July 2018

THE FIRST PERSON PERSPECTIVE: LANGUAGE, THOUGHT, AND ACTION

Pengbo Liu
University of Massachusetts Amherst

THE FIRST PERSON PERSPECTIVE: LANGUAGE, THOUGHT, AND ACTION

A Dissertation Presented

by

PENGBO LIU

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2018

Philosophy

THE FIRST PERSON PERSPECTIVE: LANGUAGE, THOUGHT, AND ACTION

A Dissertation Presented

by

PENGBO LIU

Approved as to style and content by:

Joseph Levine, Chair

Louise Antony, Member

Alejandro Pérez Carballo, Member

Dilip Ninan, Member

Seth Cable, Member

Joseph Levine, Department Chair
Philosophy

DEDICATION

To Yiwei and Yichen.

ABSTRACT

THE FIRST PERSON PERSPECTIVE: LANGUAGE, THOUGHT, AND ACTION

MAY 2018

PENGBO LIU

BA, WUHAN UNIVERSITY

M.Phil, THE CHINESE UNIVERSITY OF HONG KONG

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Joseph Levine

What it is to have a first person perspective? How do we come to understand our own perspective in the world? How do we take into account other people's perspectives in our social and linguistic interactions? This dissertation is an exploration of these issues. But instead of approaching them in the abstract, it aims to shed light on these difficult questions through a series of case studies. First, I examine the role of the first person perspective in our agency, and explain the sense in which it is essential for action. Next, drawing on recent work in psychology, I propose an model of the development of temporal self-understanding in young children. Lastly, I develop a two-level pragmatic theory of epistemic modals.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

CHAPTER 1
INTRODUCTION

The first person perspective enjoys a prominent theoretical place in many areas of philosophy, linguistics and psychology. For instance, the perspective-dependency of linguistic expressions, such as indexicals (e.g., “I”, “here”, “now”) and epistemic modals (e.g., “might”, “must”), is one of the most discussed topic at the interface of formal semantics and philosophy of language. In the philosophical literature about propositional attitudes and agency, it is widely believed that first personal attitudes are in some sense essential to our intentional actions. More generally, the first person perspective is often said to be offering us a special cognitive access to the world and our place in it, and partly because of this, its practical and theoretical role is inexplicable from an objective point of view and in third personal terms.

On the other hand, many foundational questions remain to be answered: what it is to have a first person perspective? How do we come to understand our own perspective in the world? How do we take into account *other* people’s perspectives in our social and linguistic interactions? This dissertation is an exploration of these issues. But instead of approaching them in the abstract, it aims to shed light on these difficult questions through a series of case studies. Specifically, I will focus on three issues: the problem of the essential indexical (Chapter 2), the development of self-understanding in young children (Chapter 3), and the pragmatics of epistemic modals (Chapter 4). Given the nature of these problems, my general approach is interdisciplinary, and I will draw widely from empirical research in psychology and linguistics.

The aim of this introductory chapter is twofold: first, to offer a brief overview of the issues, arguments, and conclusion of each chapter; second, to situate the chapters, each of which has a relatively specific focus, into a larger context and to make their interconnections clearer than they otherwise would seem.

1.1 The First Person and Agency.

Many philosophers are sympathetic to the general idea that the role of the first person perspective in human agency is special and cannot be fully captured in objective, third-personal terms. The popularity of this idea can be traced to the important works of John Perry (1979) and David Lewis (1979) in the 1970s. Roughly put, one of their key points is that the motivational and explanatory role of first-personal beliefs cannot be exhausted by third-personal beliefs, and, as a result, any adequate theory of belief has to make room for the distinct contributions of first-personal beliefs. For example, imagine that a bear begins to chase me when we are walking in the woods. I curl into ball and you run away to get help. It is possible that we share the same relevant beliefs about the way the objective world is (e.g., who is chased by the bear, what are the relevant features of the woods and the bear, etc.), as well as the same desires about how the situation should turn out (e.g., I am saved). Still, our first-personal beliefs are in some sense different (i.e., the belief expressed by "I am chased by a bear" can only be properly attributed to me), and this crucial difference explains why we react differently to the same situation. Hence, they argue, first-personal attitudes are essential to the explanation of intentional actions.

However, this familiar line of thought has recently been challenged by authors. (Cappelen and Dever, 2013; Devitt, 2013; Magidor, 2015) The objections can be divided into two basic kinds: firstly, it is said, indexical attitudes are not always essential to causation and explanation of action; secondly, even if indexical attitudes are indeed essential to some actions, these cases are nothing but instances of a general phenomenon (e.g., Frege's puzzle, the opacity of explanation, etc.). In other words, the putative problem of the essential indexical has nothing to do with indexicality *per se*.

To understand what is at stake in this ongoing debate, I first consider the question: what it is to have a first person perspective? This is a surprisingly difficult question, and theorists who invoke this notion often mean different things by it. Hence, a necessary step towards the final resolution of the debate is a clearer understanding of the first person phenomena at issue. In this chapter, I propose an account of first personal attitudes and offer a novel explanation of the problem identified by Perry and Lewis. Along the way, I address the challenges raised by those who are skeptical about the distinctiveness of the first person perspective.

In response to the first kind of objection, I distinguish between two kinds of indexicality: implicit indexicality and explicit indexicality, as I shall call them. The former consists in the ability to represent things from one's spatiotemporal location, whereas the latter depends on the explicit representation of the self as such. Using examples from comparative and developmental psychology, I argue that implicit indexicality is a more primitive ability and is prevalent in the daily interactions between organisms (including humans and lower-animals) and their surrounding environments. Explicit indexicality, on the other hand, is more cognitively demanding, for here the self is not indirectly implicated in one's action (as in the case of implicit indexicality), but is explicitly represented as such. Furthermore, I argue that the first type of objection mentioned above rests upon a conflation of the two importantly different kinds of first-personal representations.

In response to the second kind of objection, I argue that skeptics about the essentiality of the first person perspective have misconstrued the phenomenon that Perry and others are trying to capture. The issue can be made clearer if we consider an a priori question concerning the explanatory force of beliefs: why are some beliefs explanatory of actions in the first place? In pursuing this question, I argue that the so-called problem of the essential indexical should be characterized as what I call the Explanatory Asymmetry between indexical and non-indexical thoughts. The basic idea of the asymmetry is this: the explanatory force of non-indexical thoughts can be explained in indexical terms; by contrast, the explanatory force of indexical thoughts cannot be explained in non-indexical terms. Importantly, this asymmetry has nothing to do with the opacity of explanation or Frege's puzzle. Therefore, the problem of the essential indexical is not merely an instance of a general problem and is indeed worthy of special attention and treatment.

1.2 Self-Recognition in Psychology

After distinguishing explicit from implicit indexicality, I then take a closer look at explicit indexicality and the nature of self-representation. In Chapter 3, I critically examine the psychological models of mirror self-recognition (MSR) and delayed self-recognition (DSR), and then propose an account of the development of children's temporal self-understanding (roughly, their conception of the self as an entity that persists in time).

Although we rarely think of it this way, recognizing oneself in a mirror is actually a remarkable capacity that we share with only a few other species. Perhaps unsurprisingly, then, the famous mirror mark test (Gallup, 1970; Amsterdam, 1972) is often regarded as the litmus test for self-awareness in non-human animals and young children. While there are some controversies with regard to what kind of self-awareness is measured by the test, it is generally agreed that mirror self-recognition (MSR) at least requires the capacities (1) to form an explicit representation of the *bodily* self, (2) to match the bodily self-representation to one's mirror image, and 3) to properly integrate the visual information about the mirror image with the proprioceptive, kinesthetic and somasthetic information about one's own body.

The empirical literature, however, often fails to distinguish two related but importantly different interpretations of MSR. On the first interpretation (which I call *the correspondence account*), MSR is essentially a product of *cross-modal comparison*: the subject is aware of the cross-modal correspondence between her bodily self-image and her mirror reflection, and is able to exploit that correspondence in her self-directed behaviors. The second interpretation, which I call *the reference account*, is perhaps more intuitive, and is also more popular among developmental and comparative psychologists. It construes MSR as a form of *self-identification*: the subject recognizes that the mirror image is an image *of* the self (as if mentally pointing to her mirror reflection and think: that is *me*!). The key difference is that the reference account, but not the correspondence account, implies that subjects who reliably pass the test can appreciate the referential property of one's mirror reflection.

I then proceed to argue for the correspondence account on both theoretical and empirical grounds. From a theoretical point of view, the correspondence account postulates significantly less cognitive structures and capacities than the reference account, and, partly because of this, it is also more parsimonious. More importantly, empirical evidence from neuropathology and the developmental psychology of symbolic understanding weighed strongly against the reference account: First, patients with "mirror sign delusion" (Phillips, 1996; Breen et al., 2000) can pass the mirror test (and perform self-directed acts in front of a mirror, such as combing), but they characteristically fail to recognize their mirror image as image of themselves; Second, the reference account presupposes that the subjects who exhibit MSR have achieved what the psychologist Judy DeLoache (2004; 2011) calls *dual representation* (roughly: the ability to represent both the mirror-image *and* the abstract, representational relation the image bears to their own body). However, the large literature on children's symbolic understanding indicates that children have not acquired this capacity until they are 2.5 to 3 years old—at least a year after they can reliably pass the mirror test.

However, many evidence suggest that young children's self-understanding is importantly limited in that they seem unable to grasp the *temporal* dimension of the self. For example, in a series of *delayed* self-recognition studies (which use video recordings or photographs instead of mirrors), psychologist Daniel Povinelli and his colleagues (1996; 2001) argue that 2-year-olds who have passed the mirror test still have significant limits in their self-understanding: the 2-year-olds recognized themselves in the videos, which were recorded just a few minutes ago, but somehow failed to appreciate the relevance of the earlier events they saw to their present states. For many psychologists, this and other related experiments suggest that very young children lack the sense of the self as an enduring entity, i.e., an entity that extends in time with a past and a future–in Povinelli's terminology, 2-year-olds only have the concept of "the present self", but not "the extended self". The later emerges only when they are around 4 years old.

This trajectory raises a developmental question. How do young children become aware of the temporal dimension of the self? That is, how do they come to realize that their past self (e.g., recalled in memory) and their present self are the same ? After considering and rejecting the cognitive model proposed by Povinelli, and the social-cultural model proposal by Katherine Nelson and Robyn Fivush (Nelson and Fivush (2004); Fivush (2011)), I propose an alternative account, i.e., the social mirror hypothesis. Just like physical mirrors enable us to see reflections of aspects of our physical self that we are otherwise unable to see, other people, especially our interlocutors, enable us to gradually appreciate the temporal aspects of the self. More specifically, the hypothesis postulates three key elements: (1) a bodily self-representation, (2) episodic memory, and (3) parent-guided joint reminiscing about the past, which enables the child to connect her present state with her past experiences and explicitly identifies with her past self. In virtue of the bodily self-representation, young children are able to conceive of themselves as objective entities in the world, but the representation itself is *a*temporal. In virtue of their episodic memory, they are able to recall the events that happened in the past, but their own presence in the past is not explicitly reflected in the content of the memory. It is through linguistic communication, especially in the form of parent-guided talks about children's past (in which they are referred to with proper names and pronouns), that the objective presence of the children in the past, as well as the connection between their past and present, is made explicit. Hence, the hypothesis implies that the contribution of social interaction and linguistic communication is not contingent factor; rather, it is central to the development in that it makes it possible to *combine* the two (relatively) primitive mechanisms in a way that eventually brings about temporal self-understanding.

1.3 Perspectives and Communication

The previous chapter has brought to the fore the importance of linguistic communication for the development of a full-fledged conception of the first person. This chapter continues this line of inquiry, and deals more directly with language and communication, and in particular with the understanding and communication of the perspectives of others.

It is impossible to overstate the importance of perspectives in our linguistic communication. What our words mean often depends on where we are, what background information we have, when we are speaking, etc. For a conversation to proceed smoothly, the listener needs to be aware of (or even to some extent share) the speaker's perspective. But it would be silly to ignore the differences. Those unable to see things from others' perspectives, or to appreciate that others might have a perspective different from their own, will often find it hard to understand what others mean, let alone engage in fruitful exchanges. In this chapter, I will approach this general issue with a case study: our intuitive judgments about bare epistemic possibility claims (BEPs), such as “The keys might be in the car.”

Contextualism and relativism are the two major semantic theories of epistemic modals. According to the former (Kratzer, 1991, 2012), epistemic modals are quantifiers over epistemic possibilities, whereas the domain of possibilities is determined by the context of utterance: the context selects a particular body of information, and the truth-value of a bare epistemic modal claim, whose general logical form can represented as *Might/Must* p, is determined by the compatibility of p (the prejacent of the modal claim) with that body of information. Relativists (Egan, 2007; Stephenson, 2007; MacFarlane, 2014), by contrast, hold that a) an epistemic modal claim asserted in a context is always evaluated relative to (the information available to) a judge or assessor, and b) since different judges/assessors may have different bodies of information, an epistemic modal assertion could be true relative to one judge/assessor, but false relative to another.

The debate between contextualists and relativists has come to a stalemate. Each group claims that there are problematic cases that the other group cannot explain. To advance our theorizing, I think it is important to empirically examine our linguistic intuitions about such cases and their psychological basis. For example, recent experimental studies (Knobe and Yalcin, 2014; Khoo, 2015) discovered an interesting phenomenon: in some contexts, people tend to think that it is appropriate to reject/retract a BEP, while also judging that they are true, or at least not false. This is surprising, for normally if one has said something true then one should not reject or retract it. However, in the case of epistemic modal claims, it seems that folk intuitions about the appropriateness of rejection/retraction and the truth of a claim come apart.

To explain this phenomenon, I propose an account of BEP that focuses on the communicative function of BEP. The guiding question is: what are we doing when we utter a BEP in a discourse context? According to my account, the communicative function of BEPs is twofold. When a speaker asserts that it might be the case that p, she is performing two speech acts: (1) suggesting that some objective fact, p, should be mutually recognized as a live and significant possibility, (2) indicating the perspective-dependency of her suggestion or proposal: although p is not taken to be an established fact, it is nevertheless a reasonable suggestion, given the information available to the speaker. Of these two functions, the first is primary, and the audience generally responds to it by either accepting or rejecting the suggestion; the second is less salient, but is nevertheless crucial for a cooperative communication.

My two-level account of epistemic modals implies that evaluators of epistemic possibilities normally would take into account the perspective of the speaker. That is, the proper assessment of epistemic modal claims requires perspective-taking. To examine this aspect of communication, I draw on research in social psychology, and in particular the anchoring-and-adjustment model of perspective-taking, according to which people adopt others' perspectives by serially adjusting from their own. Consistent with this model, it is widely agreed among psychologists that both human children and adults are subject to a common cognitive bias, sometimes called "epistemic egocentricism" or "the curse of knowledge"(Royzman et al., 2003): when we predict what others would think or do, and when we interpret what others mean, we are inclined to over-impute our own privileged information (i.e., information only accessible to us) onto them.

Given this cognitive bias, my account of epistemic modals predicts that the more difficult it is for one to adjust to the speaker's perspective, the more likely it is for one to evaluate the speaker's claim on the basis of one's own privileged knowledge (and the listener's evaluation will likely change once the listener acquires more information about the speaker's perspective). This prediction can be tested. Therefore, in the final sections of this chapter, I discuss some general obstacles to perspective-taking and several specific scenarios in which our egocentric bias could significantly affect our evaluations of epistemic modals. These cases, I argue, lend further support to my account.

In these chapters, we will explore a variety of philosophical and empirical problems, many of which are only remotely related. But stepping back, it is not too hard to see the common thread that underlies these investigations. We start with some rather primitive perspectival phenomenon, such as the egocentric spatial representations of nonhuman animals and human infants, then to the explicit sense of the self as such, some aspects of which are, arguably, distinctively human—one case in point is the ability to understand the self as temporally extended. If the argument in chapter 3 is on the right track, language and communication are essential to the development of our temporal self-understanding. This turn to language and communication marks the transition from one's own perspective to the perspectives of other people: how perspectives are expressed and understood in linguistic communication? So in chapter 4, We will examine this general issue through the lens of epistemic modals. Of course, many of the ideas developed here are just the beginning of a systematic study of the nature and role of the first person perspective, and there are still a lot of questions to be answered; but at the very least, I hope that what emerges from these explorations is an empirically-informed, multifaceted, and cogent picture of the first person perspective in our mental, agential and social lives.

CHAPTER 2

ESSENTIAL INDEXICALITY AND ACTION

You are being chased by a bear. You curl into a ball and I run away for help. We have the exact same beliefs about objective facts of the world, and about what should be done in this situation. But why do I act differently from you? Easy: because I believe that you, not me, are the one being chased at the moment. And you believe the opposite. Hence, it seems that to make sense of our actions in the situation, we have to invoke our respective indexical or de se attitudes: what we do is not merely matter of our beliefs about the objective world, but also a matter of our beliefs about ourselves as such.[1]

Examples like these have led many philosophers (Castañeda 1966; Perry 1979; Lewis 1979) to think that indexical or *de se* attitudes have a special significance in our mental and agential life. There are different ways to explicate the significance of such attitudes, but one important consensus seems to be that their role in the causation and explanation of actions is in some sense essential, irreducible or ineliminable. This role in turn, it is often claimed, has to do with the fact that indexical or *de se* thoughts are manifestations of the first-person perspective[2], or that they are inherently connected with self-awareness, etc.

[1] This example is adapted from Perry (1977)

[2] In this chapter, I use "indexical thought" and "de se thought" interchangeably, although strictly speaking, not all indexical expressions are first-personal (e.g.,"they", "then"). I focus on a narrower group of indexicals , such as "I", "here" because the problem of the essential indexical or *de se* attitudes, as discussed by Perry and Lewis, is only about this group.

Recently, however, a number of authors[3] have challenged the essentiality of indexical attitudes. The objections can be divided into two basic kinds: firstly, it is said, indexical attitudes are not always essential to causation and explanation an action; secondly, even if indexical attitudes are indeed essential to some actions, these cases are nothing but instances of a general phenomenon (e.g., Frege's puzzle, the opacity of explanation). In other words, they have nothing to do with indexicality *per se*.

Although ultimately I will argue against these authors, I do think that they have raised some important, and often neglected, issues. While much of the literature have taken for granted the essentiality of indexical thoughts, it is often unclear exactly *in what sense* they are essential, and *what it is about them* that make them essential. The primary goal of the chapter, then, is to address these questions. More specifically, I will propose an account of indexical attitudes and a novel explanation of the phenomenon identified by Perry and Lewis. What emerges from these is, I hope, a clearer picture of the first-person perspective and its psychological role, which in turn would help us meet the challenges mentioned above.

This chapter will proceed as follows. In section 1, I sketch the background of the debate and distinguish several ways of understanding the essentiality of indexicality. In section 2, I distinguish between two kinds of indexicality: implicit indexicality and explicit indexicality. The key difference is that the latter, but not the former, presupposes explicit representations of the self, or self-concept. I then respond to the first objection to essential indexicality, i.e., that indexical belief is not always necessary to motivate or explain actions. This objection, I argue, rests upon a conflation of the two kinds of indexicality. In section 3, the longest section of this paper, I refute the second objection, i.e., the problem of the essential indexical is just an instance of a general

[3] Tiffany (2000); Spencer (2007); Cappelen and Dever (2013); Devitt (2013); Magidor (2015). I should note that the two-part challenge is most explicitly laid out in Cappelen and Dever (2013); Magidor (2015), while others seem to focus just on the second part of the challenge

problem. My main contention is that the opponents of essential indexicality, or the *de se* skeptics (to borrow a term from Ninan 2016), have misconstrued the phenomenon that Perry and others identified. The phenomenon should be characterized, not as a kind of opacity, but as what I will call **Explanatory Asymmetry** between indexical and non-indexical thoughts: the explanatory force of non-indexical thoughts can be explained in indexical terms, but explanatory force of indexical thoughts cannot be explained in non-indexical terms.

2.1 Perry's Argument in "The Problem of the Essential Indexical"

De se exceptionalism is the view that de se or indexical attitudes play an essential role in agency, and partly because of that, they pose a special challenge to some otherwise well-established views about propositional attitudes.[4] *De se* skepticism denies all these. It claims that de se attitudes are not essential for action, and, furthermore, that the challenge they pose is just an instance of a general problem that has nothing to do with de se attitudes *per se*.

