard to see why Beauty’s evidence should be at all relevant to her de dicto credences. In support of this, they invoke some version of Titelbaum’s Relevance-Limiting Thesis: mere changes within one’s position in the world should not affect one’s credence about the world. I have showed how we can both stick to a version of the Relevance-Limiting Thesis and embrace relevantism about Sleeping Beauty. It’s true that subjects shouldn’t change their non-self-locating credences just because their position in the world changes. But, as Beauty loses track of time, she acquires a new kind of uncertainty: this puts her in the position of genuinely learning something as she wakes up. This is what is responsible for the change in her credences.

One interesting moral of the foregoing is that both sides of the Sleeping Beauty debate turn out to be partly right. Antirelevantists are right that one’s credences in non-self-locating information cannot be affected by a mere change in one’s position in the world. Relevantists are right that the cognitive processes Beauty undergoes are

²³Here is a sketch of a proof. We’re given that

\[
Pr_{\text{Sun}}(\text{Heads}) = Pr_{\text{Sun}}(\text{Tails}) = .5
\]

By the appropriate version of the Relevance-Limiting Thesis (RTL), on Monday after relocation (say at time MON (r)) we still have

\[
Pr_{\text{Mon}}(a)(\text{Heads}) = Pr_{\text{Mon}}(a)(\text{Tails}) = .5
\]

Now, by assumption, Beauty assigns non-zero credence both to the proposition that one day has passed since Sunday, and to the proposition that two days have passed. It follows that:

\[
Pr_{\text{Mon}}(a)(\text{Heads & Tuesday}) = \varepsilon
\]

where \(0 < \varepsilon < 1\). Hence we get that the credence Beauty should have once she updates on what she learns upon waking up (time MON (r+1)) is

\[
Pr_{\text{Mon}}(a+1)(\text{Heads}) = Pr_{\text{Mon}}(a)(\text{Heads}/\neg(\text{Heads & Tuesday})) =
\]

\[
\frac{Pr_{\text{Mon}}(a)(\text{Heads} \land \neg(\text{Heads & Tuesday}))}{Pr_{\text{Mon}}(a)(\neg(\text{Heads & Tuesday}))} = \frac{Pr_{\text{Mon}}(a)(\text{Heads & Monday})}{Pr_{\text{Mon}}(a)(\neg(\text{Heads & Tuesday}))}
\]

\[
\frac{0.5 - \varepsilon}{1 - \varepsilon} < 0.5
\]

(Thanks to Bob Stalnaker for discussion about this point.)

²⁴This claim is made by, among others, Lewis (2001), Halpern (2005), Meacham (2008), Hawley (2011).
epistemically relevant to her non-self-locating credences. I regard it as a virtue of my account that it can vindicate both sides to this extent.

3.6 Update and communication

3.6.1 The problem of content and the problem of calibration

There is an obvious analogy between intertemporal update and the communication of de se attitudes. Intertemporal update involves transmission of information between different temporal slices of the same agent; communication involves transmission of information between different agents at the same time. But there are also important differences. In this section, I highlight the latter and argue that communication of de se attitudes involves no analog of relocation. My purpose is twofold. First, I want to clarify the relationship between the two phenomena: this task is important in its own merit, especially in the light of the growing literature on both issues. Second, I want to compare my proposal to that recently put forward by Moss (2011), whose account turns just on an extended parallel with communication.

De se attitudes pose a challenge for our theories of communication. It’s intuitive to think of communication as the transfer of information between two agents (call them ‘Sender’ and ‘Receiver’). Sender’s making an assertion which expresses her belief that $p$ normally results in Receiver forming a belief with the content that $p$. This is the picture underlying standard accounts of communication and assertion (see e.g. Stalnaker (1978)). But this model is ill-fitted for centered worlds content. Take a variant on Perry’s (2006) cake example. Perry, but not you, believes that he’s going to eat the cake on the table between the two of you. Hence, in all centered worlds compatible with his belief state, Perry’s personal alternative is going to eat the cake on the table. If Perry transmitted to you the content of that belief, you would come to believe that you are going to eat the cake. But this is clearly not the expected outcome of Perry’s uttering I’m going to eat the cake.