Although I will eventually take issues with de se skepticism, the aim of this particular section is rather modest. It is to trace and clarify the alleged problem (or at least one problem) associated with de se attitudes, as it arises in John Perry's early work. His classic paper "The Problem of the Essential Indexical" is often cited as a major source of de se exceptionalism, but his argument in the paper, as will show below, is entirely compatible with de se skepticism. First, the problem he identified is a general problem that does not depend on the special features of de se attitudes per se; second, and perhaps more tellingly, the solution he proposed is also a general one, which does not give any special treatment of de se attitudes.

[4] This view is often associated with Castañeda (1966); Perry (1979); Lewis (1979) and their followers. The terms "*de se* exceptionalism" and "*de se* skepticism" are from Ninan (2016)

Let me unpack these claims. The problem of the essential indexical, as Perry construed it (1979), is a problem for the then-orthodox view of belief, according to which a belief is a relation between a subject and an objective proposition (the truth value of which does not vary with individuals and times). As it turned out, however, the real problem he identified is not these elements of the orthodox view, but an additional, albeit implicit, assumption of the view. The assumption can be briefly put as follows: in so far as we want to explain or rationalize an action in terms of the agent's beliefs, the explanatory import of beliefs is fully captured by the content of the belief, i.e., the objective proposition she believes.[5] This assumption is problematic because, according to Perry, many actions cannot be explained in this manner.

The example of being chased by a bear is a case in point. On Perry's favored account of belief content, you and I believe the same thing about the relevant aspects of the world (i.e., you are chased by a bear), even though the belief would be expressed in different ways in language (e.g., you would use the first person pronoun but I would not). Moreover, both of us desire that the same objective state of affair would obtain (i.e., you are safe), and may even agree that we should be done in order to achieve that goal (i.e., that you curl into a ball and I run away for help). From an objective point of view, then, our relevant attitudes in this situation are identical as far as their contents are concerned. Still, we respond to the situation differently.

[5] In a more recent article, Perry clarifies his target:

> I think, however, that it is natural to think of the received doctrine of belief as amounting to a bit more than theses 1[i.e. belief is a relation between subjects and propositions] and 2 [i.e.,propositions have objective truth-values] entail . . . This more is that the proposition believed captures the elements of belief relevant to rationality. . . The augmented received doctrine . . . holds that once we descend to the layer of qualitative propositions, or de dicto beliefs, or Fregean propositions—-different pictures, perhaps, but more or less equivalent for our purposes . . . we will not need a further distinction between what is believed and how it is believed to capture the rational consequences of belief. (2006, 208-209)

Or consider the famous messy shopper case. Perry told the story that he once followed a trail of sugar in a supermarket, searching for the shopper with the torn sack. Later, to his surprise, he realized that the messy shopper was actually himself. So he stopped following the trail and started to rearrange the torn stack in his cart. The newly acquired belief, expressed by “I am making a mess”, is essential to the explanation of this action. However, and this is Perry’s main point, the explanatory contribution of this belief is not a matter of its object or content. After all, the content of the belief is just a de re proposition, an ordered pair consisting of the shopper and the concept **is making a mess**. However, when he was searching for the shopper, Perry might have already 1) believed that John Perry was making a mess (e.g., through the testimonies of others) and 2) forgot that he himself was John Perry. Crucially, according to Perry, the belief John Perry is making a mess’ has the same content as his indexical belief expressed by “I am making a mess”, because the semantic content of“John Perry” and “I”, as used by Perry, is identical.

Now this assumption about contents of these two beliefs is not entirely uncontroversial. Hence, Perry spent a large portion of the article examining and refuting several attempts to construe their contents differently. Whether or not he succeeded in this task need not concern us here. For our purposes, what matters is the general shape of the problem as Perry understood it: what, if anything, is so special about indexical beliefs? It is thus helpful to see his proposed solution to the problem. Roughly put, he argued that the change of behavior is not due to the acquisition of a new belief that has a different content, but a new way of believing the same content, or a new type of belief state. So, for example, earlier when he was searching for the shopper, he believed that John Perry is making a mess, he was in one type of belief-state (under the guise of “John Perry”, as one may put it); but when he came to realize the messy shopper is actually himself, he was in a different belief state and believed the old de re content in a new way (e.g., from a first-person perspective, under the guise of "I", etc.).

In many ways, Perry's notion of belief state can be viewed as the mental counterpart of Kaplanian"character". The character of an expression is a function from a context to the expression's content at that context. The character of "I", for example, roughly corresponds to "the speaker of the expression". Hence, "John Perry" and "I" (as uttered by Perry) pick out the same individual, although the proper name and the indexical have different characters. Likewise, we can believe the same content in different ways, in virtue of being in different types of belief states. Furthermore, the same indexical expression can pick out different individuals, depending on the relevant features of the context (e.g., who speaker is). Likewise, when you and I are in the same type of belief state, the contents of our beliefs might differ. This is the case when both you and I have the indexical belief, say, "I live in Boston": your belief is about you, and my belief is about me.

Now it is important to emphasize that Perry's analysis is not about any special features of indexical beliefs per se, but rather, it concerns a general aspect of belief (i.e., belief state) that, he claimed, any complete theory of belief should take into account.[6] Hence, the main contrast he wanted to highlight in the paper is not indexical vs. non-indexical, but belief state vs. belief content. So, one might naturally wonder, what is so essential about indexicals?

[6] He concluded the paper with the following remark:

> To say that belief states must be distinguished from objects of belief, cannot be individuated in terms of them, and are what is crucial for the explanation of action, is not to give a full-fledged account of belief, or even a sketchy one... But... the problem of the essential indexical should teach us that no philosophy of belief can be plausible that does not take account of the first. (1979, 49-50)

At one point, Perry said: “These indexicals are essential, in that replacement of them by other terms destroys the force of the explanation, or at least requires certain assumptions to be made to preserve it.” (1979, 35)This seems to suggest the following thesis

- **The Essentiality of Indexical Expressions (EIE):** The explanatory role of indexical expressions cannot be fulfilled by non-indexical expressions.

How does **EIE** relate to his more general point about ways of believing or belief states? The answer seems to be this: in describing the contents of one’s thoughts/beliefs, the use of distinct but co-referential expressions typically reflects different ways of thinking or believing about the same entity. So, for example, using “I” to self-refer reflects one’s first-personal way of thinking about oneself. Adopting the“belief state” terminology, we can call the mental states that correspond to indexical expressions such as “I”, “here”, “no”, etc., indexical states, and call the mental states expressed by, or corresponding to, non-indexical expressions non-indexical states. For the ease of exposition, I will just focus on “I” and the corresponding states in the subsequent discussion, but my points can be easily extended to other indexical states.

Now Perry’s reasoning can be restated as follows: to explain certain actions, it is necessary to invoke the agent’s indexical states (or first-personal way of thinking), for the explanatory force will be diminished if we only appeal to the agent’s non-indexical states (or third-personal way of thinking). Further, this is in turn because that these actions (e.g., the messy shopper’s change of behavior) are caused or motivated by the agent’s indexical thoughts/beliefs. Hence, there are two closely related theses that underwrite **EIE**:

- **Essentiality of Indexical States (EIS):** The explanatory role of indexical states cannot be fulfilled by non-indexical states.

- **Essentiality of Indexical States* (EIS*):** The motivational role of indexical states cannot be fulfilled by non-indexical states

Suppose **EIS** and **EIS*** are correct, what does this tell us about the (allegedly) irreducible explanatory role of belief states in general? Here we need to recognize an additional assumption of Perry's argument, namely, that the contents of propositional attitudes are publicly accessible: different believers can access the same belief content. For example, he claims that the content of his first-personal belief "I am making a mess" and the content of his third-personal belief "John Perry is making a mess" are the same de re proposition. Now suppose, for reductio, that invoking attitude contents alone is sufficient for an adequate action-explanation. Then a third-personal explanation (an explanation that only invokes the agent's third-personal way of believing) could do just as well as a first-personal explanation, provided that the relevant propositional attitudes share the same contents. But the messy shopper case and the like show that sometimes third-personal explanations are insufficient. According to Perry, then, we need to invoke belief states, or ways of believing, to explain (at least some) actions in addition to belief contents.

Understood this way, the role of the puzzle cases involving indexical beliefs are used to illustrate a general point about the (belief-desire) action explanation, and in particular, about the distinct explanatory role of beliefs states. Crucially, this general point is entirely compatible with one major complaint of *de se* skeptics: that there is nothing particularly problematic about indexical attitudes. Given the way he analyses and addresses the problem, there is no principal reason to think Perry's general point can only be illustrated with indexical beliefs.

That being said, it remains possible that the contrast between indexical and non-indexical states is still important and interesting in our theorizing about agency and action explanation. Indeed, as I will show below, once we shift our focus to indexical states, we will be in a better position to appreciate the role of the first person perspective in our mental and agential life, and, consequently, to understand what is at stake in the recent debates surrounding *de se* attitudes.[7]

On my reading, then, the insight of Perry, Lewis and others is that the *first-personal way of thinking* is critically important for motivating and explaining (at least some of) our actions. I do think they are onto something important and philosophically significant; however, I also think that the literature tends to over-intellectualize the phenomenon (or at least one of the phenomena) they have identified. Specifically, the first-personal way of thinking that is essential for action can be manifested in more primitive but no less important mental states, which do not involve explicit thoughts about the self as such. In the next section, I will elaborate on this point. This in turn will help us to respond to some of the skeptical challenges (section 3).

2.2 Babies, Rats, and Two Kinds of Indexical States

As noted above, we can understand Perry and others as arguing for the importance of first-personal ways of thinking in motivating and explaining some of our actions (**EIS** and **EIS***). The irreplaceability of indexical expressions (**EIE**) should then be seen as derivative: it is derived from the fundamental roles such states play in our mental and agential life. However, as I will argue below, the first-personal way of thinking does not have to be manifested only in the use of indexical expressions. More specifically, I will show that, properly understood, having a first-person perspective does not require any, linguistic or otherwise, representation of the self as such. To

[7] Another related thesis, endorsed either implicitly or explicitly by some, is what I shall call:

- **Essentiality of Indexical Content (EIC):** The explanatory role of first-personal/indexical content cannot be fulfilled by third-personal/nonindexical content.

Now it should be clear that Perry himself is not committed to **EIC**, for he thinks, as just mentioned, that the content of indexical beliefs are de re propositions. Many theorists, inspired by the Perry-style examples, postulate non-standard contents: properties (Lewis 1979), reflexive contents (Higginbotham 2003), etc. I think this does not really solve the problem (if there is one), but I will leave this topic for another occasion. See Stalnaker (1981); Holton (2015) for the problems with Lewis' approach, and Recanati (2007) for problems with Higginbotham's approach.

clarify this point, I will first distinguish between two kinds indexical states, only one of which requires some explicit representation of oneself (e.g., the first-person pronoun or its mental counterpart, a self-concept).

For a preliminary sense of this distinction, consider the representation of space. Philosophers and psychologists have distinguished between egocentric and allocentric representation of space.[8] An egocentric representation uses an egocentric frame of reference; that is, it represents objects (e.g., their locations, directions, and distances) with respect to the perceiver's own body or parts of the body (eyes, heads, torso, hands, etc.). An allocentric representation, on the other hand, locates objects within a frame of reference independent of the perceiver's position. Both types of spatial representations are available in human spatial cognition, and, while they are distinct from each other, they can be connected and combined (Wang and Spelke 2002; Burgess 2006). As I use the term, creatures that can form egocentric spatial representation have first-person perspective. Crucially, while the first-personal representation is egocentric, it does not depend on the ability to explicit represent the bodily self as such. That is, one can represent things in the world from one's own perspective without representing oneself as such. In still other words, having a first-person perspective does not presuppose the ability to entertain thoughts about the self as such, i.e., the kind of thoughts we typically express with the first person pronoun and the like.

The underlying idea of the distinction is, I believe, both intuitively plausible and empirically supported. Something like it can be found in Perry's own work (1998)[9]; it also bears some resemblance to what the psychologist Michael Lewis calls "the mechanism of the self" and "the idea of the self"(2003). To be more concrete, let us

[8] See Klatzky (1998) for a review of psychologists' use of these concepts. For a classic philosophical treatment of egocentric representation of space, see Evans (1982). Recent discussions of the issue include Grush (2000); Schellenberg (2007)

[9] see also Campbell (1995); Recanati (2007)

consider another example. Young children start to use first-person pronouns around 18 months of age, but long before then, they can represent the world from their spatiotemporal locations and interact with their surroundings in relation to their standpoints. Reaching behavior is a case in point. Infants of 4-to-5 months old have learned to combine information acquired from various sensory modalities (vision, audition, proprioception, etc.) and take into account the information when they reach out for objects, and to calibrate their actions in relation to their distances from the objects (Yonas and Granrud 1985). Moreover, they also take into account their degree of postural mobility, and reduce their attempts to reach for graspable objects if the objects are far enough to jeopardize their balance (Rochat et al., 1999). It is possible, then, that very young infants (and many non-human animals) understand what is within (and out of) reach, but as the philosopher John Campbell (1995) has pointed out, this does not imply that they can form a detached, explicit representation of themselves and their place in the (physical) world. Likewise, in analyzing such behaviors, psychologists often attribute to infants some primitive sense of their own bodies, locations, relations to other things and individuals, etc., but they do not think young children have explicit representation of the (bodily) self until at least a year later. (see, for example, Rochat 2009; Lewis 2003).

Many non-human animals can form and use egocentric spatial representations, even though, as the available evidence suggests, they are not capable of representing themselves as such. An extensively studied topic in comparative and developmental psychology is spatial reorientation. In a classic experiment (Cheng, 1986), rats were first introduced to a rectangular-shaped enclosure and shown a food reward hidden in one corner (e.g., corner A; see Figure 1a). After being disoriented (e.g., rotated in a black room), the rats were then reintroduced to the room and allowed to search for the food. The results showed that the rats searched equally often at the correct corner (A) and the opposite corner (C)

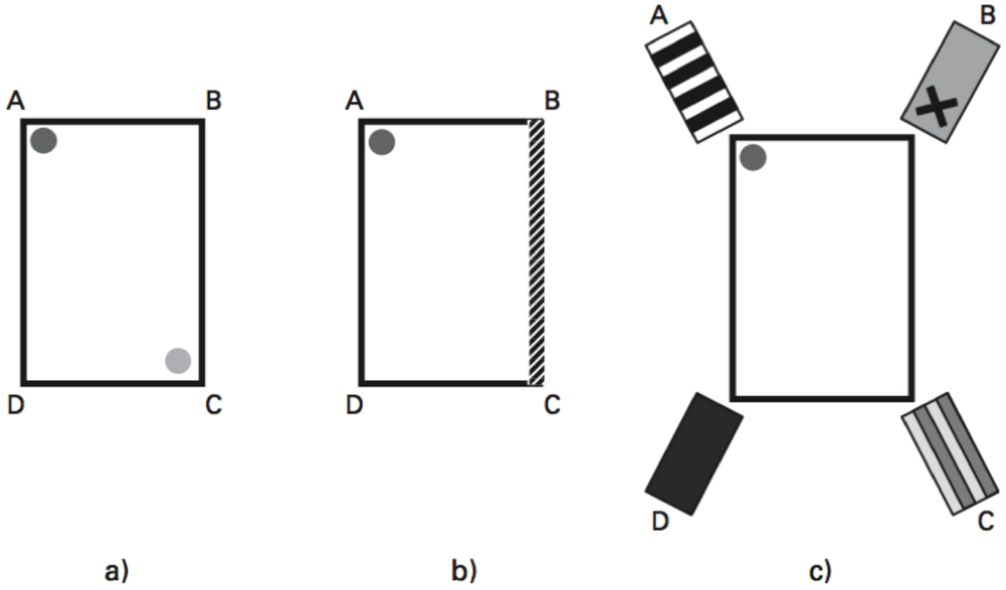

Figure 2.1. Spatial Reorienation

It is easy to see that the searching pattern makes perfect sense, since geometrically, A and C are equivalent: they are both corners to the right of a long wall. According to the standard psychological interpretation of these experiments, rats can use two types of geometric cues: the first is sometimes called geometric sense (of right and left), and the second metric information about length (Vallortigara, 2009). More strikingly, similar search patterns were observed when disambiguating non-geometric features were present: for example, when geometric information and non-geometric information together (e.g., one long wall was colored blue while all others are white, as in 1.b), or non-geometric information alone (e.g., each corner was attached to a different panel, as in 1.c, or with different olfactory cues), can uniquely identify the correct location. It appears, then, that in these experiments, rats only rely on geometric information. [10]

[10] Subsequent studies, using the same paradigm, have show that a wide range of animals, such as human children (Hermer and Spelke 1994), fish (Vargas et al. 2004), chicks (Chiandetti and

For our purposes, the most interesting thing is that rats' geometric sense (of left and right) clearly is first-personal: the properties of left and right always depend on, or are relative to, the perspective of the subject. Just as Perry's indexical belief (that he is making a mess) is essential to explain why he stopped his search, the rats' first-personal representations (that the food is on its right and towards the end of the longer wall) is also essential to explain their search patterns. However, psychologists do not (and probably should not) thereby attribute any explicit representation of the self, or self-concept, to them.

Notions such as *left*, *right*, *within reach* have immediate and crucial implications for our actions, and they are also indexical in the sense that their meanings and practical significance essentially depend on the location and capacities of the subject. The ability to represent these relations[11], however, does not presuppose the ability to conceive the self as such.

These examples, then, can motivate a distinction between two kinds of indexicality of mental states: *explicit* and *implicit* indexicality, as I will call them. To be more precise: a thought is explicitly indexical if it involves a self-concept, by which I mean a first-personal mental representation of oneself, a representation that corresponds to, or is normally manifested in, the competent uses of first-person pronouns.[12] Implicit indexicality, on the other hand, does not presuppose self-concept, or any representation of the self as such. Instead, an implicitly indexical thought is, or involves, an

Vallortigara 2008), are capable of encoding this geometric information. These results lead some theorists to postulate an innate geometric module for spatial representation (see Spelke et al. (2010))

[11] In Campbell's term, the "practical grasp" of these notions and their implications for action does not require a detached, reflective picture of the self as such.(1995, 41—51). He also talks about "causal indexicals", representations whose reference depend on the causal powers of the subject. What I call implicitly indexical representation might count as a sub-category of causally indexical representations

[12] For brevity and vividness, I sometimes speak of self-concepts as the mental counterparts of "I". A more precise way to put this is perhaps that one's competent uses of "I" express or manifest one's self-concept, as I want to be noncommittal with regard to the issue of whether self-concepts are mental indexicals. See Millikan (2012); Recanati (2012) for discussions of this issue.

egocentric representation of the world, i.e., it represents the world from one's point of view. The involvement of the self here is implicit, in that the self serves as the anchoring point of one's egocentric representations of the environment.

To further clarify the sense of explicit and implicit as used here, consider two cognitive systems that represent the same rule in importantly different ways. The first system explicitly represents the rule *if* p*, then* q, in that it stores the rule, in symbolic form, in its repertoire of inferential rules. The second system represents the rule only implicitly, in that whenever p is represented as an input, it automatically produces a representation of q as an output. In this way, the system implements or realizes the rule. Consequently, a major difference between the two ways of representing is rule is that the first is more flexible, because the explicit rule is available for a wider range of application. For example, the rule can be used to infer that p is not the case when the system accepts not-q. It can also be combined with other rules to perform more complex computations, etc. Similarly, implicit indexicality does not consist in any particular explicit representation (e.g., self-concept); it is realized in the cognitive-motor mechanism of the agent as a whole, and is manifested by its interactions with its surroundings: the agent is able to directly engage with their environment, through perception and self-initiated action. What information the agent receives and what it can do is partly determined by features of the embodied agent itself, e.g., its spatial (and temporal) location, its anatomy, its movement, etc. [13]

[13] It is worth noting that the kind of interaction at issue here is not just stimulus-response. Implicitly indexical representations are, first and foremost, *representations*. An entity that can form such representations not only registers information about its environment (like thermostats), it is also able to combine the information acquired from different channels (vision, audition, olfaction, etc.), integrate these information with stored information, and act upon its environment to bring about a desired consequence. For example, the rats in the experiment described above integrated the perception of the enclosure, their geometric sense of left and right, their memory of the location of the food, etc., and acted upon their room (i.e., searching at corners A and C) in order to get what they wanted: the food. The fact that a rat can display this intelligent search pattern is amazing, but at the same time, to do all this, they do not need any representations of themselves.

A creature capable of forming implicitly indexical representations has, we might say, a point of view, or a perspective. It interacts with the world from that perspective: the contents of its perceptual states, beliefs (or belief-like states) and desires/ goals (at least partly) depend on that perspective. It is, then, impossible to overstate the importance of implicit indexicality for its actions and, ultimately, survival . A creature equipped with explicitly indexical representation has the additional ability to represent itself as such. It has, in other words, a self-concept. What does the self-concept do? This is an enormously complex question that deserves much more space than allowed here, so I will only briefly describe several ways that the self-concept typically function in human beings (although some of them are likely to be shared by other higher animals): the self-concept enables one to conceive of oneself in a first-personal way, as the agent of one's action, the subject of one's thoughts and experiences, etc. It also helps the subject to better situate itself in the physical and social worlds, along with other objects and persons, to make plans for future actions (Barth et al., 2004), to locate past experiences in its autobiographical memory, and to construct a narrative about the self (Nelson and Fivush, 2004). These capacities deeply enrich our conceptions of ourselves, and thus facilitate more complex interactions with the physical and the social world. Most non-human animals, and probably young human infants too, do not have self-concepts in this sense.

Obviously, the above characterization of the nature and function of self-concept is extremely sketchy, and many important questions remain open: what is the connection between implicit and explicit indexicality? Can there be some intermediate kind of first-personal capacity that is more advanced than the primitive implicit indexicality but does not require a self-concept? A complete investigation of the first-person perspective needs to spell out the details, but for our purpose, however, we only need to have a clear sense that implicit and explicit indexicality are indeed distinct types of phenomena, although both can claim to be first-personal in some sense. With this distinction in hand, we can now address the first challenge of the *de se* skeptics.

2.3 Implicit Indexicality, Action, and Action-Explanation

To recapitulate the story so far: in the first section, I argued that a close reading of Perry's problem of the essential indexical reveals that his fundamental concern is not indexicality per se, but the importance of belief states (or ways of believing) in explaining actions. The examples he uses to illustrate the problem (e.g., the messy shopper case) are important only because they highlight the role of indexical states in actions and action-explanations (**EIS** and **EIS***). I then distinguished between implicit and explicit indexicality. Given this distinction, **EIS** and **EIS*** each have two readings. To show indexical states are not essential or indispensible for (motivating and explaining) actions, then, one would need to make the case for both implicit and explicit indexicality: one would need to show, in other words, that neither implicitly nor explicitly indexical are essentially for (motivating and explaining) actions. However, as we will see shortly, in raising their challenge, *de se* skeptics tend to ignore implicit indexicality all together.

As we have noted, it is far from clear that explicitly indexical thoughts are needed for all actions. Implicit indexicality, on the other hand, seems to be more fundamental: most, if not all, of our actions involve the interaction with things around us, and thus the ability to represent these things from our perspectives and act upon them is indispensible for to get around in the world. Therefore, I suspect, when theorists claim that indexical thoughts are in some sense necessary or essential for action or action-explanation, they have in mind *implicitly* indexical thoughts, rather than *explicitly* indexical thoughts.

Furthermore, it seems that in at least *some* of the puzzle cases that *de se* exceptionalists use to make their case, the fact that the subject has an explicit *de se* belief or thought is inessential. For example, *de se* exceptionalists claim that sometimes, as in the bear-chasing example, two agents have the same (relevant) objective beliefs and desires, but act differently. To account for the difference, they continue, we must invoke the agents' respective *de se* attitudes. But consider a variation of the example: a car suddenly comes around the corner, and you are about to be hit. I am standing on the sidewalk, and seeing the car run into you, I close my eyes. Luckily, you instantly notice the coming car and jump out of the way just in time. Like the bear-chase example, (let us assume) we have the same relevant objective beliefs and desires. Still, we act differently. What explains the difference? It is unlikely that you have the time to form the belief that you are about to be hit by a car (or the desire that you not be hurt by it). And it is unnecessary: what matters, very crudely put, is just that your perceptual system registers that the car is coming (and from a particular direction) and your motor system react appropriately, mediated by some fast, unconscious, and perhaps wired-up mechanisms of fight-or-flight response.

One might think that the low-level states and processes involved are too trivial to be philosophically interesting. Indeed, when they discuss this sort of examples, Cappelen and Dever, and Magidor often emphasize that many behavioral differences associated with indexical and non-indexical attitudes are simply due to physiological differences (so again, the relevant differences show nothing distinctive about indexical attitudes *per se*). The assumption seems to be that if we expand the causal role of (mere?) physiological factors, we would thereby diminish the role of psychological factors, including the first-personal or perspectival states. This may or may not be true, but in the present variant of the bear-chasing example, the important thing here is that the low-level states that are responsible for our different fast reactions are not merely physiological. Rather, they are representational and perspectival: your perceptual system represents the coming car from a particular direction (relative to your body), and thus the resulting motor commands is determined by where you stand

and where the car is coming from, relative to your location (e.g. if you jump to the car, you will be badly injured, even killed). Hence, while we both see the car running towards you, we occupy different spaces and see it from different perspectives. The fact that our behavioral responses to the danger are not typical intentional actions (and more like instinctive reflexes) does not imply that perspectival difference is irrelevant.

In a schematic form, then, my worry is that *de se* skeptics wrongly assume that the causal/explanatory significance of the self is either explicitly representational, e.g., when the self is represented by the agent, or is merely physiological. The previous section can be seen as an attempt to pinpoint the middle ground between these extremes, where the involvement of the self, though implicit, is nevertheless essential to the action-guiding representations. *De se* skeptics ignore this middle ground, and thus conflate explicit and implicit indexicality.

To get a more concrete sense of this conflation, let us now turn to some examples. They are taken from Cappelen and Dever's book, *The Inessential Indexical*; however, the problem they exemplify, I believe, is a general one.

2.3.1 Implict Indexicality and the Cause of Action

Let us begin with the cause of action. Cappelen and Dever discuss a remark in another paper by Perry (1998). While, as I have shown above, his classic paper "The Problem of the Essential Indeixcal" aims to identify and solve a general problem facing the traditional account of belief, here Perry specifically contrasts objective or perspective-independent representations with perspectival representations, and stresses the importance of the later:

> Consider a transaction with a fax machine. To press certain button on it, I have to move my fingers a certain distance and direction from me. It isn't enough to know where the buttons were relative to one another, or where the fax machine was in the building or room. I had to know where these things were relative to me. (1998, 87)

Perry's point seems to be that to use the fax machine, I need to represent the buttons and the fax machine from my perspective, and act upon them from that perspective. To this apparently simple and straightforward point, Cappelen and Dever respond:

> . . . in order to act, our bodies have to be brought into engagement with the world around us. That much is trivial. But what is not trivial is that for that to happen, we must represent ourselves in relation to the objects we engage with . . . On cognitive-representational level, it's just all about the buttons and their objective position in space. Then a bunch of neurons fire and our bodies end up doing the right thing. . . For motivation 2 [the motivation that bodily movement requires indexical thought] to be persuasive, it needs to come with some reason to think that it gives out at exactly the right spot: while *the self is still involved in representation.* (2013, 43; emphasis added)

It is worth noting that Cappelen and Dever's goal in the book is quite ambitious: they contend that perspectival, first-personal or indexical representations (they treat these terms equivalently) have little philosophical significance, and that these representations do not have any important role in action, action-explanation, belief, perception, etc., that many assume they have (2013, 1-3). However, throughout their book, they often start out by citing some examples purporting to show the essentiality of *indexical* or *perspectival* representation, and then move on to argue that indexical expression is not essential for an adequate explanation of the relevant actions.[14] But as should be clear by now, this sort of objection trades on a crucial ambiguity, because "perspectival representation" could mean either explicitly indexical representation or *implicitly* indexical representations. This passage is a case in point. We have seen that a perspectival representation, understood as a representation from the first-person perspective, need not contain a representation of the self. In the fax machine case, it seems that I do need to know the spatial locations of the buttons

[14] Occasionally, they move on to argue that indexical contents are not essential or special in any philosophically significant way. But to say that indexical representations are special does not imply that their contents are special. Perry apparently thinks that indexical thoughts play a distinct role in our actions, but he also says that the content of such thoughts are just de re propositions.

from my perspective, in relation to me (i.e., *implicit* indexicality), but I do not have to represent myself or my body as being in such and such spatial relation with the buttons (i.e., *explicit* indexicality)

Curiously, it is clear from the text that Perry himself uses this example to elucidate what he calls knowledge *concerning* oneself (as opposed to knowledge *about* oneself), which, just like implicit indexciality, does *not* require self-representation.[15] So it seems that Cappelen and Dever not only misconstrue Perry's point here, but also fail to provide a compelling case against the claim that implicit indexicality is necessary for actions.

2.3.2 Implicit Indexicality and Action-Explanation

The next example concerns the explanation of action. According to Cappelen and Dever, the advocates of essential indexicality hold the following:

- **Impersonal Incompleteness Claim (IIC):** Impersonal action rationalizations (IAR) are necessarily incomplete because of a missing indexical component. (2013, 37)

However, they did not explain what they take an "indexical component" to be. If an indexical component is just an indexical expression, then **IIC** is similar to **EIE** (i.e., the explanatory role of indexical expressions cannot be fulfilled by non-indexical expressions). This should immediately raise the equivocation worry we have seen

[15] In the end of the very section, Perry says

> When we perceive how the world is around us and act upon it, we need to judge what distance and direction things stand relative to ourselves. But we do not need to keep track of who it is that we are judging things to be in front of or to the left of, at least as long as we are basing our actions on simple perceptual knowledge. In this case, our knowledge concerns ourselves but need not involve an explicit representation of ourselves. (1998, 89)

earlier: even if Cappelen and Dever manage to refute **IIC** (or **EIE**), they have not yet undermined **EIS** (The explanatory role of indexical states cannot be fulfilled by non-indexical states.), for the latter thesis concerns indexical states in general, not just those explicitly expressed by indexical expressions.

But even if we read **IIC** more charitably, taking "indexical component" to be indexical representations in general (which encompasses *both* explicit and implicit indexicality), there is still another difference between **IIC** and **EIE**: **EIE** says that the role of indexical expressions cannot be fulfilled by non-indexical ones, but it is silent on whether impersonal rationalizations are *necessarily* incomplete. **EIE** only implies that to the extent that indexical expressions do play an explanatory role, they are *irreducible* to, or *irreplaceable* by third-personal expressions. **EIE** is thus compatible with the claim that third-personal explanations of actions can *sometimes* be complete.

I think **EIE** better captures the spirit of essential indexicality that de se exceptionalists have in mind. **IIC** is too strong. It seems that whether an explanation is complete is largely context-sensitive: it often, if not always, depends on the interests and background assumptions of the relevant context. Hence, one might worry that **IIC** is a straw man that no supporters of essential indexicality should accept anyway. But let's bracket this worry and see whether the objection itself withstands scrutiny. Cappelen and Dever object to **IIC** by counterexamples: there are, they contend, impersonal explanations of actions that are nevertheless complete. For reasons mentioned above, I am not particularly keen on defending **IIC**. Still, I strongly doubt that their generalizations are real counterexamples. One of their examples is the following:

- **Useful (U1):** Any criminal engaged in a criminal act who has a belief she would express by "Superman is right around the corner" will, *ceteris paribus*, run. (2013, 46)

They argue that **U1** can be used in a complete explanation of action (say, why a particular criminal runs), and that explanation is indexical free. Now it's true that **U1** does not explicitly use any indexical expression, but it does seem to invoke some indexical or perspectival representation, in the following sense: representing Superman as around the corner is, at least [16] *implicitly* indexical, for the corner is represented *relative to* the subject's own location. That is, the corner is not just a particular corner on an objective, perspective-independent map that the criminal has in her mind, since her own location is clearly implicated, although not explicitly mentioned, when she forms the belief that Superman is around the corner. The corner is the one near the crime scene, which the criminal is aware of, otherwise she won't feel the urgent need to run at all. Therefore, in **U1**, the belief ascribed to the criminal is at least *implicitly* indexical. In other words, when we, as theorists, use **U1** to explain her action, we tacitly assume that the subject's awareness of its surroundings from her own perspective is key to her action: that is, we presuppose the subject's implicitly indexical representations play a critical causal role in her action. Therefore, this is yet another instance of the equivocation between implicit indexicality and explicit indexicality we have encountered before.

2.4 Explicit Indexicality, Opacity and Asymmetry

2.4.1 The Skeptical Challenge

Now let us turn to the role of indexical expressions in action explanation; here, I think, de se skeptics have a stronger case. Although their objection ultimately fails (or so I will argue), engaging with the objection will help us to see more clearly what

[16] One might even plausibly maintain that the criminal cannot think that Superman is around the corner unless he thinks that Superman is around the corner near him, thus entertain a self-concept that is not explicitly expressed in the description of his thought. If so, so much worse for Cappelen and Dever, for **U1** invokes, or at least assumes, explicit indexicality. Additionally, it can be plausibly maintained that "the corner" here is an incomplete description, and it can only be completed by some indexical element (thanks to Dilip Ninan for the suggestion)

is at stake in this debate. The following seems to be the consensus among many *de se* skeptics: in general, beliefs with identical truth-conditions tend to have different cognitive and behavioral significances, and thus different explanatory imports. Therefore, even if in some cases, the explanatory role of indexical expressions cannot be fulfilled by non-indexical expressions, it has nothing to do with indexicals *per se*. For illustration, consider the following example:

Superman Lois Lane knew that Clark Kent was at Times Square. He just called and asked her to join him to watch the New Year's Eve firework show. She declined the invitation, saying that she's tired and would rather stay at home. A few minutes later, however, the live news reported that Superman showed up at Times Square, apparently to celebrate the New Year with the public. Thrilled at the prospect of seeing Superman at the turn of the year, she immediately turned off the television, put on her coat, and rushed to her car.

According to the *de se* skeptics, this is perfectly analogous to the messy shopper case. Lois Lane believed that Clark Kent was at Times Square and that Superman was at Times Square, but only the latter caused her to leave her apartment. If we assume, with Perry and others, that these two beliefs have the same content, i.e., the de re proposition constituted by the superhero and the property of being at Times Square, then what explains her behavior is a change in the way of believing.[17] So we need to invoke Lois Lane's "Superman"-ish way of believing to explain her actions, just like in the messy shopper case, we need to invoke the shopper's indexical way of believing to explain his actions.

[17] I've been following Perry's broadly Russellianism about propositions here. For Fregeans, the "sense" or "mode of presentation" correspond to "Superman" and "Clark Kent" are presumably different, so the beliefs that Clark Kent is at Time Square and that Superman is at Time Square have different contents. But according to de se skeptics, adopting a Fregean position is not helpful here, for presumably, "I" and "the messy shopper" also have different senses, so again there is still nothing particularly problematic about indexicals.

Therefore, *de se* skeptics maintain, the messy shopper case and the like show nothing particularly problematic about the role of indexicals in the action-explanations. According to Cappelen and Dever, the lesson we should draw from these problematic cases is **Indexical Opacity**, which is just an instance of **Generic Opacity**

- **Indexical Opacity:** There's a set of indexicals, I-SET, that cannot be substituted salva veritate in action-explanation contexts by any other expressions.

- **Generic Opacity:** Co-referential referring expressions cannot be substituted *salva veritate* in action-explanation contexts. (2013, 33)[18]

I am inclined to agree that **Indexical Opacity** is one lesson we can draw from the Perry-style examples, and by itself it shows nothing distinctive about indexical thoughts. In fact, I think this point is hardly surprising, especially for Perry himself. As I argued above, his main point in "The Problem of the Essential Indexical" is a general one that that does not single out indexical beliefs as particularly problematic. We often have different concepts that pick out the same entity, or, to put thing in Perry's terms, we often have different ways of thinking about the same entity, and it

[18] Likewise, Evan Tiffany says

> . . . the argument from the inability to substitute *salva actione* for the indexical doesn't reveal anything unique about indexicals. The form of the argument was that if the first belief in each group explains an action, then substitution *salva veritate* for any of the terms in the proposition believed is not guaranteed to be substitution *salva actione*–This result, I suggest, should be neither new nor surprising, for, as was pointed out by Frege and expanded upon by Quine, normal inferential links between propositions in opaque contexts are not preserved. (2000, 40-41)

And Cara Spencer says

> The received view is that indexical belief presents a special problem for the traditional picture of belief, and Perry's arguments show us what this problem is... I deny that the problem has anything special to do with indexical belief narrowly construed. Perry's problem arises quite generally for what is commonly called singular belief. [2007:179-180]

Similar points can be found in Devitt (2013); Magidor (2015) is only natural that this cognitive difference results in behavioral differences. In this regard, Lois Lane's "Superman"-ish and "Clark Kent"-ish way of thinking about the man, and Perry's explicitly indexical way of thinking ("I am making a mess") and third-personal way of thinking ("John Perry is making a mess!") about himself are indeed on a par.

But does it follow that there is *nothing* distinctive about explicitly indexical thoughts in the explanation of actions? I think not. In what follows, I will argue that there is more to the Perry-style examples than the generic opacity of explanation. Specifically, I will argue that, first, **Indexical Opacity** fails to capture a critical difference between the Superman case and the messy shopper case, and second, the best way to characterize the idea that indexical thoughts, as opposed to non-indexical thoughts, are special or even fundamental to our actions, is not in terms of the opacity of explanation, but in terms of what I will call the Explanatory Asymmetry between these two kinds of thoughts.

2.4.2 The Explanatory Force of Beliefs

Let us take a closer look the source of the explanatory force of beliefs (or ways of believing) in these examples. We have seen that, in the **Superman** case, Lane's belief (1) is explanatory of her action, but her belief (2) is not.

(1) Superman is at Time Square.

(2) # Clark Kent is at Times Square.[19]

The reason for this explanatory difference, we have assumed with Perry, has to do with Lane's ways of thinking about Superman/Kent. But how, exactly? What is it

[19] "#" indicates that the sentence or the belief expressed by the sentence is not explanatory in the context.

about these two distinct ways of thinking about the same individual that results in the motivational and explanatory differences?

Obviously, it is not because the term "Superman" somehow carries more motivational weight than the term "Clark Kent". Rather, the reason has to do with the cognitive (and affective) significances the agent herself, i.e., Lois Lane, associates with the man known to her *as* Superman: what properties she ascribes to him, how her conception of him relates to her other thoughts, emotions, plans, etc.[20] Perhaps the reason Lois Lane wants to see Superman is that she regards him as the most admirable man in the world, or her favorite superhero, etc. The motivational force, or the practical relevance, of her belief (1), then, rests on these cognitive and affective associations.[21] On the other hand, the features she associated with the same man, under her "Clark Kent"-ish way of thinking about him, are a rather different set of features (e.g., a meek, bumbling colleague who seems to hide something about himself), which are unable to motivate her to accept his invitation.

Now for the explainers of a particular action, it is not, of course, always necessary to invoke this sort of associations, since, as I granted earlier, action-explanations are context-sensitive. As long as the relevant background information is readily available, simple explanations (e.g., Lois Lane wanted to see Superman and she believed that he was at Times Square) may suffice. The point is that these associations, when revealed,

[20] Note that I am not committed to a Fregean picture of proper names. The issue here is not the semantics of the linguistic expressions, but the psychological significances (both cognitive and affective) that the subject associate with an entity, represented in a certain way, that bear on the actions of the subject.

[21] It might be contended that the explanatory difference between (1) and (2) is due to some simple facts about Lois Lane's desires: e.g., she wants to see Superman, but not Clark Kent. However, this just puts the question a step back: what it is about the "Superman" and "Clark Kent" such that she wants to see the same man conceived of in one way, but not when it is conceived of in another way? Of course, Fodorians might answer by appealing to the syntactic features of SUPERMAN and CLARK KENT in Lane's language of thought, but I have to bypass this issue here because, first, I don't want to be committed to any particular view of mental representations; second, and perhaps more importantly, I doubt that the Fodorian answer, if fleshed out, is really incompatible with my account of the explanatory differences between (1) and (2).

can help to make the agent's actions more intelligible, and, in this particular case, to explicate the motivational and explanatory differences between beliefs (1) and (2) for someone who lacks the relevant background information. The important thing to note here concerns the form of the explication: in principle at least, the relevant associations may either be non-indexical ("the most admirable man in the world") or indexical ("my favorite superhero"), depending on how the object is conceived by the agent herself.

Furthermore, we might be justified in attributing to the agent propositional attitudes that make explicit these associations. We might, for example, ascribe beliefs expressed by (3) and (4) to Lois Lane when she comes to believe that Superman is at Times Square:

(3) One of the most admirable men is at Times Square.

(4) My favorite superhero is at Times Square.

Unlike (1), these two beliefs are not de re. But for Lois Lane, the motivational force of the beliefs (3) or (4) may well be what underwrites the motivational force of her belief (1). Indeed, to someone who lacks the relevant background information, beliefs such as (3) and (4) can be more informative and illuminating.