The problem posed by communication of de se attitudes concerns content. More precisely, it’s the problem of defining a notion of content that (a) somehow still captures, or maps onto, the de se element and (b) can do work in a theory of communication, given that different agents occupy different positions within the world.²⁵

²⁵Just searching for such a notion of content is what most attempts at a solution do: see Stalnaker (2008) and (2011), Egan (2010), and Moss (2011). A variant strategy consists in giving up on the notion of content transferred in communication, and rather defining a procedure for mapping contents uttered by Sender into contents used in update by Receiver (see Torre (2009) and Ninan (2010)). But also in this case, it’s clear that the problem concerns coordination of the contents in play in communication.
This problem has an analog for the case of intertemporal update. Here we need to explain what content is passed on, via memory, between temporal slices of an individual. For example, we want to explain what content is shared by the belief that you expressed yesterday as *It's sunny today* and the one you express today as *It was sunny yesterday*. Again, what is needed is a notion of content that somehow captures the *de se* element and also allows for transfer of beliefs through time. But this is not the only problem posed by intertemporal update. In addition, we must explain how subjects change their beliefs about their position in the world as this position changes. This problem doesn't concern the content passed along via memory, but rather the relative position between different temporal slices of an individual. To use effectively the information that your earlier slices pass along to you, you must know how distant they are from you; in other words, you must be able to calibrate your position with respect to them. To give a model of update, we must also model this calibration process.²⁶

To see that the content problem and the calibration problem are distinct, it's useful to consider an example. Take the following scenario:

*The Mad Experimental Philosopher.* A mad experimental philosopher has kidnapped you to run an experiment on you. First, he will inject a drug that will severely hinder your time-tracking capacity. Then he will put you to sleep for a quantity of time determined by a random number generator. As you wake up, he will make you fill questionnaires investigating how losing track of time affects your opinions on causation.

The mad philosopher's drug obviously hampers some of your cognitive processes. But which ones? Clearly, your previous credences in *de dicto* and *de re* propositions are unaffected. Suppose that the day of your kidnapping was a sunny day. Hence you came to believe that *that* day it was sunny. As you wake up, you still retain that belief. Similarly for any other demonstrative beliefs you held before being put to sleep; your memory is untouched. But of course, because of the drug, you're unable to relate *that* day to your current position in time. Hence you can't form new *de se* beliefs, or new *de re* beliefs concerning the distance between this moment and *that* day.

The mad philosopher case shows that transferable content alone is not enough to perform intertemporal update. Your temporal slices are perfectly able to communicate and share temporal beliefs: the notion of *de re* content I defined in section 3.3 is designed specifically for this purpose. But something more is needed for intertemporal update. Your later temporal slice needs to know how she is related to your earlier one: i.e., how distant in time she is from the earlier slice. Without any knowledge of this sort, you're unable to update effectively.

²⁶I borrow the term from Stalnaker (2011). Notice that Stalnaker is not fully explicit about the distinction I want to draw.
The upshot is that communication provides, at best, a partial analogy for intertemporal update. In fact, there is no analog of the calibration problem in ordinary communication: speakers invariably have common knowledge of their respective positions. As you speak to me at time $t$, it’s common knowledge between us that you are the speaker at $t$ and I am the individual you’re speaking to at $t$. The mere fact that we’re engaging in communication provides each of us, as a default, with a way of locating the other with respect to ourselves. Hence all we need to account for communication of *de se* beliefs is defining an appropriate notion of content.\(^{27}\)

### 3.6.2 Comparison with Moss’s account

As I hinted already, Moss’s (2011) account of update is based on an extended analogy with communication. The slogan summing up her account is “engaging in *de se* update is like engaging in a conversation with your past self”. There are a number of similarities between her account and mine. First, Moss’s account also uses a notion of *de re* content, though one that is quite different from the one I defined in section 3.3\(^{28}\). Second, Moss contemplates, besides genuine learning, a kind of updating procedure, which she calls ‘black box updating’ which seems reminiscent of relocation. (In essence, what you learn in black box updating is what you would learn about the passage of time if you were secluded from any kind of evidence coming from the external world.) So the two accounts have obvious points of contact.