These discussions about Superman have revealed at least three things: first, we can account for the explanatory force of a non-indexical beliefs, such as (1), by appealing to the cognitive (and/or affective) significance associated with Lois Lane's "Superman"-ish way of thinking about Superman/Kent; second, depending on the actual situations, these associations can be fleshed out in either indexical ("my favorite superhero") or non-indexical ("an admirable man") terms; third, we can attribute to the agent certain further propositional attitudes, such as those expressed by (3) and (4), that make explicit these associations, and thus make explicit the underlying explanatory force of the original explanatory belief (e.g., the belief expressed by (1)).

On the other hand, however, a similar account of the explanatory forces of indexical beliefs does not seem readily available, at least not in a similar way. We can see this by reconsidering the messy-shopper example. Assuming that, while he was searching for the messy shopper, Perry had believed that α ("α" is an indexical-free expression that designates himself, such as "John Perry" or a definite description that uniquely picks him out) was responsible for the sugar on the floor; but since he didn't know that he was α, it was only when he acquired the relevant indexical belief ("I am making a mess!") that he stopped to rearrange the torn sack. Therefore, of these two beliefs, only one can informatively explain the change of behavior:

(5) I am making a mess.

(6) α is making a mess.

So far the case is analogous to the Superman/Kent case above. However, now suppose we ask parallel questions: why does belief (5), as opposed to belief (6), change Perry's course of action, given that "α" and "I" in this context refer to the same individual? In other words, why is his indexical way of thinking about himself, as opposed to the way of thinking corresponding to "α", explanatory of his action? What are the special cognitive and affective significances he associates with "I" that cause and explain his action?

Unlike in the previous example, these questions seem hardly intelligible. Someone unfamiliar with the Superman story may sensibly wonder about the cognitive and effective significances Lane associates with "Superman" [22] (thus finds her belief expressed by (1) not explanatory of her action), but presumably, the significances one associates with the first-person pronoun are largely the same: "I" as uttered by

[22] Or more precisely: the cognitive and effective significances Lane associates with Superman/Kent known to her as Superman, or under the 'guise' of Superman.

S refers to S, who is the subject of S's thoughts and experiences, the agent of S's actions, etc.

But perhaps there is a richer and more substantive dimension to one's explicit "I"-thoughts. Perhaps, for example, the reason that belief (5) changes Perry's action is due to his self-conception: he self-ascribes the property of being a gentleman, and he started to rearrange torn sack because that IS what a gentleman would do in that situation. So, we might say, this is why his belief (5), but not his belief (6), explains his action. The term "α", after all, is not directly connected with his self-conception.

However, this sort of explanation is incomplete unless we assume that Perry takes himself to be the agent that has that property (i.e., is a gentleman). That is, if we try to explicate the cognitive significance of "I" by linking it with a property P the agent self-ascribes, we still have to retain, at least tacitly, an indexical component (e.g., she takes herself to be P) in order to fully explain her action. Unlike the **Superman** case, a purely objective, indexical-free explanation cannot do the job.

Here is another way to put the contrast of this case with **Superman**. When an agent comes to believe an "old" proposition (i.e., a proposition she has believed already) in a new and non-indexical way, the motivational and explanatory force of this new belief, or new way of believing, can in turn be explained by the cognitive and affective significances the agent herself associates with the "res" (i.e., the object the belief is about) under this way of thinking. Moreover, depending on the context, these significances can be cashed out in either indexical or non-indexical terms. So, for example, under Lois Lane's "Superman"-ish way of thinking, Superman/Kent is one of the most admirable men in the world, or is her favorite hero, etc. Consequently, we can ascribe to her new indexical or non-indexical beliefs, such as beliefs (3) and (4), which make explicit these associations in order to clarify or further explain her action.

On the other hand, however, when an agent comes to believe an "old" proposition in a new and indexical way (e.g., when Perry learned that he himself is making a mess), the motivational and explanatory force of the new way of believing cannot be explained in exclusively non-indexical terms. Relatedly, it is hard to see what new non-indexical beliefs alone can contribute to a better and more elaborate explanation of her action, in the way that Lane's belief (3) or (4) helps us to understand her action. This contrast, then, is what I call **Explanatory Asymmetry**:

- **Explanatory Asymmetry:** The explanatory force of non-indexical beliefs can be elucidated in non-indexical (and indexical) terms, but the explanatory force of indexical beliefs cannot be elucidated in non-indexical terms.

This, I think, is what many have in mind when they maintain that indexicals are in some sense irreplaceable (even if it is not necessary) in the explanation of (at least some) actions, and this is indeed what **EIE**, repeated below, meant to capture:

- **Indexical Expression Essentiality (EIE):** The explanatory role of indexical expressions cannot be fulfilled by non-indexical expressions.

Now we are in position to step back and take stock. First of all, **EIE** is a natural way to characterize essentiality of indexical expressions. As I take it, even de se skeptics can agree with **EIE**. However, on their view, **EIE** just amounts to **Indexical Opacity**, and shows nothing particularly problematic about indexicals or indexical thoughts, for it is an instance of a general problem, say, **Generic Opacity**. But now **Explanatory Asymmetry** offers us a second, and more interesting way of interpreting EIE and the contrast between indexical and non-indexical thoughts. On this interpretation, the phenomenon that inspires the classic papers of Perry, Lewis and others is not merely an instance of a general problem of opacity.

Unlike (Indexical or Generic) opacity, Explanatory Asymmetry is not primarily about whether certain expressions can be substituted salva veritate in action-explanations. Rather, it concerns a prior question: why are certain beliefs explanatory, while others that have the same *de re* content are not, in the first place? Indeed, as we have seen, substituting an expression with another sometimes can preserve, or even enhance, the explanatory force of the agent's beliefs. In the Lane/Superman example, what underwrites the explanatory force of Lane's belief (1) is probably something like (3) and (4), so if we substitute "Superman" in (1) with "one of the most admirable men in the world", we can explain why her belief (1) is explanatory of her action. What Explanatory Asymmetry suggests, on the other hand, is that indexical and non-indexical thoughts have, at least sometimes, different types of explanatory imports.

Therefore, the distinctiveness of indexicals in action-explanations does not consists in the fact that indexical cannot be substituted *salva veritate* in action-explanation. Although the messy shopper case can be seen as an instance of **Generic Opacity** (or Frege's problem), and in this respect, it is analogous to the **Superman** case, it also shows something more, namely, the **Explanatory Asymmetry** between indexical and non-indexical thoughts. This asymmetry is, I suspect, what underlies the feeling that indexical thoughts are in some sense more fundamental or essential (than non-indexical thoughts) to our actions.

2.5 Why Asymmetry: Some Concluding Thoughts

The recent debates surrounding indexical or de se attitudes suggest two possible ways to understand the so-called "essentiality" of such attitudes. The first is to construe it as a kind of necessity. Advocates of this type of "essential indexicality" hold that indexical attitudes are, in some sense, necessary for intentional actions. As noted above, I am skeptical about essential indexicality in this sense, at least about the essentiality of explicitly indexical thoughts. The second, which is closer to Perry's original meaning of the term, is to construe it as a kind of irreplaceability. As should be clear by now, I am more sympathetic with this type of essential indexicality, and the arguments in the preceding section can be seen as an attempt to clarify its nature and significance.

Suppose my arguments so far are on the right track, a natural question to ask at this point is: why do indexical and non-indexical thoughts exhibit this asymmetry? Unfortunately, it is impossible to fully address this question within the scope of this chapter. So let me conclude with some remarks that, I hope, at least point to the general direction of an adequate answer to this question.

First, the discussions in the previous section suggest that it is misleading to take one's self-concept as merely another (perspective-free) concept for oneself. Interestingly, this echoes a point that Anscombe made decades ago: in some non-trivial sense, the first person pronoun is not a name for oneself (1975). For if it is, then the **Superman** and the messy shopper case would be no different, and there shouldn't be any asymmetry between indexical and non-indexical thoughts. Hence, the difference between indexical and non-indexical thoughts is, perhaps, a difference between perspectives, not (just) a difference between terms and their significance.

Second, the distinct character of the first person can be traced to the fact that our competent uses of it expresses our self-awareness, and the differences between indexical thought and non-indexical thoughts in our actions reflect the differences between our self-awareness and other-awareness in our deliberation, intentional actions, etc. In a way this should not be very surprising: as embodied agents who interact with the world from our own perspectives, our self-awareness is tied to our actions in a way that our other-awareness is not. However, it is a daunting task to spell out the exact role of self-awareness in agency. As noted above, I do not think that we always need to invoke an agent's explicitly indexical thoughts about herself to explain her actions. But in cases where we do (as in the messy shopper case), this is presumably because the agent's self-awareness is crucial for her deliberation, plan, and decision, and correspondingly, for us to make sense of the rationale behind her action. After all, the use of "I" in one's utterance, or employment of self-concepts in thinking and deliberation, often reflects not only the agent's having a first-person perspective, but also her agent's awareness of that perspective. If this is the case, then any adequate account of the Explanatory Asymmetry should include a discussion of self-concepts, self-awareness, and their roles in human agency.[23]

Third, and perhaps more importantly, knowing the indexical attitudes of an agent typically enables us to directly appreciate the agent's point of view on her own action (i.e., her self-understanding of her own action), and seeing things from her point of view often renders the agent's action more intelligible than when we are only aware her attitudes characterized in a third-person way. This is in part because knowledge about the agent indexical beliefs can put us in a better position to empathetically understand her reasons for action (to put ourselves in her shoes, so to speak), and to use our own cognitive resources to make sense of her action and the underlying practical reasons.[24] Lane's non-indexical beliefs (Kent-beliefs, or even Superman-beliefs) may have no explanatory import for one unless one is aware of the respective practical significances of the referent, under these guises, for her—that is, unless one is able to understand the action and underlying reasons from the Lane's point of view. In the absence of certain crucial background information, nonindexical beliefs sometimes fail to illuminate the agent's self-understanding of her action, which in turn can impedes an observers' understanding of the action. By contrast, if the agent performs an action in part because of her indexical or de se attitudes, and we are informed about the relevant attitudes, then it is much easier to see their relevance to the action: after all, as self-aware animals and competent speakers, we know what it

[23] Again, by this I do not imply that all exercises of human agency require self-awareness (see Doris (2015) for a rebuttal of this view), although some more sophisticated or advanced forms of human agency may at least presupposes the possession of self-concepts and capacity of self-awareness.

[24] Perhaps in a way not unlike mindreading by simulation Goldman (2006)

is to entertain an indexical thought, and thus are in a better position to know or infer about the potential relevance of the agent's indexical attitudes to her action. This is because, as noted earlier, while different speakers may associate different features with a proper name, the basic and practically relevant properties of the self, conceived as the self, is very much the same (e.g., the agent of one's own action, the subject of one's own thoughts and experiences, etc.) Likewise, for someone who wonders why a customer in the supermarket suddenly stopped his chart to rearrange the stack in it, the indexical belief ascription "He believes he himself is making a mess" is typically more adequate and informative than the corresponding non-indexical belief ascription "He believes John Perry is making a mess" (the latter could be adequate and informative, though a bit bizarre, provided that one knows that the agent also believes he is John Perry. In that case, the non-indexical belief ascription can provide information, albeit indirectly, about Perry's self-understanding of his own action.) Now it is an open question whether all informative action-explanations should be able to induce empathetic understanding of the actions, but for our purposes, it only needs to be admitted that for the explainer, knowledge about the agent's indexical beliefs offers a convenient and straightforward way to empathetically understand the action at issue, and therefore, it typically helps to render the action intelligible, or more intelligible, than the mere knowledge about the agent's perspective-independent, third-person beliefs, other things being equal.[25] In short, indexical beliefs provide a direct access to agent's first-person perspective on her action, e.g., her deliberations, plans, etc., and thus are often illuminating in a way that non-indexical beliefs are not. This difference, then, might be another underlying source of **Explanatory Asymmetry**.

[25] This, of course, does not imply that knowing an agent's indexical attitudes always suffice to render the action intelligible. After all, one's indexical attitudes might be too unfamiliar or idiosyncratic for others to make sense of.

CHAPTER 3
THE PSYCHOLOGY OF SELF-RECOGNITION

In the previous chapter, we have distinguished two kinds of first-personal thought. In this chapter, we will take a closer look at explicit indexicality and the concept of the self as such by investigating the nature of the self-recognition. Self-recognition presupposes a concept of the self, but the self is a multi-facet entity: among other things, the self is the subject of experiences and thoughts; it is the agent of intentional actions; it is a bodily entity that exists in physical space; it is a continuous entity that persists in time, etc. Many of the different aspects of the self are functionally and neurologically dissociable (Klein and Gangi, 2010) An exhaustive study of the many dimensions of the self is thus beyond the scope of this thesis. Instead, this chapter focuses on the psychological structures and mechanisms that underwrite two forms of self-recognition that have figured prominently in recent psychological research on the self: mirror self-recognition and temporal self-recognition.

3.1 Mirror Self-Recognition

Although we rarely think of it this way, recognizing oneself in a mirror is actually a remarkable capacity that we share with only a few other species. For many comparative and developmental psychologist, it is a capacity closely associated with self-awareness. Indeed, the famous mirror mark test, independently developed by Gallup (1970) and Amsterdam (1972), is often regarded as the litmus test for self-awareness in non-human animals and young children (Courage et al., 2004; Parker et al., 2006)

Although there are many variants of the test, the basic procedure is rather straightforward: first, an experimenter unobtrusively places a marker on the subject, on some part of the body that is not directly visible (e.g., the forehead); then the subject is led to a mirror to see its mirror image. The subject's voluntary reactions to its mirror image are recorded, and some particular kinds of behavior are taken to be the evidence for mirror self-recognition (MSR). For example, if the subject attempts to wipe off the mark while or after observing its image in the mirror, she is regarded as exhibiting MSR, and thus have passed the test. Passing the test in turn indicates that the subject has at least some form of self-awareness. Other behaviors can be even more telling. In his classic 1970 study, Gallup found that chimpanzees not only started to touch the marked region once they see themselves in the mirror, but also used the mirror to inspect and explore areas of their body that they otherwise cannot see.

Over the years, Gallup's study has been replicated many times, and the mirror test, or some version of it, has been performed on other species. The evidence for MSR has been found in great apes (with the possible exception of gorillas); however, most other nonhuman primates that have been studied failed the test [1], and outside the primate order, there have been only a few instances of success (Reiss and Marino, 2001; Plotnik et al., 2006). In the case of human children, Amsterdam's initial finding was that children as young as 20 months can recognize their mirror images. Today the general consensus among developmental psychologists is that from 18 to 24 months of age, most children are capable of MSR (Lewis and Brooks-Gunn, 1979; Moore, 2006; Rochat, 2009). Moreover, it is also generally agreed this development marks the onset of awareness of the self as such in humans.

[1] For review, see Anderson and Gallup Jr (1999); it is perhaps worth noting that recently there are some preliminary evidence that some gorillas and monkeys exhibit MSR after elaborate training (Posada and Colell, 2007; Roma et al., 2007; Chang et al., 2015)

On the other hand, however, there is still some controversies with respect to what *kind* of self-awareness is measured by the mirror test. Gallup himself has offered a very rich interpretation of MSR, according to which passing the mirror test indicates a variety of complex mental abilities, such as introspection and mindreading (Gallup, 1998; Gallup et al., 2002). His interpretation of the experiment and its significance, however, has been subject to vigorous and, I think, decisive critiques (Mitchell, 1997a,b; Morin, 2011; Suddendorf and Butler, 2013). The main criticism is that mirror self-recognition, introspection, and theory of mind are conceptually distinct and functionally dissociable. For example, many autistic children can pass the mirror test; but they fail the standard ToM tasks, and also show serious difficulty in reflecting on their subjective experiences (Mitchell, 1993; Hobson, 1995).

Critics of Gallup generally favor a more moderate hypothesis, according to which the kind of self-awareness involved in MSR is only the awareness of the *physical* or *bodily* self: that is, the awareness of one's body on the basis of proprioceptive, kinesthetic and somasthetic information. As we will see shortly, this claim is still subject to different interpretations, but at the most general level, the idea is that the central capacity involved in MSR is the capacity to form and integrate two representations of the self: to pass the mirror test, one would at least need (1) an explicit representation of the bodily self, and (2) the ability relate the bodily self-representation to the image one sees in the mirror in such a way that the visual information can be integrated with the proprioceptive, kinesthetic and somasthetic information so as to guide one's action. A version of this hypothesis, put forward by the psychologist Robert Mitchell, is the *Kinesthetic-Visual Matching theory.* According to his theory:

> ...passing the mark test is a result of kinesthetic-visual matching, that is, a capacity for matching between the kinesthetic, proprioceptive, and somasthetic sensations of one's own body's position and one's own bodily feeling, and visual images of one's own body and others' bodies. In this view, the organism's ability for kinesthetic-visual matching is an explanation for its passing the mark test.(Mitchell, 1997a, 41).

Another important researcher in the field, Daniel Povinelli (Povinelli, 1995, 2001; Barth et al., 2005) embraces the same basic idea.In particular, Povinelli and his colleagues emphasize the role of the explicit representation of the body in MSR:

> Once an organism can hold in mind a representation of the current state of its body, it is in the position to begin to form explicit relations among objects of perception (e.g., the image in a mirror) and the body image. One such relation that seems critical to exhibit the patterns of behavior that are taken to be criterial for self-recognition in mirrors, is an equivalence relation: that thing (image in the mirror) is equivalent to my body.(Barth et al., 2005, 21)

The explanation of MSR along these lines is intuitively plausible, and its intuitiveness, perhaps, explains its popularity among contemporary psychologists. However, from a theoretical perspective, it remains incomplete: it is unclear what the"equivalence" or "matching" relation amounts too. Here things become more obscure, partly because the theorists who endorse a broadly kinesthetic-visual matching theory rarely spell out the details about the nature of the relation and the cognitive mechanisms or processes subserve it. As a result, it is sometimes even unclear that they really mean the same thing.

Broadly speaking, there are two ways of interpreting the basic idea of the kinesthetic-visual matching story and its cousins. On the first interpretation, one is able to recognize her mirror image because one detects the perceptual and cross-modal *similarity* or *correspondence* between one's bodily self-image and the the image in the mirror, and especially the movement synchrony and contingency. When different sensory modalities are recruited in MSR, there might be enough cues, such as cross-modal information (the structure, shape, of the body etc.) and temporal synchrony of the movements (e.g., the movements one feels from the inside, and the movements of the mirror image one sees), that enable the agent to match one source of information with another (e.g., proprioceptive and visual)—that is, to regard them as similar to, or correspond to, each other. Once the similarity or correspondence relation is established, the agent can use the visual image in the mirror as a suggestive medium to guide its self-exploratory actions. Note that, on this interpretation, the agent who recognizes her mirror image does not have to be committed to, either explicitly or implicitly, the *identity* between one's own body and the body one sees in the mirror. That is, the "equivalence“ relation the undergirds MSR is not a relation of identity, but a relation of resemblance or correspondence. I will hence call this the *correspondence* account of MSR.

On the second and, I think, a cognitively more demanding, interpretation, MSR is fundamentally (or in any case relies on) a form of *self-identification*. That is, when one recognizes oneself in a mirror, one not only becomes aware of the correspondence between the somasthetic/proprioceptive stimuli and the visual stimuli, one also thereby *identifies* oneself with the person in the mirror *as* one and the same thing (i.e., numerical identity). Some remarks by Mitchell seem to support this reading:

> the organism matches its kinesthetic sensations to its visual image in the mirror, and therefore predicts that, because these are the same, the mark on the nose of the image in the mirror indicates a mark on its own nose. (Mitchell, 1997a, 41)

For Mitchell, then, realizing the resemblance or correspondence relation between one's body image and the mirror reflection is apparently insufficient for MSR—a further step, which we might call *self-identification*, is needed to pass the mirror test. Admittedly, this seems to be the more intuitive and straightforward interpretation of MSR: what else is self-recognition, if not the recognition that the mirror image *is* an image of oneself? For this reason, when we recognize ourselves in a mirror, the recognition can be simply expressed as "That's me!"

In the developmental psychological literature this intuitive idea is typically fleshed out in terms of *reference* or *representation*: on this construal, MSR is the recognition that the mirror image is an image *of* the self.[2] Therefore, this apparently very attractive interpretation of MSR presupposes that the subjects who display MSR understand, at some level, *representation*, or more specifically, the representational property of the mirror (e.g., the image in the mirror refers to things "outside" the mirror). Henceforth, I will call this interpretation of MSR the *reference* account. The reference account of MSR attributes to young children who passed the mirror test the ability to recognize their mirror image as the reflection of themselves. The psychologist Philippe Rochat (Rochat, 2009) clearly endorses this account when he says:

> . . . individuals [who pass the mirror test] demonstrate the ability to refer to the specular image as referring to their own body. In other words, they refer the silhouette they see reflected in the mirror to precise regions of their own body they cannot see directly (e.g., their forehead). This would be impossible without some kind of a representation of the own body that is mapped onto what is seen in the mirror. Therefore, this behavior indicates that the individual sees the mirror reflection *as standing for* this representation. It is identified *as referring to* the body experienced and represented from within, not anybody else's. Identity is used here in the literal, dictionary sense of "recognizing the condition of being oneself, not another" . . . (Rochat 2009, 98; my emphasis)

It is worth noting that, like Mitchell, here Rochat moves from matching of two representations ("mapped onto") to identifying with the mirror reflection: the *referent* of mirror image is identified with physical self. Presumably, then, he also thinks that mere similarity or correspondence relation is not sufficient for passing the test. But why so? Is the move from resemblance or correspondence to reference so natural and smooth that one always implies the other?

[2] Hence, to say that the subject identifies the image as herself does not mean that subject somehow believes that the image she sees in the mirror is a real person, who is numerically identical with herself. Those who pass the mirror test do not confuse mirror reflections with reality. Besides, some evidence indicate that children do not fully understand numerical identity until they are at least 3 years old (Perner et al., 2011), more than a year after they could reliably pass the mirror test.

To be sure, it is true that many iconic symbols, e.g., pictures, scale models, etc., represent objects at least partly in virtue of the resemblance relation they bear to the represented objects. Hence, advocates of the reference account of MSR can (and do) agree that subjects who pass the mirror test are able to recognize the relevant resemblance or correspondence relation between their bodies and the mirror images. But, as the quoted passages above seem to indicate, for them that this is only one step toward the representational understanding of the mirror images. Moreover, they seem to believe that the inference from correspondence to reference is utterly unproblematic. As noted above, although Mitchell labels his theory the visual-kinesthetic matching theory (and "matching" sounds very much like "correspondence"), it seems ultimately that what accounts for MSR is the recognition of the relevant *reference* relation, as opposed to the *correspondence* relation, between one's body and the mirror image.[3] So apparently he has in mind a richer conception of "matching" than merely connecting two perceptual stimuli by their perceptual similarity or correspondence. However, the question is that, again, it is unclear why recognizing the resemblance and correspondence is not itself sufficient for MSR. It is also not obvious that once a subject recognizes the resemblance or correspondence relation, she would thereby infer that the visual image *refers* to her body that she experiences from the inside. Hence, in the next section, I will raise some considerations against the reference account of MSR.

3.2 Against the reference account

In the previous section, I distinguished two explanations of MSR that are often lumped together in the literature. To adjudicate between them, it is important to keep two questions separate: 1. What does cognitive capacities does the mirror test

[3] See also his (1997b, 31)

measure? 2. What thoughts or judgements attributed normal human adults best express their MSR? The reason for separating them is that, although it might seem natural to attribute certain thoughts or judgements to adults when they recognize themselves in a mirror, it does not follow that the ability to make such judgements is necessary for, or essential to, MSR. That is, it does not follow that young children or nonhuman animals who pass the mirror test can form these judgements. After all, our question is what cognitive capacity enable them to pass the mirror test, not what *we* are able to judge or infer when we see ourselves in front of a mirror.

In this section, I will raise three considerations against the reference account, and for the correspondence account, of MSR: one methodological, and two empirical. The methodological case in favor of the correspondence account of MSR is relatively straightforward. First of all, representing the correspondence relation is less cognitively demanding than representing the relevant referential relation in the mirror test. The latter, but not the former, requires that the subjects have some understanding of representation, e.g., the mirror image *refers to* the self. If both can explain MSR, as I argued earlier, there is no need to posit that young children have this more advanced ability, unless, of course, there are other empirical data that strongly suggest that the young children do in fact have this ability. However, as I will show shortly, the empirical evidence actually points toward the opposite direction. The second methodological point is closely related to the first: the correspondence account is theoretically more parsimonious. As we have seen, many advocates of the reference account have in effect already conceded that the children who pass the test are able to recognize the correspondence between their mirror reflections and their bodily self-images. But they maintain that, *in addition*, these children are also able to infer on this basis that the mirror images *refer to* themselves. Since the correspondence account does not require this additional capacity, it is more parsimonious. Other things being equal, we should accept the more parsimonious account.

But perhaps other things are *not* equal. Perhaps, that is, the reference account does have other important explanatory advantages over the correspondence account. One such advantage might be the unifying power of the reference account: while the MSR does not, strictly speaking, require any understanding of the representational property of the mirror images, perhaps other behaviors of young children may suggest that, by the time they exhibit MSR, they already can understand between the representational relation between one thing and another. If so, then the ability is not too demanding for them after all. What's more, now the reference account can provide an unified explanation of these behaviors and MSR. The unifying power would then be a reason to favor the reference account.

Therefore, the appeal to parsimony, by itself, hardly settles the issue. For the methodological considerations to have more than *prima facie* force, then, we would need substantive reasons for thinking that the additional postulation of the reference account is not only redundant, but also unwarranted. So in the remainder of this section, I will present just such evidences. While it may not be conclusive, it does raise serious challenges to the reference account, and, as a result, shift the burden of proof to its supporters. The evidence comes from neuropathology and developmental psychology, respectively. To preview: there are pathological cases where the subjects can pass the mirror test (or perform self-directed acts in front of a mirror) but they persistently deny that the mirror images they see are images of them. The second, and in my opinion even more important, set of evidence comes from developmental psychology: the age at which children start to understand and actively use representations is much latter than the age when they can pass the mirror test. However, this evidence, joined with methodological considerations mentioned above, constitute a strong reason against the reference account.

3.2.1 Mirrored-Self *Misidentification*

Psychiatrists and cognitive neuroscientists have recently discovered an unusual delusion, sometimes called "mirror sign delusion" in the literature (Phillips, 1996; Breen et al., 2001). The delusion is characterized by the inability to recognize one's own image in the mirror, even though the patient's ability to recognize others in the mirror usually remains intact. An actual case might help to bring about the surprising and puzzling features of this delusion. Phillips (1996) described a patient EF, who suffers from global dementia

> She was unable to recognize her reflection in the mirror, calling her reflection "my friend". Although she used the mirror when washing and grooming, she continued to maintain that it was her "friend" whom she was viewing in the mirror and who was also engaging in washing and grooming routines. She became anxious when unable to look in a mirror, but noticeable reassured when able to see her reflection once again. When asked, she did not think that her "friend" looked identical to her but did agree that here was some resemblance. She was unable to give any biographical information regarding her friend.(156)

Although EF apparently misidentified her reflection in the mirror, she was able recognize her possessions and her room as her own. In addition, she referred to herself as "Nellie", the name EF was called by her family, and was able to recognize old photographs of herself ("that's Nellie"), although she was less certain about more recent photographs of herself. Hence, despite the fact that she was unable to recognize her mirror reflection, it seems EF still preserved some concept of herself. Importantly, she was able to use mirror to guide her self-directed behavior, such as combing and washing. These tasks are equivalent to reaching for the hidden marker in the mirror test. Ramachandran (2007) also reported that some patients, who insisted that their reflections in the mirror were "someone else", were nevertheless able to pass the mirror test.

Some other patients with this delusion are documented in Breen et al. (2001) (see also Breen et al. 2000). Unlike EF, these patients demonstrated intact autobiographical *semantic* memory and semantic knowledge about the mirror. One patient, TH, seemed to have what is known as “mirror agnosia” (Ramachandran et al., 1997), a condition of confusing a mirror reflection as a real object. TH thus exhibited a curious dissociation between the semantic knowledge about mirrors and his ability to interact with mirror reflections. In one experiment, he looked into a mirror and saw an object being held behind him, and when he was asked to take the object in his hand, he “reached towards the mirror on each occasion, scratching on the mirror surface or attempting to reach into or behind the mirror, instead of correctly reaching behind his shoulder to take the object.” (Breen et al., 2001, 248)

It is hypothesized that TH’s mirror agnosia contributed to his mirror sign delusion. However, another patient in the study, FE, did not have mirror agnosia, although he had facial processing deficits. Like TH, he insisted that his reflection was not him, although he admitted that they look similar. He continued to shave before a mirror. According to (Breen et al., 2000), when FE was asked whether the person in the mirror was bald, “FE tilted his head forward so that the top of his head would be visible in his reflected image in the mirror.” (85)

Although the delusion is usually associated with right hemispheric dysfunction and frontal damage, the cause of the symptoms is likely to be multifarious. For our purpose, these studies have shown that, as a matter of fact, subjects who can pass the mirror test (or its equivalents) may nevertheless sincerely and persistently deny that the reflection is identical with, or is an representation of, him or herself. This is *prima facie* evidence that the realization that one’s mirror reflection *refers* to oneself is not necessary for passing the mirror test. Hence, these cases cast doubt on the key assumption of the reference account, i.e., recognizing the reflection *as* (referring to) oneself is required to self-directed actions guided by mirror. Moreover, they remind us that the inference from the observed similarity and correspondence between one’s mirror image and oneself to the recognition that the image refers to oneself is by no means trivial or automatic, even for otherwise rational human adults.

Of course, the phenomenon of mirror self-misidentification is as puzzling as it is intriguing, and for all that has been reported, it is still possible that, while the patients verbally deny that their mirror reflections are them, at some level the patients still tacitly believe that the reflection *is* or *refers to* them. Perhaps, for example, the formation of reportable explicit beliefs is disrupted by influences down the stream, e.g., certain affective or cognitive disruptions, resulting in the bizarre confabulation. A stronger evidence, then, would have to show that even at the tacit level, it is unlikely that subjects who can pass the mirror test necessarily have the capacity required by the reference account. I believe evidence of this sort does indeed exist, and this is the subject to which we now turn.

3.2.2 Young Children's Symbolic Understanding

According to the reference account of MSR, children as young as 1.5 to 2 years old are able to pass the mirror test, largely because they regard the reflections as *referring* to themselves. Hence, this account assumes that these children understand, at least implicitly, the representational property of mirror reflections, i.e., that the reflections not only have properties of their own, but also *represent* other things. In the words of psychologist Judy DeLoache, they achieve *dual representation* (2004; 2011): the ability to represent both the mirror-image *and* the abstract, representational relation the image bears to their own body.

Achieving dual representation, however, is a difficult challenge for young children. DeLoache's important work on children's understanding of symbolic artifacts (e.g. scale models, pictures, etc) nicely illustrates this. In the classic object-retrieval task (DeLoache, 1987), children were asked to use a scale model to find an object in a large room. The model was a highly-realistic replica of the room. It contained miniature furnitures that looked very similar to the actual furnitures in the room, and were placed at parallel spatial locations. The children received extensive orientation in which the experimenter described and demonstrated the relation between the room and the model; they were also asked to take each of the miniature furniture to the room and to compare it to its larger counterpart. In the symbolic retrieval stage, the children first observed an experimenter hide a miniature toy in the scale model, and then were told that the larger version of the toy was hidden in the corresponding place in the room itself. Then they were taken to the larger room to search for the toy. After that, they returned to the model and were asked to find the miniature toy that was hidden earlier. Both 2.5 year-olds and 3-years olds were very successful in finding the miniature toy, indicating that they remembered where the toy was originally hidden, and were also motivated to find it. However, only the 3-year-olds were equally successful in retrieving the target toy in the real room (75% passed the test). It seems, then, the 2.5-years-olds were unable to use the scale model as a source of information for retrieving the larger toy, presumably because they failed to appreciate and exploit the model(symbol)-room(referent) relation.

Similar patterns of behavior have been observed since DeLoache's original study (for a review of this literature, see DeLoache 2011; Callaghan and Corbit 2014). One particularly interesting finding is that, when the status of the model as an independent object is reduced, children's performance improves. For example, when the model is placed behind a window, so that children cannot directly manipulate it, more 2.5-year-olds pass the task (DeLoache, 2000). Their performance also improves when they are led to believe that the scale model is an actual room that has been shrunk by a "shrinking machine". This improvement makes sense because when children treat the model as the actual room, the need for dual representation is eliminated, and as a result, the cognitive demand is reduced. According to DeLoache, it is generally difficult for young children to hold in mind two representations of a single entity. And with regard to symbolic understanding in particular, she claims:

> . . . the concrete features of a symbolic artifact can interfere with young children's ability to notice its relation to what it stands for. A realistic scale model like those used in our research is a highly salient, attractive, interesting object in and of itself. It affords and even invites direct physical activity–playing with the items of furniture contained in it, for example. This makes it hard for young children to treat it as standing for something other than itself. (2000, 321)

DeLoache's hypothesis predicts that if the symbol is less physically salient, then children will be in a better position to focus on its referential properties. This is consistent with the results of the "shrinking room" test mentioned above, as well as other variants of the retrieval task (e.g., that models placed behind a window and beyond the child's reach). Moreover, the object properties of pictures and videos are less salient and less interesting than that of scale models: children can (and love to) play with scale models, but there's not much they can to about a picture or a screen other than just looking at or pointing to it. So, presumably, young children would perform better in the retrieval tasks when the medium of information are pictures or videos. This prediction is borne out as well: it is found that 2.5-year-olds successfully pass versions of the retrieval tasks that use 2-D pictures instead of realistic, 3-D models (DeLoache and Burns, 1994). Likewise, similar patterns of performance have been observed in video versions of the tasks (Schmitt and Anderson, 2002; Troseth et al., 2006). For our purpose, it is important to note that the performance of 2-year-olds is still very poor in these tasks, even with multiple trials.

Now one might reasonably question whether the young children who pass the test genuinely have a concept of the abstract *referential* relation. After all, all these "symbols" (scale models, pictures, videos) are iconic. It is thus always possible that, instead of understanding their referential property, the children merely rely on their *similarity* to the actual room. That is, to pass this test, the children only need to exploit the similarity or correspondence relation between the symbols and the actual room; they do not have to take the former to be *referring to* or *standing for* the latter. Against this important worry, there is evidence that some 2.5-year-olds can appreciate the similarity between the actual room and the model (e.g., they can readily match a toy furniture in the model with the corresponding actual furniture in the room), but they failed the retrieval test nonetheless (DeLoache, 2011). Therefore, it seems that appreciating the similarity or correspondence is not sufficient for successful performance in the task. [4]

As noted earlier, the reference account implies that children who are capable of MSR achieve dual representation: they not only represent their reflections in the mirror, but also represent the abstract referential relation between the reflections and themselves—that is, the children are able to treat their mirror reflections as referential or symbolic. Hence, in this respect the mirror test and the symbolic retrieval test are structurally similar: both require the use of symbolically mediated information in one's exploration and search. It is thus a little surprising that there is little, if any, communication between these two research paradigms. This does not mean, of course, there is not substantial difference between the tasks. For example, the referential relation between models (and pictures) and the objects is based exclusively on their perceptual similarities, but mirror reflections also carry an additional temporal and dynamic cue: the temporal contingency of the perceived movements of the reflection and one's proprioceptive feedback. No such contingency is available in the symbolic retrieval test.

Now defenders of the reference account of MSR might insist that for this very reason, the mirror test is easier, and the success of younger children in this test

[4] At any rate, this worry poses no threat to my defence of the correspondence account of MSR, since on that account, appreciating the relevant resemblance and correspondence between the mirror reflection and oneself is sufficient for passing the mirror test.

suggests that before their second year, they have already achieved dual representation. However, the appeal to the contingency of movements cuts both ways: after all, the contingency cue may simply be taken as a cue for correspondence relation, as opposed the referential relation, between the bodily self and its mirror reflection. It is question-begging, then, to simply assert that passing the mirror test indicates symbolic understanding. In addition, the mirror test is perhaps in one respect more challenging for young children, for the referential relation has to be grasped through cross-modal comparison (e.g., visual and kinesthetic/proprioceptive), whereas in the standard symbolic retrieval tasks, both the symbol and the referent are represented visually. Hence, all things considered, there is no reason to favor the reference account of MSR, even if we grant that there are important differences between the mirror test and the symbolic retrieval test.

To recapitulate: while the reference account of MSR presupposes that subjects who succeed in the mirror test are capable of what DeLoache calls dual representation, the research on children's symbolic understanding suggests that children do not have this capacity until 2.5–3 years of age—-at least a year after they can reliably pass the mirror test. Of course, absence of evidence is not evidence of absence. It is entirely possible that children do in fact understand referential relations much earlier the current evidence indicates, and that better experimental designs will eventually vindicates this. However, given the evidence available now, this possibility seems remote. At the very least, there is no uncontroversial and independent evidence that 1.5-to-2-year-olds have the kind of capacity required by the reference account of MSR. This, again, casts serious doubts on the account.

3.3 The Self in Time

MSR is atemporal: it requires a mental representation of the bodily self, and the ability to match this representation with the mirror reflection of oneself; but it does not require a conception of the bodily self as a *temporally extended* entity. Presumably, to have this richer conception of the self, one needs some conception of time or persistence, and perhaps also the ability to remember events from one's past. But how, exactly, does this transition from an atemporal to a temporal conception of the self occur in humans? That is, how do young children start to understand themselves as entities that persist in time? In this section we will explore this question. While it is not my goal to propose a *complete* developmental account of temporal self-understanding, I do want to suggest that there are several core elements that make the transition possible. This chapter proceeds as follows: like in the previous section, I first review some experimental data that bear on the issue of temporal self-understanding, with special focus on delayed self-recognition (DSR); next I critically examine some psychological accounts of the development of temporal self-understanding and show their respective shortcomings. Lastly, I propose an alternative account, which, for reasons to be explained, I call the *social mirror hypothesis*, and discuss some of its implications.

3.3.1 A Puzzle: Delayed Self-Recognition

While the temporal dimension of the self may strike us as obvious, it is surprising how much difficulty young children seem to have in grasping it. To illustrate, consider the experiments done by psychologist Daniel Povinelli and colleagues (Povinelli, 1995, 2001; Povinelli et al., 1996). The basic design of their experiments is a modification of the original mirror test, but instead of a mirror, the researchers use videos and recent photos. In the literature, this design is often referred to as delayed self-recognition (DSR) test or paradigm, but it is important to note that the main research question addressed by the experiment is not whether young children can recognize themselves in videos or photos, but whether they can appreciate the relevance of their past experiences, as recorded in these media, to their present state.

In the delayed video version of the test, for example, a child first played a novel game with a familiar individual, and then an experimenter came to praise the child by patting on her head. During the final pat, the experimenter placed a large, brightly colored sticker on the child's head. Several minutes later, the child was shown the video recording that depicted the previous events, including "(a) the child playing the game, (b) the experimenter placing the sticker on his or her head, and (c) several ensuing minutes of the child with the sticker on his or her head." (2001, 77) It turned out that, after seeing the video, none of the 2-year-olds subjects reached up to their heads to search for the sticker, and only 25% of the 3-year-olds did. Most of them, however, had no problem recognizing themselves in the video: when the experimenters pointed to the image and asked them "Who is that?", most children answered correctly, either using their proper names or the first person pronoun (e.g, "That's me! "). In contrast, the majority of the 4-year-olds reached up and removed the sticker on their head once they saw the video.

The photo version of the test is similar. A child was shown two Polaroid photographs of herself, one of which was taken when the experimenter placed the sticker on her head, the other was taken later and the sticker was still on her head. The subjects are 3-year-olds and 4-year-olds. Again, most of the 4-year-olds reached for the sticker when the saw their photograph; most of the young 3-year-olds (35–42 months), however, did not, even after verbal prompting. Moreover, the data showed a clear developmental trend between the 32 months and 53 months. Another interesting finding was that, when they were asked the identification question (e.g.,"Who is that?"), the 4-year-olds almost exclusively used the first-person pronoun, whereas the 3-year-olds used proper name as much or more often than the first-person pronoun. So there seems to be some connection between children's competence with the first-person pronoun and their ability to appreciate the relevance of their past to their present.

Based on these findings, Povinelli and colleagues conclude that children younger than 4 lacked the ability to relate the past self with the present self.[5] According to their model of the development, which will be discussed in more detail in the next section, 2-to-3-year-olds do have an explicit representation of the self, but that representation is of *the present self*, i.e., the self of here and now. On a simple reading of their proposal, the present self is, primarily, the *bodily* self, although occasionally they also seem to suggest that other aspects of the self (e.g., mental states) are represented as well. At any rate, they clearly regard the proprioceptive/kinesthetic dimension as the primary or most salient dimension of the present self. 5-year-old children, on the other hand, have acquired the representation of what Povenelli et.,all label as *the proper self.* The notion is analogous to Neisser's(1988) concept of *the extended self*: the self as an entity that extends forward and backward in time. And it is partly because they can represent the temporally extended self, that children of this age are able to appreciate the relevance of the past to the present, and the present to the future.