At the same time, Moss’s account is insufficient to get a model of intertemporal update. Her account solves the content problem, defining a notion of *de re* content that is transferred between different temporal slices of a subject. But the calibration problem is basically left untouched. Here is a summary of the account that Moss herself gives:\(^{29}\)

> In black box updating, you form beliefs on the basis of information you get from your previous self. Getting information from your previous self is just like getting information from other agents. Each *de se* proposition you used to believe is equivalent with some *de dicto* proposition . . . This kind of

---

\(^{27}\)To be sure, we could describe contrived cases of communication where an analog of the calibration problem arises. To do this, we would need a special situations where actual individuals assume different identities throughout a conversation. For example, we might play a theater game where a number of actors read lines from a corresponding number of characters in a scene, and at given intervals actors switch characters, according to a pre-established rotation. To successfully take part in the game, actors should keep track of the rotation of identities: this process is structurally to the calibration performed in intertemporal update.

\(^{28}\)In fact, the background metaphysics in Moss’s account is haecceitistic (Moss, p.c.), on a par with what happens for Stalnaker (2008).

\(^{29}\)Caveat: Moss uses ‘*de dicto*’ as a blanket label for *de dicto* and *de re*. In particular, the proposition equivalent to a *de se* proposition is actually a *de re* proposition.
**3.7. CONCLUSION**

A de dicto proposition is something you can currently believe. Furthermore, you currently have some de se beliefs about your relation to your previous self. So you can also currently believe some de se propositions: the consequences of your current de se beliefs and your old de dicto information. (p. 12)

So far as I can see, Moss simply stipulates that “you currently have some de se beliefs about your relation to your previous self”. Nothing in her account explains what is the process that generates these beliefs. But this problem—the problem of calibration—seems to be the very heart of the problem of intertemporal update. Calibration is what is distinctive about this kind of update and what makes giving an account of the phenomenon hard. We don't have a genuine account of intertemporal update until we have an account of how later temporal slices of a subject individuate their position with respect earlier slices.

Hence the main advantage I claim for my theory over Moss is completeness. I have tried to give a detailed account of the calibration problem. In doing this, I make two main moves that differentiate my account from hers. First, I single out a different kind of update operation which essentially consists in a massive shift of credences through agent’s doxastic states. Second, I argue that this operation has very different effects from standard update operations like conditionalization. If it is performed effectively, its outcome is retaining the subject's level of certainty in temporal information. If it is performed ineffectively, it leads to a loss of certainty in de re information. These two moves are the main elements distinguishing my theory from Moss. Moreover, just these two moves are responsible for the philosophical consequences of the theory: on the one hand, the vindication of a version of the Relevance-Limiting Thesis; on the other, the conciliatory account of the Sleeping Beauty problem.

3.7 Conclusion

In his original discussion of self-locating belief, David Lewis writes:

[I]t is interesting to ask what happens to decision theory if we take all attitudes as de se. Answer: very little. We replace the space of worlds by the space of centered worlds . . . All else is just as before. Whatever the points of the space of possibilities may be, we have probability distributions over the space and assignments of utility values to the points. (1979, page 534)

Much literature on self-locating update takes for granted that Lewis is wrong, at least with regard to credences. Self-locating information is problematic for standard Bayesianism, the claim goes; our standard update rules require drastic revision. I disagree. Lewis
is essentially right: we can switch from possible to centered worlds and keep using standard Bayesian machinery to model update. Lewis simply overlooks that this machinery needs to be supplemented. Subjects’ position within the world changes: as a result, self-locating information that is true of them at certain points is not true of them at others. Keeping track of these changes requires a cognitive process very different from learning. The goal of my paper has been introducing this process, explaining its functioning, and exploring some of its consequences.