Importantly, according to Povienelli, young children's lack of temporal self-understanding is a manifestation of their general difficulty in understanding the relevance of the past to present. This point is supported by the result of another experiment conducted by Povinelli's group (Povinelli et al. 1999), which concerns children's understanding of *event order*: events are temporally ordered, and, normally, something's more recent past has more direct causal implications for its present state. Children younger than 5, however, seem to have great trouble understanding the significance of the ordering of events that unfold in time. In their experiment, only children as old as 5 succeeded at levels strongly exceeding chance in correctly relating the present location of a pup-

[5] Even though they were as *motivated* as the older children to remove the sticker: in the photo version of the test, for example, 85% of the 3-year-olds who did not reached for the sticker when they saw their photographs did eventually reach up for it when they later saw themselves in a mirror.

pet with the most recent relevant event (i.e., an experimenter hiding the puppet in a box), as opposed to the relevant event in a more distant past concerning the same puppet. In short, young children seem unable to comprehend the general relevance of the temporal order of past events to the present state of the world. This is further confirmed by some recent studies by Teresa McCormark and Christoph Hoerl (2005; 2007), who used a different design.[6]

Additional evidence comes from the social-cultural approach to the development of autobiographical memory (Nelson and Fivush, 2004; Fivush, 2011). In general, starting from 4 to 5 years of age, children almost exclusively use the first-person pronoun to refer to themselves; moreover, they have become competent narrators: compared to 2-3-year-olds, they give more temporally organized and elaborate accounts of their past experiences, and use more temporal markers (e.g., *when*, *before*,*after*). Relatedly, Lagattuta and Wellman (2001) observed that, from 3 to 5 years of age, children become increasingly competent at *explaining* other people's current emotions in terms of their past experiences and thoughts.

[6] One of their experiments is helpfully summarized below:

> . . . children were introduced to a novel box and told that pressing the blue button caused a toy car to appear in the window of the box, whereas pressing the red button caused a marble to appear in the window of the box. There was only ever one object in the window at any one time. The box was then covered with a screen, and children were told that two dolls each pressed one of the buttons, and were also told the order in which the dolls did their button pressing, but did not see the button pressing itself. When the screen was removed, the dolls were left beside the buttons that they had pressed, and children were asked what was currently in the window... To answer this question, children had to use information about the order in which the button-pressing events had occurred in order to figure out what toy was in the window right now. It was not until children were 5 that they could answer this sort of question correctly. (McCormack 2015, 659)

Hence, as a general observation of the development in temporal cognition that takes place during this period of life, Povinelli is right, I think, that the development of temporal self-understanding is probably a manifestation of, or is dependent upon, the development of the general capacity to appreciate the relevance of the past to the present. The question, however, is what explains this development. In particular, how do children acquire the capacity to link the past with the present in appropriate ways, and acquire a sense of the self as an temporally extended entity? Even if we agree that young toddlers's self-understanding is limited in that they have not yet understood the self as extended in time, and thus cannot properly relate the past with the present stage of the self, we still want to know the underlying causes of this important development. While Povinelli and his colleagues have not spelled out a specific model of the development, some of their remarks do seem to suggest an attractive proposal that nicely situates the development of temporal self-understanding within the context of more general cognitive developments. In the next subsection, then, I will consider this proposal.

3.3.2 Povinelli's Cognitive Model

According to Povinelli and his collaborators, the ability to relate the past self with the present self is, first and foremost, an instance of the general ability of "sustaining multiple, and contradictory representations of the same object or event" (2001, 87), which has nothing to do with their conception of the self *per se.* So older children can pass the DSR test largely because they can hold in mind and compare two distinct representations of themselves. Elsewhere, they suggest that another general cognitive difficulty that might stumble young children's performance in the DSR test is the understanding of *causality*: as mentioned earlier, some of their experiments seem to indicate (Povinelli et al., 1999) that it is not until 5 years of age that children start to understand the systematic causal relations between past events and present events.

Putting these points together, their proposal seems to be that older children can pass the DSR test because they are able to (1) form and hold in mind two representations of the self (one refers to the self at present and one to the self in the past), and (2) connect these two representations in ways that capture the causal relation between the past and the present stages of the self.

I believe these two aspects that younger children fall short are undoubtedly important factors for an adequate explanation of their performances in DSR tests. However, I think their model is underdescribed in important ways. First of all, it is worth noting that in some other contexts, young children do seem to possess the general capacity for constructing and operating different representations of the same object. We have seen earlier that MSR requires the subject to hold in mind a representation of the body and a representation of the mirror image, and recognize the correspondence relation between the body and the image. Perhaps more relevantly, children start to engage in *pretense play* (Leslie, 1987; Perner, 1991) in their second year of life. To treat one thing *as if* it were something else is to represent that thing in two different ways: for instance, when a child holds a banana up to his ear and mouth and says ”Hello! How are you?”, she is not confused; rather, she represents the banana as if it were a telephone, but she is also aware that it is a banana (and that's why pretense is fun). Indeed, MSR and pretense (along with means-end reasoning, success in hidden displacement tasks, etc.) are often taken to be strong evidence that by the end of their second year, human children are capable of entertaining two representations (a *primary* representation and a *secondary* representation, as they are called in this literature) of the same entity (Perner, 1991; Suddendorf and Whiten, 2001). Hence, Povinelli's explanation of young children's failure in the DSR test is at least importantly incomplete: he also needs an account of the relevant differences between the DSR test and other tasks that tap onto the same general capacity (that is, if his theory is correct), otherwise it would seem mysterious why young children who exhibit the general capacity at issue in other tasks nevertheless fail the DSR test.

Moreover, even if Povinelli et al. are right that the ability to relate the past with the present depends on the general capacity of forming distinct representations of the same entity, it is unlikely that this is the whole story. After all, the general capacity at issue has nothing to do with the understanding of time in general or of the temporal dimension of the self in particular. But it seems that this sort of understanding is essential to the ability to connect the past to the present, and to the conception of an extended self. As a result, the model Povineli and his colleagues propose cannot differentiate judgements of *synchronic* identity versus *diachronic* identity: after all, representing (either synchronic or diachronic) identity relations, on their model, just consists in holding in mind two representations of the same entity and connecting them in an appropriate way. A better model of our understanding of the extended self should capture the distinctively temporal dimension the self.

The second part of Povinelli's model concerns causal cognition, and thus might be seen as filling the gap between the general capacity to form distinct representations of the same object and the more specific *temporal* understanding that enables one to link one's personal past with the present. This is because, according to Povinelli et al. (1999) and McCormack and Hoerl (2005, 2007), the conception of causality is an essential element of temporal cognition: to properly locate an event in time is to see it as causally depend past events. It follows that to adequately represent the self as temporally extended, one has to understand that the present state of the self causally depends on its past states. On this view, the ability to causally connect the past with the present is precisely what 2-to-3-year-olds lack. One indication of this, as noted earlier, is that 2-to-3-year olds are unable to understand the causal significance of the temporal order of events. If, at a more distant time T1, a puppet was in the blue box, and at a more recent time T2, the puppet was moved to the red box, then assuming that there was no intervening changes, it is more probable that the puppet is in the red box now. But when 2-to-3-year old were asked to identify the current location of the puppet, only a few succeeded, as shown by the experiments of Povinelli et al. (1999). Perhaps, then, young children's defects in causal cognition is what explains their failure in DSR tests.

Their theoretical point about the importance of causal cognition is well taken. Indeed, I am inclined to agree that, for a full-fledged conception of time, *some* notion of causality is perhaps indispensable. However, I am not convinced by their explanation of the experiments. This is because it is not obvious that *causal* inference is really needed to explain the reported experimental results. To see this, it would be useful to distinguish *temporal* order from *causal* order. The temporal order of events or states consists in the chronological relations between events or states in time ("before","after"); the causal order of events or states is the sequence of successive events or states, as they unfold in time, that are connected by causal chains. The later arguably presupposes the former (that is, understanding temporal order is necessary for understanding causal order), but the reverse does not hold. An event that occurred at an earlier time may have no causal impact whatsoever on an event of a later time. Hence, the representation of the temporal order of events does not require the representation of their causal relations, if there is any; indeed, it is possible to represent the temporal order of events without being able to represent causal relations *at all*. Importantly, the awareness of the temporal relations between events can account for the success in the tasks designed by Povinelli et al. (1999) and McCormack and Hoerl (2005, 2007). All it takes to pass the tests, it seems, is some fallible but simple heuristic rules such as "the more recent the state of an object is, the more similar it is to the present state of the object" or "the an object is likely to remain where it was in the most recent past". The children who correctly located the puppet in the red box might do so simply because they had learned such rules (and because they remember the last time they saw the puppet it was moved to the red box). Therefore, no causal inference is strictly required to pass this kind of test. Likewise, the 4-to-5-year-olds who passed the DSR test might rely on such rules as well: one possible explanation of their success is that they realized that in the recent past one experimenter placed a sticker on their forehead, and inferred that they have a sticker on their forehead now, because the present state or location of the sticker was (probably) similar to its state or location in the recent past.

Furthermore, there are ways to mentally encode the temporal order of events that do not presuppose any understanding of causality. For example, an important *subjective* clue for the temporal relation between events is their vividness, strength or accessibility in memory, but these features have nothing to do with causality as such. Indeed, there is plenty of evidence indicating that both adults and children frequently locate events in time in precisely this way (reviewed in Friedman 2004/). For our purpose, it is important to note that the maturation of this capacity coincides with improved performance in the DSR test: it is not until 4-to-5 years old that children start to reliably form accurate judgements of relative recency of past events: i.e., which event occurred in the more distant or recent past (but only if the events occurred no more than a few months ago, otherwise even 8-year-olds are at chance level.) Therefore, it is not obvious that *causal* inferential capacity is really needed to explain children's performance in the DSR test. On an alternative interpretation, what younger children fail to appreciate is not how their personal past *causally* bears on their present, but that the things, including themselves, generally tend to remain the way they were in the (more) recent past (assuming that is no intervening events that happen to them). The latter explanation implies that, to pass the DSR test, the child needs a conception of the self as something that exists both in the past and the present, but the connection between the two stages of the self does not have to be represented as *causal.*

To sum up: on Povinelli's cognitive model, the sense of the extended self is formed on the basis of two general capacities, i.e., the capacity to form and operate on two distinct representations of the same entity, and the capacity to draw causal inference from past events to present events. I've argued that this model faces some empirical and theoretical problems. On the one hand, there is plenty of evidence that children younger than 4 do have the first capacity; on the other hand, the second capacity is not necessary for passing the DSR test. It is thus unclear how these general capacities are essential to pre-schoolers' gradually developed sense of the extended self. Moreover, Povinelli's model leaves the developmental question untouched, as it does not even try to explain *how* the the sense of the extended self emerges in children.[7] Are the general capacities at issue learned capacity? Or are they in some sense innate? In the next section, we will consider a developmental model that pays special attention the *how* question .

3.3.3 Fivush & Nelson's Social-Cultural Model

In a number of influential articles and books, Robyn Fivush and Katherine Nelson developed a social cultural theory of the development of autobiographical memory (Nelson and Fivush 2004; Fivush and Nelson 2006; Nelson 2009; Fivush 2011). In their earlier work, autobiographical memory (AM) was defined as "declarative, explicit memory for specific points in the past, recalled from the unique perspective of the self in relation to others."(Nelson and Fivush 2004, 488). Now although their theory is not primarily about the issue that we are concerned with, namely, children sense of the temporally extended self, these two things are obviously and closely related. Indeed, it seems that the ability to form AM about one's personal past presupposes

[7] Another worry is that, to understand that the present self causal depends on the past self, it seems one has to assume that these two selves are identical in the first place. If this is the true, then Povinelli's appeal to causal cognition, like Fivush&Nelson's Social-Cultural Model to be reviewed below, already presupposes, rather than explains, the fact that the children at issue have temporal self-understanding.

the sense of self as temporally extended: for the relevant memory to be genuinely autobiographical, the subject has to recognize that the self recalled (i.e., the past self) and the self doing the recalling (the present self) are identical. This connection between AM and the extended self is made more salient in Fivush's more recent definition of AM: AM is "a sense of a self as continuous in time linked across specific experiences placed on a personal timeline that stretches back into a personal past linked to the present and projected into the future."(2011, 570) At the very least, then, we can assume that the sense of the temporally extended self is an essential component of AM.

Unlike Povinelli's theory discussed above, Fivush and Nelson's theory emphasizes the critical role of social interaction and linguistic communication in the development of AM, especially in the form of *parent-child talk* about the past. But it should be noted at the outset that their model is rather complex: many cognitive capacities and mechanisms, such as episodic memory, theory of mind, the acquisition or possession of complex mental and temporal concepts, etc. also contribute to the development of AM. Moreover, in principle at least, social-cultural models are compatible with cognitive models of the sort proposed by Povenelli. These models could in fact be complementary: while the cognitive models focus on the underlying cognitive structure and mechanisms of temporal self-understanding, the social-cultural modoels stress contributions of social context and interaction to developments of the relevant capacities (Nelson and Fivush 2004, 501). Moreover, claims about the underlying structure and capacity of a cognitive phenomenon often have developmental implications about how the phenomenon emerges, and vice versa. Therefore, these two approaches need not be mutually exclusive.

I will not try to summarize Nelson&Fivush's theory of AM in its entirety. Rather, I will just highlight the parts that are most relevant to children's ability to understand themselves as continuous, or temporally extended, entities. Their basic contention is that social interaction is essential to the development of AM. On the other hand, however, they also concede that, well before effective social-linguistic communication kicks in, very young children already have what they call pure episodic memory (PEM): that is, young children can reliably recall specific events they experienced, often after days and even weeks.[8] What distinguishes AM from *pure episodic memory* (PEM) is that PEM is the memory of the experienced events happened at a particular time in the past, and unlike AM, the self as such is not involved in the content of PEM. That is, my PEM of a particular past event does not represent the event as something that *I* have experienced in past. In still other words, PEM is more about the objective events one has experienced or encountered than about the subjectivity of the experience. For this reason, their conception of episodic memory departs importantly from the tradition founded by Endel Tulving (2002), which defines episodic memory as *essentially* involving awareness of the self (i.e., autonoetic awareness).

On the other hand, however, it is not clear whether, on Nelson & Fivush's model, social interaction is *necessary* for children's sense of the self as such. But it is clear that, on their model, such interaction is essential for children's *temporal* understanding of the self, and their ability to locate themselves in time. An especially important form of interaction is parent-guided, joint reminiscing about events of the past, which parents (mothers in particular) and their children routinely engage in. Drawing on the large literature on joint reminiscing and its influence on children's developing autobi-

[8] FIvush summarizes some of the relevant findings:

> . . . by 9 months of age, infants demonstrate recall of previously seen sequences of actions through deferred imitation, a methodology in which infants are asked to recall previously experienced events through action . . . By the end of the first year, infants can reliably recall complex sequences seen only once before even after delays of up to several weeks. Thus, by 12 to 18 months of age, typically developing infants possess complex memories of even quite novel experiences and can reliably demonstrate recall of these experiences over substantial delays.(2011, 565)

ographical skills (Hudson 1990; Harley and Reese 1999; Bauer 2007), Nelson&Fivush argue that social interaction and linguistic communication have a defining and formative impact on children's sense of the past in general, and their *personal* past in particular. [9]

More specifically, on their model there are at least three ways in which parent-guided reminiscing influences children's sense of the temporally extended self. First, such conversations enrich children's memory about the past event. One important factor is the elaborativeness of the parent. When elaborate parents verbally prompt their child to describe his or her past experiences, they tend to ask more open-ended questions, and their questions are more specific and less repetitive than the questions asked by less elaborative parents. With the assistance of their elaborative parents, children can recall more details, and thus have a richer memory of their past experiences than those who have less elaborative parents. (Reese et al., 1993; Welch-Ross, 2001) .

Second, parent-guided talk about the past also contributes to the *organization* of children's memory of their personal past. Again, the elaborative skills of parents play a crucial role. An elaborative parent is very good at weaving together pieces and fragments of experiences the child recalled, and thus helps the child to construct a coherent and well-organized narrative. Correspondingly, her child's verbal recalls also tend to be more coherent. By contrast, the verbal recalls of children whose parents are less elaborative tend to be more impoverished and fragmentary (Fivush and Nelson 2006). Of particular interest here is the parent's use of temporal expressions, such "when", "before", "after", "first... then...", etc. In the early stages of development, children seldom use these expressions on their own to locate events in time. Rather, it is almost always parents who supply the temporal ordering and organization with the

[9] As they put it, "the ways in which parents, especially mothers, structure strong and enduring influences on how children come to construct their own narrative life history" (2004, 497)

linguistic temporal markers, and children just repeat or confirm what their parents say (Hudson 1990; Haden et al. 2001). As Nelson & Fivush put it, "adults provide the linguistic scaffold, or framework, that helps children to organize their experience, both as it is occurring and in retrospect, and it is this organization that allows children to both represent and subsequently verbally recall the event in a coherent and meaningful fashion." (2004, 494) It is likely, then, that this particular aspect of parent-guided reminiscing facilitates children's understanding of the temporal order of past events.

A third way in which joint reminiscing helps to build children's sense of their personal past is less straightforward. According to Fivush & Neslon (2006), the parent-guided talk about the past typically invokes many expressions that refer to mental states. The purpose of such talk is not just to recall their past mental states, but also to make sense of the past by making explicit the causal connections among various mental states, and between mental states and the events or states of affair that the child has witnessed. (e.g., the child was and still is sad, *because* she saw a wounded cat). However, the exact route from the mental state talk to children's sense of the temporally extended self is elusive to me, and from what I understand, even if the former somehow influences the latter, it is perhaps indirect. In any case, their point that is most germane to our present concern seems to be that, by talking about their past mental states, children learn to grasp the causal link between their past experiences, thoughts and emotions and their current states. This enables them to construct rich personal narratives that make sense of their actions in mental state terms, especially in emotional terms. Partly in support of this, they cite several lines of research that purport to show that "mothers who talk more about emotion during shared reminiscing early in the preschool years have children who include more emotion in their personal narratives later in the preschool years", and that "mothers' use of narrative evaluation that includes emotional reactions and feelings during early mother-child reminiscing (e.g. ‘wasn’t that fun?’ and ‘that was neat!’) predicts childrens’ later use of evaluation in their own independent personal narratives” (2006, 241) [10]

To sum up: on Fivush & Nelson’s social-cultural model, parent-guided talk about the past is crucial for the development of children’s autobiographical memory, because it facilitates the recall of specific details of their personal past, provides the temporal structure for their memory, and reveals the causal links between their past and current mental states. As a result, children develop a richer, better situated and more accurate memory of their personal history.

Fivush & Nelson’s social-cultural model contains many interesting ideas about the development of autobiographical memory. However, in so far as the development of temporal self-understanding is concerned, it seems importantly incomplete. The fundamental problem is that their account of joint reminiscing seems to *presuppose* that the child who engages in the conversation is already able to grasp, to some extent at least, that the person whose past experiences she remembers (with the help of her mother) is *herself*. That is, she already understands that she exists both in the past

[10] Perhaps unsurprisingly, on their model parent-child conversations about children’s mental states are also directly related children’s theory of mind or mindreading capacities. For example, one of their claim is that, it is through parent-guided conversations that children start to understand the representational nature of memory (and other, related mental states): what one seems to remember may not reflect what actually happened in the past, and different people can have different memories about the past, and have different thoughts and feelings about a past event that they both remember (2006, 239). It is not obvious to me how these points are related to children’s sense of the extended self, but even if we grant that the development of autobiographical memory is closely related to theory of mind (see also Perner 1991; Perner et al. 2007), it is far from clear how theory of mind contributes to the understanding of the self as an continuous entity that exists in time. At any rate, there are now plenty of evidences suggesting that even children who are less than than two-year-old already have some basic theory of mind capacities (Onishi and Baillargeon 2005; Southgate et al. 2010; Baillargeon et al. 2010). The nature of such capacities, and how (and whether) they are different from the mindreading capacities of older children and adults are heatedly debated (e.g., Apperly and Butterfill 2009; Perner and Roessler 2012; Carruthers 2013), but it is relatively safe to assume that children have some understanding of the representational of the mind long before they grasp the temporal nature of the self–and before they effectively participate in conversations about their personal past.

(i.e., in the remembered situation) and in the present (i.e., in the current situation, in which she is talking with her mother about their past experiences).

Let me explain. According to Fivush & Nelson's model, by leading the conversations about the past, parents provide their children with rich details and temporal scaffold of children's past, which in turns enhances the children's autobiographical skills. However, in order to understand and actively participate in such conversation, the child has to understand that her parent is talking about *her* in the first place. This in turns means that the child already takes for granted that there is an earlier stage (or stages) of her where her experiences departs from what she is experiencing at the moment of conversation. In other words, if joint reminiscing were to influence the child's autobiographical memory in the way Fivush & Nelson proposes (e.g., uncovers specific details of the child's experience and supplies the temporal organization for the memory), the child has to have some sense of herself as temporally extended in the first place: at the very least, she is able to regard the recalled events as belonging to her own past. Hence, having some sense of the extended self is a precondition for autobiographical memory, and also for the capacity to effectively engage in the parent-guided talk about the personal past. Similarly, the child's ability to link her past mental states with the present situation presupposes her ability of recognizing that those past states were hers (even though she is not in those states now). The mother certainly could help the child to appreciate the relevant causal links by making them more explicit, but only if the child is able to identify with her past self.

In short, Fivush & Nelson's model presupposes, rather than explains, young children's temporal self-understanding. Unfortunately, this feature of their model, though perhaps not a problem in itself (for their model is primarily concerned with the development of autobiographical memory), is not made as clear as it should be. As an illustration, consider the following passage from Fivush and Nelson (2006). Background: one major aim of that paper is to show "how specific talk about past internal states, as embedded in narratives, allows the child to *construct* a temporally extended understanding of self and other". (240; my emphasis). Here their point appears to be that parent-guided talk about the past, in the form of narrative, enables the child to develop temporal self-understanding (that is, to understand herself as temporally extended). A closer look, however, suggests their model is not really about the development of children's temporal self-understanding *per se*. Rather, the model in fact assumes that children's temporal self-understanding is already in place. For instance, in their explanation of the contributions of mental state talk in joint reminiscing, they claim:

> ...in reminiscing about the past, the child is confronted with multiple ways in which past internal states may be related to current internal states. Some of these internal states may differ between then and now and some may be the same. Thus, in these conversations, children may learn how to relate past self-understanding to current self-understanding. Such discussions allow children to come to an understanding that the life of the mind, as expressed in mental state and emotion terms, continues to exist even when the event is over, that thoughts and feelings about events exist over time, and continue to influence the interpretation and evaluation of the past event in the present.(241)

Their point here seems to be that conversations about (past and present) mental states enable the child to better understand the persistence and effects of mental states. This may well be true, but at best it shows that conversations about mental states, like conversations about details and temporal orders of past events, greatly *enrich* the child's understanding of aspects of her past (e.g. her past mental states) and how they relate to the present. But it is by no means clear how this allows the child to *construct* a conception of herself as temporally extended in the first place. Rather, that conception is presupposed: to (better) understand how one's past and current mental states are related, one has to identify the past self and the current self in the first place—one has to assume, that is, it is the same subject that extends from the past to the present.

Therefore, while I agree that in general, social interaction and linguistic communication are key to many of our cognitive developments, and that Fivush&Nelson's framework rightly emphasizes the importance of joint reminiscing for the development of autobiographical memory, I think they have not yet adequately explained how children come to see themselves as temporally extended.

3.4 The Social Mirror Hypothesis

Let me take stock. In section 4.3, we began with a developmental puzzle: young children, who have no problem passing the mirror tests, have great difficulty with the DSR test, apparently because they can not connect their past experiences with their present state. The understanding of the self as temporally extended emerges only until they are 4 to 5 years old. Why do they fail the DSR test, and how is temporal self-understanding developed? Then I examined two models that purport to address this. I argued that although they contain interesting and plausible ideas, both have crucial shortcomings as an explanation of the cognitive development at issue.

In this section, I will propose an account of the development of children's sense of the extended self. As will be clear, it draws on Fivush & Nelson's models in some ways, but with a more direct focus on children's temporal self-understanding and a more explicit emphasis on the way *natural language* transforms our cognition. It is also worth noting at the outset that, since the empirical study that directly bears on the issue is still in its infancy, the account will be, to a large extent, speculative. As all the authors we have discussed acknowledge, the understanding of the self as temporally extended is a difficult achievement, and probably involves a myriad cognitive mechanisms and resources. Hence, I certainly do not pretend that my account is complete, and I expect that future empirical research will fill in the necessary details.

3.4.1 Memory, Language, and the Past Self

In general, I agree with Fivush & Nelson and many other psychologists that natural language has a profound influence on human cognition. Some of our fundamental cognitive capacities are practically impossible without the help of nature language. Importantly, this should not be taken to be the obvious point that language is a great tool of communicating information that we otherwise would not have known independently. This is certainly true, and is also what Fivush & Nelson have in mind when they talk about the contribution of social-cultural interactions to the development of autobiographical memory. However, it seems to me that they overlook a deeper or more structural influence that language may have on our cognition. Through language we learn not just more information about the world and ourselves, but also some basic representational structures or frameworks that make much of these knowledge possible. This, of course, does not mean that language does all the work; some innate cognitive structures or mechanisms might be crucial as well. But the role of natural language is indispensable in how it transforms our innate endowment. What I want to highlight, then, is the interaction between natural language and innate cognitive capacities that may jointly produce the conception of the temporally extended self.

To illustrate what I have in mind, consider another foundational human cognitive capacity: number cognition. The psychologist Elizabeth Spelke argues that our natural number concepts are a joint product of our innate endowments ("core systems") and our natural language, and it is (only) through natural language that we are able to effectively *combine* the core systems so as to construct natural number concepts (Spelke 2017). As she puts it:

> ...natural number concepts arise through the productive combination of representations from a set of innate, ancient, and developmentally invariant cognitive systems: systems of core knowledge. In particular, natural number concepts depend on a system for representing sets and their approximate numerical magnitudes...and a set of systems that collectively serve to represent objects as members of kinds. None of these core systems is unique to humans, but their productive combination depends on

> the acquisition and use of a natural language. Because both the core systems and the language faculty are universal across humans, and because children master their native language spontaneously, natural number concepts emerge universally, with no formal or informal instruction. Because language is unique to humans, so is our grasp of the natural numbers. (148)

This is not, of course, the place to examine Spelke's proposal. [11] But her point about the cognitive and developmental function of natural language is a valuable one, and I want to suggest something analogous in the case of children's understanding of the temporally extended self. The basic idea is that, through the acquisition of and competence with first person pronouns and names, children are able to productively combine various innate capacities, and gradually (since the competence in the relevant forms of language comprehension and production takes time) develop a sense of the self as extended in time. Hence, like in Fivush&Nelson's model, social interaction and linguistic communication, mediated by language use, are essential for this development, but the way this changes our cognition is more profound on my account.

I will call my account the "social mirror hypothesis", if only to capture the social and interactive dimension of the development: just like physical mirrors enable us to see reflections of aspects of our physical self that we are otherwise unable to see, other people, especially our interlocutors, enable us to gradually appreciate the temporal dimension of the self (and presumably other aspects too, such as one's social identity).

Briefly put, the social mirror hypothesis claims that the emergence of the temporal self-understanding relies on three components. The first has already been mentioned: natural language. But specifically, it is the linguistic expressions that are used to explicitly refer to the child herself. The other two components are arguably innate,

[11] See Carey (2009) for a different view that also stresses the transformative impact of external symbolic systems on our cognition. For Carey, however, what plays the role is not natural language, but counting procedures.

or in any case largely independent of social-cultural learning[12]: a self-representation of the bodily self, and episodic memory of one's past experiences.

Entering the second half of their second year, children have developed at least a primitive form of self-representation, i.e., bodily self-representation, as manifested in their success in the mirror test. At around the same time, children start to use first-person pronouns to self-refer (Oshima-Takane et al. 1999) and display some self-conscious emotions, such as embarrassment and shame, which are often associated with their exposed body (Lewis 2003). Hence, by this age, children are able to conceive of themselves as, first and foremost, a physical object that (1) is located in the objective space, (2) can be perceived by others, and (3) can be systematically explored with or without the help of a mirror. Now since many non-human animals, such as other great apes, can pass the mirror test, bodily self-representation is innate. Still, for normal human beings, it is the first step of the understanding of the *objective* self (James 1890; Moore 2006): the self as an object in the external, physical world.

However, as we have seen above, the temporal dimension of the self has not been understood by young children of this age. The bodily self-representation itself is *atemporal.* If Fivush & Nelson are correct, then parent-child talk (which is still very fragmentary for children at this age) about their shared past experiences certainly might help them better understand their past, but how, exactly? The question from our discussion of Fivush &Nelson's proposal is precisely this: how do children come to understand that it is *their* past experiences that are being talked about? After all, this understanding seems to presuppose that they are able to identify themselves with the subject whose past experiences are being discussed.

This is the place where particular features of referential expressions become important. Well before they can understand the temporal dimension of the self, young

[12] And are liked shared by non-human animals.

children have been exposed to proper names and personal pronouns, and most of them have started to correctly use these linguistic expressions (Macnamara 1982; Hall 1999). With their competence with these expressions comes particular conceptions of the their referents. For example, even very young children tend to assume that proper names, unlike adjectives, are used pick out unique individuals, although they can be prodded into accepting that two individuals can bear the same name with sufficient explanation or contextual cues (Bloom 2000). While this linguistic assumption is not explicitly temporal, it might facilitate the understanding that the referent of a proper name could bear different properties at different times. So, if a child knows her name, and her parent uses the name to denote her in their joint reminiscing, then, presumably, the child would be able to infer that her parent is talking about her (the child), even though many things said *about* her concern her experiences at an earlier time and thus no longer hold. In this way, she would start to appreciate the connection between the self at the time of conversation with the (past) self that is explicitly mentioned in the conversation.

More important perhaps is the use of personal pronouns. Indeed, it would be unusual if in a face-to-face conversation between a mother and her child, the child is referred to with a proper name (or, for that matter, the third-person pronoun). This is because that this way of talking, though certainly possible, is dissociating or disengaging, for it is as if the speaker is talking about someone who is not present. Hence, parents are much more likely to use the second-person pronoun, e.g., "you", and the first-personal plural, e.g. "we", in such talks. It might initially seem that, since pronouns shift references across contexts, they are harder for children to learn than proper names, which typically have a fixed reference. However, the empirical research in fact does not support this supposition: it seems children start to understand personal pronouns and proper names at around the same age, i.e., the second half of their second year (Bloom 2000).

A particularly significant aspect of the use of personal pronouns is that, as noted above, it engages with the child in a way that the use of proper names does not. Trying to help the child to recall their trip to the zoo, a mother says"We have been to the zoo, remember? You liked the panda a lot." Crucially, she is not just *reporting* what the child did in the past; rather, she is getting the child to remember her own past. The communicative function of personal pronouns is thus twofold: by using them the mother makes explicit that she is both talking *to* the child and *about* the child. This in turn helps the child to relate her present self to the remembered self, and to become aware of the fact that, although some experiences or events are no longer present, they were still *her* experiences or events. Therefore, while the use of proper names in parent-child conversations is perhaps not uncommon, personal pronouns are more convenient and useful in getting the child to appreciate the connection between her presence in the present (when she is being talked to) and her experiences in the past. Correspondingly, on the side of language *production*, the child who more frequently uses the first person pronoun to refer to her past experiences is more likely to have a better understanding of her continuous existence over time. Suppose she looks at a video that depicts a past episode of her life, and utters, "That's me!", it would seem that she genuinely self-recognizes herself: she can *directly* relate herself to the child in the video, and thereby take the experiences of the child as her own. By contrast, if the child uses her proper name to refer to herself in the video, it is possible that the her self-recognition is not very different from her recognition of others, e.g., by associating a proper name with stored information about certain *objective* features (such as her looks, her dress, etc.) In other words, it perhaps should not be seen as robust self-recognition at all, for she might fail to directly appreciate how the child in the video is related to *herself* (even though she gets her name correct).

So by virtue of having a self-concept of their bodily self, young children can conceive of themselves as objective entities in the world (and such a self-concept presumably plays an important role in learning their names and the first person pronoun); in addition, joint reminiscing helps to bring about the apparent connection between their present self with experiences of their past self. However, this connection cannot be fully appreciated unless young children themselves can actually recall the part and parcel of their past. That is, if the previous events mentioned in the conversation leave no traces in their memories, then even if they can understand their names and personal pronouns, they would not be able to effectively participate in *joint*-reminiscing about their past. Moreover, without at least some memories about the relevant past events, the content of the conversation would seem, in an important sense, alien to them: it does not resonate with what they can remember. In that case, the conversation would not be an effort of *joint* reminiscing, but only a case of (parent) reporting and (child) repeating.

Hence, another cognitive capacity is called for: the child's ability to recall the specific events—that is, her episodic memory. At first glance, the appeal to episodic memory might seem problematic, partly because, as hinted earlier, the very notion of episodic memory is contested: according to a minimal conception (Nelson and Fivush 2004; Fivush 2011; cf. Tulving 1972), episodic memory is the memory of the *what*, *when* and *where* of specific past events that occurred in particular contexts, and it is not necessary that the events are remembered *as* what the subject herself has experienced; by contrast, on a much richer definition (Tulving 2002; Markowitsch and Staniloiu 2011), episodic memory inherently involves some kind of self-consciousness ("autonoetic consciousness"), or at least, its phenomenology has a distinctively subjective dimension (McCormack 2001). Although there is quite a lot of discussion of this richer notion in recent years, to me it remains unclear what the subjective dimension of episodic memory consists in: is it just a sense of familiarity, a feeling of conscious recollection, or a special mode of presentation that confers on the remembered experience a distinctive subjective feeling, e.g., rendering it an experience-for-me?

In other words, it is not clear how the self figured in this type of episodic memory. Moreover, in the empirical literature, the two senses of episodic memory are not always delineated, and this problem is partly responsible for the heated but sometimes confusing debate about whether episodic memory is uniquely human (Tulving 2005; Suddendorf and Corballis 2007; Allen and Fortin 2013), or, more relevant to our concern, whether children younger than 4 can form episodic memories: Tulving himself holds that children cannot form episodic memories prior to 4 or 5 years of age (2005; see also Perner and Ruffman (1995)); others, perhaps using the minimal conception similar to what Fivush & Nelson call "pure episodic memory" (PEM), contend that episodic memory emerges much earlier in childhood (Bauer 2007; Tustin and Hayne 2010), even before the end of their first year. [13]

Fortunately, these hard conceptual and empirical issues need not concern us here. After all, we are primarily interested, not in autobiographical memory *per se*, but in the development of children's understanding of themselves as entities that persists in time. This understanding requires an objective and temporal conception of the self: the self is something that exists both in the past and in the present. Nothing in developmental literature on episodic memory indicates that this understanding is already in place before 4 to 5 years age (and there are plenty that suggests the contrary, as reviewed above). It is without question, however, that before they reach the full-blooded understanding of the temporally extended self, young children are able to verbally recall their personal past in parent-guided conversations, even though in children younger than 4 years of age it tends to be fragmentary and disorganized. It is safe to conclude, then, that these children have at least minimal or pure episodic memory (Bauer et al. 2000), and can verbally recall some pieces of their remembered

[13] Besides the lack of clear criteria, another general methodological difficulty for this area of research is how to make the elusive phenomenology of episodic memory tractable in empirical tests—especially for non-human animals that are unable to convey verbally their subjective experiences.

scenario. For the sake of argument, I will agree with Tulving and others that children younger than 4 do not yet have the richer kind of episodic memory that has a distinctive subjective dimension (whatever that amounts to), and thus are unable to form the explicit thoughts about the role of the self in the remembered scenario. That is, I will assume that young children typically do not regard the things they remember *as* things that they have experienced.

On the present account, the explicit sense of the presence of the self in the recalled past is brought about by linguistic tools of reference. Specifically, by talking to the child about her personal past and referring to her with a proper name and/or personal pronoun, the parent (or other speakers who participate in joint reminiscing) enables the child to explicitly appreciate that it is *her* past, that she is the subject or the witness of the recalled episodes. This in turns facilitates the child's understanding of the self in time, as the conversation makes it clear that the child is not *only* present at the moment of conversation, but also was present in the remembered past. [14] Metaphorically put, then, the picture is this: by referring to the child with a proper name or pronoun, joint reminiscing is like "mental time travel machine" that takes the child's objective awareness of the self, which she has in virtue of the atemporal bodily self-representation, to a (collectively) remembered past, and thereby explicitly anchors or locates the child's self in the past. In other words, what joint reminiscing does is explicitly placing the self in the past, so that the child can come to see that it is *their* experiences that are remembered and discussed. In this way, their objective presence in the past becomes salient to her, so that she is now in a position to appreciate the connection between their past and present. On the social mirror

[14] In other words, while the child might be able to recall episodes of events, the memory is only implicitly indexical (see Chapter 2): they are memories of events experienced from her perspective, but the self as such is not represented in the memory. Rather, her objective presence or role in the past episodes is understood with the assistance of the parent.

hypothesis, joint reminiscing is a social mirror through which the child comes to sees her *objective presence* in the past, and to *identify* with her past self.

In this section, I have sketched a model of the development of children's sense of temporally extended self. The model has three key elements: a bodily self-representation, episodic memory of past scenarios, and joint reminiscing about the past.[15] In virtue of the bodily self-representation, young children are able to conceive of themselves as objective entities in the world, but the representation itself is *a*temporal. In virtue of their (pure) episodic memory, they are able to recall the events that happened in the past, but their own presence in the past is not explicitly reflected in the content of the memory. It is through linguistic communication, especially in the form of parent-guided talks about children's past (in which they are referred to with proper names and pronouns), that the objective presence of the children in the past, as well as the connection between their past and present, is made explicit. According to the social mirror hypothesis, these components jointly produce the understanding of the self as an entity that persists in time. [16] Analogous to Spelke's account of the acquisition of natural number concepts, this model assigns to natural language a profound, transformative role: by combining two (relatively) primitive cognitive systems, it greatly enriches our basic conceptual repertoire.[17]

[15] The first two components are likely innate, or in any case are acquired relatively early in childhood; the last essentially involves social context and linguistic communication, and crucially depend on language development.

[16] Since the issue here is entirely empirical, and concerns the trajectory of normal human children, I don't find it particularly interesting or relevant whether language-based social interaction is *necessary* to the developmental process.

[17] Obviously, the details of the proposal can only be filled in with the empirical research on the memory and temporal cognition, which are still in early stage, but I hope that the model I propose here is at least plausible and suggestive.

3.4.2 Prediction and Evidence

The model I proposed in the previous section identifies three key elements for the development of temporal self-understanding. As I have shown above, there is independent evidence for the existence of the relevant cognitive and linguistic capacities (i.e., the representation of the bodily self, episodic memory of past events, and the understanding and use of proper names and personal pronouns) in young children before they can reliably recognize the causal implications of their past on their present. In this concluding section, I shall lay out some concrete predictions of the social mirror hypothesis, and see to what extent they are, or can be, empirically vindicated. To be sure, research on the temporal dimension of children's self-understanding is still in its infancy, and there are plenty of conceptual and empirical issues that remain to be explored. Therefore, the discussion here will be largely exploratory. However, I hope that, by detailing its predictions, these discussions at least help to further elucidate the hypothesis sketched above, and to make it vulnerable to empirical test. Hence, in places where I cannot find relevant empirical research that directly bears on the issue, I will offer some tentative suggestions about possible experimental designs.

At the most general level, like Fivush & Nelson's model, the socially mirror hypothesis predicts that without linguistic communication, it is practically impossible for a creature to develop the sense of temporally extended self. This is so even if one has an objective representation of the bodily self and episodic memory about one's personal past. Hence, it seems possible to adopt some variations of Povinelli's DSR test to higher animals that can pass the mirror test and have episodic memory of past events. Likewise, the test can be extended to human children whose social and linguistic abilities are impaired. To my knowledge, there is little, if any, empirical work done on non-human animals that explicitly targets the question of the temporal self-understanding. On the other hand, in recent years there have been some interesting works that examined children with autistic spectrum disorder (ASD) with DSR tests

(Lind and Bowler 2009; Dunphy-Lelii and Wellman 2012; Goddard et al. 2014). The consensus of this line of research seems to be that although children with ASD show impaired performance in some tasks (e.g., theory of mind test), their performance in DSR tests is not significantly different from typically developing children. However, this result is in fact compatible with the social mirror hypothesis, for the children with ASD selected for these studies generally have verbal and communicative capacities comparable to a normal 5-year-old. For example, they can use either proper names or personal pronouns to self-refer. Hence, it seems in principle that these children could engage in joint reminiscing about the past. It would be interesting to see how children and adults whose language competence are less developed would perform in DSR tests.[18]

Next, at the intermediate level, the social mirror hypothesis predicts that the extent to which children's specific past are referred to in conversations will likely have a crucial impact on their temporal self-understanding. More specifically: if a child, whose language competence is otherwise normal, does not frequently participate in joint reminiscing about her past, or the joint reminiscing is more concerned with features other than the child's own individual presence in the past, then (other things being equal) it would be more difficult for the child to understand herself as an entity that persists in time. The issue here is not just about parent-guided talks about the child's personal past, but also about *how* or *which aspect of* the past is verbally recalled and discussed: in particular, whether the child's individual place and role is made salient would, on the present hypothesis, exert some influence on the development of her self-understanding. Hence, this prediction is different from, or at least a lot more specific than, what Fivush & Nelson's model can deliver.