In closing, let me point to a methodological moral. Much literature on de se update focuses on epistemological issues. My starting point has been rather in philosophy of mind: I’ve built my account on a theory of the nature and the dynamics of the information involved in update. I hope I have showed that, at the very least, new and interesting possibilities are opened once we take this route. The lesson, I think, is that questions in the epistemology of de se belief cannot be detached from foundational questions about information and the relationship between thought and context.
3.7. CONCLUSION

Appendix: further notes on the Sleeping Beauty problem

As I say in section 3.5, my account falls short of determining a precise numerical value for Sleeping Beauty’s credences on Monday. Here I investigate a possibility for filling in this gap.

On Sunday, Beauty starts off with credence .5 in each of the two possibilities in her belief state (by Lewis’s Principal Principle (1980a)). Moreover, the final step in the update procedure is simply conditionalization; hence it’s clear how Beauty’s credences evolve after relocation. The missing element is in the middle: we should specify the probabilistic weights in the weighted partitions that Beauty uses for relocating. This partition involves the proposition that one day has passed since Sunday (call it ‘OneDay’) and the proposition that two days have passed (call it ‘TwoDays’):

\[ \{ \langle \text{OneDay}, p_1 \rangle, \langle \text{TwoDays}, p_2 \rangle \} \]

The question to adjudicate is what \( p_1 \) and \( p_2 \) are. Informally, the question is what credences Beauty should have, on Monday morning, about the two different hypotheses about the passing of time (given that she knows her memories from her time-tracking capacities might have been tinkered with).

Here is an appealing way of filling the gap. I don’t endorse this solution here, but it’s worthy of further investigation and deserves to be flagged. Assume an indifference principle about weighted partitions³⁰:

\[ \text{Partition Indifference. } \text{If two members of a weighted partition } A_1 \text{ and } A_2 \text{ are subjectively indistinguishable and both receive precise nonzero credence, then } A_1 \text{ and } A_2 \text{ should receive equal credence.} \]

This would make it the case that each of OneDay and TwoDays gets credence .5. By the imaging rule for relocation, we would get that the credence distribution of Beauty after relocating is

\[
Cr^*(w') = \sum_w Cr(w) \times \begin{cases} 0.5 & \text{if } w_{\text{OneDay}} \text{ is } w' \\ 0 & \text{otherwise} \end{cases} + \sum_w Cr(w) \times \begin{cases} 0.5 & \text{if } w_{\text{TwoDays}} \text{ is } w' \\ 0 & \text{otherwise} \end{cases}
\]

Since Beauty starts with credence .5 in the two sequenced worlds in her belief state and since each of these credences is split equally between the OneDay-closest and the TwoDay-closest sequenced worlds to them, we end up with the credence distribution:

³⁰For more principles of this sort in the literature, see the principle of indifference originally proposed by Elga (2000) and the modified version given by Briggs (2010).
From here, by conditionalization, Beauty ends up with credence 1/3 in the only remaining heads possibility.

This solution has a lot to say for it, but I refrain from endorsing it here. I have two main reasons. First, indifference principles like Partition Indifference should receive stronger justification than intuitive plausibility. Second, it might be that the Sleeping Beauty scenario is one where subjects should not adopt precise credences. So it might be that Partition Indifference is correct but has no specific consequences for the Sleeping Beauty scenario. Both these points deserve more extended discussion.
Bibliography


Elbourne, Paul (????). *Situations and Individuals*. Cambridge, MA: MIT Press.


Evans, Gareth (1982). *The varieties of Reference*. OUP.


BIBLIOGRAPHY


BIBLIOGRAPHY