[18] Also relevant (though not specifically about temporal self-understanding) in this connection is a study on the autobiographical memory of deaf adults who did not grow up with sign language (Weigle and Bauer 2000): compared to hearing adults, the age of their earliest memory do not differ, but their early memories are significantly sparser.

Unfortunately, to my knowledge no research has tackled this aspect of the connection between joint reminiscing and temporal self-understanding. Indirectly relevant might be some recent cross-cultural research (Mullen and Yi 1995; Wang 2001; Wang and Fivush 2005), which suggests that in joint-reminiscing about children's past experiences, eastern Asian mothers tend to be less elaborate than European American mothers, and they in general focus more on general routines and collective activities, as opposed to the experience and emotions of particular individuals. According to this line of research, the autobiographical memory reports of people from Eastern Asian cultures "include fewer references to themselves and fewer personal evaluations, relative to reports from children in the United States." (Bauer et al. 2010, 169) This difference purports to reflect the values and conceptions of the self in these cultures (e.g., communal vs. independent). Partly as a result of this difference, it is reported that the average age of earliest autobiographical memories of Americans is a few months earlier than their Chinese counterparts (Wang 2001). On the social mirror hypothesis, parents' narrative styles would have a similar effect on children's temporal self-understanding. If the child's presence and personal experiences in the past is not explicitly mentioned in parent-guided talks, then the child would have more difficulty in appreciating the connection between her past and present, and thus also more difficulty in developing temporal self-understanding. Hence, assuming the cross-cultural observations mentioned are correct, the hypothesis seems to imply that, in general, children from eastern Asian cultures would reliably pass DSR tests at a later age than children from European American cultures. This seems to me to be another promising and independently interesting issue for future study.

Lastly, the social mirror hypothesis predicts that children's linguistic competence (both comprehension and production) with proper names and personal pronouns is positively correlated with their temporal self-understanding. This is because, according to the hypothesis, the ability to understand their names and pronouns (when used to refer to them), and the ability to use such expressions to self-refer, is crucial for the effectiveness of their active participation in the joint-reminiscing of the personal past, and in particular for grasping the connection between the past, remembered self and the present self who is doing the remembering.

Here the general trajectory of children's language development seems to provide some *correlational* evidence for this prediction.(Macnamara 1982; Hall 1999): although children start to use the first person pronoun to self-refer in their second year of life, they tend to make some systematic errors initially (e.g.,using *I* to refer to the addressee; see Oshima-Takane et al. 1999), and they also frequently use their names for that purpose. It is only until 4 or 5 that children almost exclusively use the first person pronoun to self-refer (Lewis 2003)—just around the age they can reliably pass the DSR test.

Even more tellingly, in their original experiments, Povinelli and his collaborators (Povinelli et al. 1996; Povinelli and Simon 1998) observed that, in the DSR test, older children (4 to 5 years) were much more inclined to use the first person pronoun ("That's me!"), as opposed to proper names or the third person pronoun and to identify themselves and to describe the location of the sticker ("It's on my head!") in the delayed video than younger children (2 to 3 years). Especially relevant is the finding that the correlation between the type of linguistic expressions children use to self-refer and their performance in the test is significant. According these researchers,

> . . . the use of the proper name and other dissociative phrases (such as describing the sticker as being on "his/her" or "the" head) was significantly associated with not passing the mark test. Children who reached up to remove the sticker from their heads rarely used their proper name when asked to identify the image and virtually never used "his/her" or"the" to describe the image's head. (Povinelli et al., 1996, 1549)

To explain the difference in the patterns of self-reference between younger and older children, Povinelli and his collaborators draw on Nelson's earlier work (1993) and suggest that the difference perhaps indicates the onset of autobiographical memory in older children. However, it is utterly unclear how the two are related: why does the development of autobiographical memory correlate with the preference to use the first person pronoun to self-refer?

The social mirror hypothesis, I believe, promises a better explanation of the phenomenon, and for this reason this evidence lends additional support for the hypothesis. *Pace* Povinelli and his collaborators, I think what the difference in ways of self-reference reveals is not the development of autobiographical memory *per se*, but a fundamental change in one's self-understanding.[19] In the last section, I have suggested that, in parent-guided conversations about the past, personal pronouns are more convenient tools for getting children to directly appreciate the connection between their past and present existence. The first person pronoun is reflexive. If John says "I did such and such", he is not just *describing* the behaviors of some particular person in the past; rather, and perhaps more importantly, he is expressing the (tacit) recognition that the person is himself. By contrast, this self-identification would be opaque, or can be easily missed, in the corresponding statement "John did such and such", even if it is uttered by John, for John might forget who he is (Perry 1979). Likewise, when the older child points to his image in the delayed video and utters, "That's me!", he is expressing the recognition of the relevance to the image to his present self. The younger child who uses his name, say, "Sam", to identify himself in the delayed video might still posses some kind of self-consciousness (he knows that he is Sam and is able to connection certain physical or featural properties with himself), but by thinking of himself third-personally, he might have a harder time (though of course not impossible) to appreciate the relevance, causal or otherwise, of the depicted image to himself as he is speaking. While using a proper name to self-refer is not incompatible with self-consciousness, it would be oddly indirect and alienating if

[19] This is compatible with the plausible claim that certain forms of self-understanding, such as the sense of the self as temporally extended, is a prerequisite for full-fledged autobiographical memory.

someone exclusively talks about him or herself in this way.[20] The upshot, then, is that the social mirror hypothesis, since it stresses the referential role of personal pronouns in parent-guided talks about the past, can readily explain the correlation between children's use of the first person pronoun in self-identification and their success in the DSR test.

3.5 Conclusion

In this chapter, I have explored two kinds of self-recognition: mirror self-recognition and delayed or temporal self-recognition. I have argued, on both methodological and empirical grounds, for the correspondence account of mirror self-recognition, according to which mirror self-recognition is essentially a product of *cross-modal comparison*: the subject is aware of the cross-modal correspondence between her bodily self-image and her mirror reflection, and is able to exploit that correspondence in her self-directed behaviors. I agree with Povinelli and others that subjects capable of delayed self-recognition have at least two distinct representations of the self, i.e., the present self and the past self,[21] but in addition, I have proposed an alternative model of the development of children's temporal self-understanding, i.e., the social mirror hypothesis, which highlights, among other things, the transformative role of language and social interaction to our cognitive development.

[20] In the most extreme cases, it might indicates that the subject's self-consciousness is impaired in some way: recall, for example, the patient EF, who refers to herself as "Nallie" (the name she is called by her family) but apparently cannot recognize herself in the mirror.

[21] While this chapter and the empirical literature it reviews have not examined children's concept of the future self, which is, presumably, also an essential component for the full-fledged conception of the self as temporally extended, it is worth noting that the ability to remember the self in the past and projecting the self in the future is closely connected, and possibility shares a common neural basis.(Schacter et al., 2007; Conway et al., 2016)

CHAPTER 4
EPISTEMIC MODALS AND PERSPECTIVE-SHIFTING

The last chapter has brought to the fore the importance of linguistic communication for the development of a full-fledged conception of the first person. This chapter continues this line of inquiry, and deals more directly with language and communication, and in particular with the understanding and communication of the perspectives of others. It is impossible to overstate the importance of perspectives in our ordinary communication. What our words mean often depends on where we are, what background information we have, when we are speaking, etc. Suppose Tom reported, “After work, I had a few drinks at the local bar, and then went home”. Count how many words in this sentence implicate, either explicitly or implicitly, Tom’s own perspective. For a conversation to proceed smoothly, the hearer needs to be aware of (or even be able to share) the speaker’s perspective. But it would be silly to ignore the differences between perspectives (“You think stinky tofu is tasty!? What are you?!”). Those unable to see things from others’ perspectives, or to appreciate that others might have a perspective different from their own, will often find it hard to understand what others mean, let alone to engage in fruitful exchanges. But exactly how do other people’s perspectives, potentially different from one’s own, affect the meanings of linguistic expressions? How are they processed in ordinary communication? In this paper, I approach these general questions with a case study: our intuitive judgments about bare (i.e., unembedded) epistemic possibility claims (BEPs), such as “The keys might be in the car”, and “Tom might be drunk”.

This chapter proceeds as follows. Section 1 summarizes the orthodox contextualist theory of epistemic modals and the challenges it faces. Section 2 reviews some recent experimental data that raise interesting questions about our intuitive judgments about BEP. In section 3, I propose an account of BEP that promises to accommodate these intriguing data, and to solve other puzzles discussed in the literature. My approach to epistemic modals focuses on their communicative function. The basic idea of my proposal is that a BEP has two distinct communicative functions: to suggest that participants in the discourse context should take a possibility as live and significant, and to indicate the perspective-dependency of the suggestion.

Since, as will be clear later, my account implies that our assessment of BEPs is sensitive to their perspective-dependent nature, I briefly explore the psychology of perspective-taking in section 4. The hope is that the kind of interdisciplinary approach adopted here would illuminate the psychological basis of our linguistic intuitions. This in turn will benefit our semantic theorizing, as the intuitions often constitute the relevant data points. Combined with a very popular psychological model of perspective-taking, the so-called "egocentric anchoring and adjustment" model, my hypothesis will generate some concrete and testable predictions. In the last section, I will make some suggestions about how these predictions can be tested, and provide *prima facie* reasons to think that they are correct

4.1 Contextualism and Its Discontents

According to canonical contextualism (Kratzer 1991, 2012), epistemic modals are quantifiers over epistemic possibilities. The domain of possibilities, or the modal base, is supplied by the context of utterance. The context supplies the modal base by selecting a particular body of information, and the truth-value of a bare epistemic modal claim, whose general logical form can represented as *Might*/*Must* p, is determined by the compatibility of p (the prejacent of the modal claim) with that body of information.

To illustrate, consider a toy example: suppose we are looking for my keys, and I assert, “The keys might be on the table”. Suppose further that the context selects my knowledge as the relevant body of information. According to canonical contextualism, my assertion is true iff in some of the worlds compatible with my knowledge, the keys are on the table. On the other hand, if I assert “The keys must be on the table”, then that assertion is true in the context iff in all the worlds compatible with my knowledge, the keys are on the table.

Note that, in this example, the relevant body of information is stipulated to be knowledge of the speaker. This interpretation of epistemic modals is standardly called the solipsistic interpretation.(Von Fintel and Gillies, 2011) However, depending on the relevant context, non-solipsistic interpretations could be available as well. Canonical contextualism only requires the context to supply the modal base by selecting a relevant body of information, but which body of information it selects is entirely up to the context itself. Hence, in some cases, the knowledge of other participants of the conversation is incorporated into the relevant body of information, resulting in a group interpretation of epistemic modals. This flexibility, in the eyes of the advocates of canonical contextualism, greatly enhances the explanatory power and thus the attractiveness of the view.

However, critics argue that canonical contextualism is unable to accommodate many of our semantic intuitions about epistemic modal claims. Simply put, the challenge is this: no matter how flexible canonical contextualism is, it remains the case that the relevant body of information is uniquely selected by the context of utterance. But sometimes it is hard to see how a *unique* body of information, selected by the context, can capture all the relevant intuitions we have about epistemic modals. An example would make this point clearer:

> **KEYS.** Alex is helping her roommate Billy search for her keys. Alex asserts:
>
> (1) "You might have left them in the car."
> Billy responds:
>
> (2) "No; I still had them when we came in."
> Alex has no reason not to trust Billy. So she concedes:
>
> (3) "Oh, I guess I was wrong then."

In the beginning of this simple conversation, Alex brings up the possibility that Billy has left the keys in the car. Billy rejects it on the basis of his private information, of which Alex initially is not aware. Trusting what Billy says, Alex then retracts her earlier assertion.[1]

The problem arises when we ponder over the exact content of (1) and its relation with Billy's rejection and Alex's own retraction. What is the proposition expressed by (1)? On the solipsistic reading of "might", (1) expresses the proposition that it is compatible with Alex's knowledge (at the time of the utterance) that the keys are in the car. But then it is puzzling why it is appropriate for Billy to reject, on the basis of his information, Alex's assertion (1): after all, on this reading, Alex is only talking about her own knowledge. Furthermore, it is similarly unclear why Alex takes back her earlier claim, given that the key's being in the car was compatible with her knowledge back then.

So, perhaps, in asserting (1), Alex intends the group reading of "might". On this reading, (1) expresses the proposition that it is compatible with the knowledge of the group (which includes both Alex and Billy) that the keys are in the car. This would indeed solve the problems we have encountered above: Billy rejects this proposition because the keys being in the car is incompatible with his (and thus the group's) knowledge; and, after learning this piece of group information, Alex

[1] I will say more about the rejection and retraction of epistemic modal claims latter, but this seems to be a very straightforward and natural way to characterize the scenario.

impeccably retracts her earlier, false assertion. But the group interpretation faces another puzzle: what entitles Alex to assert (1) in the first place? As von Fintel & Gillies put it,

> . . . [Alex] does not seem to be within her linguistic rights to be claiming that the group's information cannot rule out the prejacent. After all, (she) does not now whether Billy has private information about the whereabouts of the keys. (2011, 116)

Of course, the solipsistic and the group readings are just two straightforward ways to implement the contextualist semantics. In principle, there could be other contextualist interpretations that are free of these problems. But it is at least not obvious what interpretation would fit the bill. At any rate, cases like this have motivated many to abandon the canon and look for alternatives. One of the most prominent alternatives that have been developed recently is semantic relativism. (Egan 2007; Stephenson 2007; MacFarlane 2011, 2014)

Semantic relativists contend that the truth-values of epistemic modal claims are not fixed once for all by the context in which they are uttered. Rather, their truth-values could vary across judges (or assessors, contexts of assessment, etc.), which is an additional parameter [2] that figures in the evaluations of epistemic modal claims. Different versions of relativist semantics have been proposed, but the basic idea can be summarized as follows: a.) an epistemic modal claim asserted in a context is always evaluated relative to (the information possessed by) a judge or assessor; b) there is no unique judge/assessor from whose perspective the claim should be evaluated; and c) since different judges/assessors may have different bodies of information, an epistemic modal assertion could be true relative to one judge/assessor, but false relative to another. So, for example, relative to the earlier Alex (given what she knows at the beginning of the conversation), the assertion (1) is true. This explains why she is

[2] In addition to standard parameters, such as world and time.

warranted in in asserting (1). However, relative to Billy and the later Alex (after her knowledge is updated in light of Billy?s response), that same assertion is false. This explains why it is appropriate for Billy to reject it and for Alex to retract it. On this view, the context of utterance sets no constraints on what information a judge/assessor has, or indeed who a judge/assessor might be. She need not be present at the context: we could be a judge/assessor of Alex and Billy's assertions, and relative to us, (1) is false. Given what we know about the scenario, we would judge the assertion to be false.

Or would we? This is an empirical question. This means that the question cannot be settled merely from the armchair, as much as we love it. Indeed, in my presentation of KEYS and the problems it poses for the canonical contextualism, I have tacitly appealed to quite a few intuitions about whether an assertion/response is appropriate or true. But are these intuitions really as robust as the critics of canonical contextualism assume? And even when they are robust, what exactly do they tell us about the semantics and pragmatics of epistemic modals? In the next section, I will review some recent empirical studies that bear on these questions.

4.2 Rejection, Retraction and Falsity

Relativists use a number of putative counterexamples, such as **KEYS**, to argue that canonical contextualism makes wrong predictions, or, in any case, cannot accommodate all the semantic intuitions elicited by such examples. Recently, a number of researchers tried to put these cases to test. Although this line of research is still in early stages, it has already uncovered some interesting results, some of which have the potential to weaken the force of the relativist challenge.

Consider, first, the following example, which is discussed by both Knobe and Yalcin (2014); Khoo (2015):

> **FAT TONY.** Fat Tony is a mobster who has faked his own death in order to evade the police. He secretly plants highly compelling evidence

> of his murder at the docks. The evidence is discovered by the authorities, and word gets out about his apparent death. Several forensic experts have carefully examined the evidence. Expert A says,
>
> (4) "Fat Tony is dead".
>
> Expert B has also carefully examined the evidence, but his assessment is more cautious. So he says
>
> (5) "Fat Tony might be dead".

In one of their experiments, Knobe & Yalcin first presented participants with the vignette above, and then asked them whether they think the assertions made by Expert A and Expert B are true or false. More precisely, they asked whether the participants agree with one of the following statements:

- (NONMODAL-TRUE) What Expert A said is true.
- (NONMODAL-FALSE) What Expert A said is false.
- (MODAL-TRUE) What Expert B said is true.
- (MODAL-FALSE) What Expert B said is false.

As it turned out, the participants' reactions in the modal case and nonmodal case were very different. On one hand, they strongly agreed with (NONMODAL-FALSE), i.e., that the nonmodal statement (3) is false. On the other hand, they were much less inclined to agree with (MODAL-FALSE), i.e., that the modal statement (4) was false. In fact, they were significantly more inclined to agree that (4) was true than they were to agree that it was false.

These findings, especially the last comparison within the modal condition, is in tension with a core assumption of semantic relativists that motivates their challenge: speakers, as judges or assessors, would regard an epistemic modal possibility claim ("Fat Tony might be dead") as true only if its prejacent ("Fat Tony is dead") is compatible with their knowledge (MacFarlane 2014, 257). In the present case, the prejacent is not compatible with the participants' knowledge, but they were strongly inclined to assess it as true.

Justin Khoo's recent study (Khoo 2015) confirmed this result. In a similar set-up, the subjects of his experiment tend to regard a non-modal claim to be false, but were much less inclined to regard the corresponding modal claim to be false. Another surprising result he discovers is that, while the subjects were not inclined to judge modal claim such as (4) to be false, they were strongly inclined to reject it by saying things like

(6) No, Fat Tony is alive. He faked his death.

Indeed, Khoo's data suggests that people were almost equally inclined to use (5) to reject Expert A and Expert B's assertions. That is, there is no significant difference between the modal condition and nonmodal condition when it comes to rejection.

Hence, people seem to think that, while some BEPs are not false, it is nevertheless appropriate to reject them. This observation can be put as follows:

- **Divergence:** When presented with cases like FAT TONY, ordinary speakers' intuitions about the falsity of a BEP and the appropriateness of rejection come apart: they are not inclined to assess the BEP as false, but are inclined to regard rejecting the BEP as appropriate.[3]

Interestingly, Knobe & Yalcin also discovered a parallel phenomenon. In another experiment, they investigated people's intuitions about retraction. Participants were presented with the following vignette, which is modeled on an example in MacFarlane (2011):

> **BOSTON.** Sally and George are talking about whether Joe is in Boston. Sally carefully considers all the information she has available and concludes that there is no way to know for sure. Sally says: "Joe might be in Boston."

[3] See also Khoo (2015, 520)

> Just then, George gets an email from Joe. The email says that Joe is in Berkeley. So George says: "No, he isn't in Boston. He is in Berkeley." (Knobe and Yalcin 2014, 14)

In the nonmodal version of the scenario, Sally says "Joe is in Boston." This experiment is designed to compare people's intuition about the falsity of a BEP and about the appropriateness of retracting that BEP. In the retraction condition, participants were asked to indicate their level of agreement, on a scale from 1 ("completely disagree") to 7 ("completely agree"), with the statement: "It would be appropriate for Sally to take back what she said." The participants in the falsity condition were asked to indicate their level of agreement with the statement: "What Sally said is false."

As in the previous experiments, participants were much less inclined to judge the modal claim ("Joe might be in Boston.") to be false than they were to judge the nonmodal claim ("Joe is in Boston.") to be false. More interestingly, Knobe&Yalcin found a divergence between people's intuitions about the falsity and appropriateness of retraction in (and only in) the modal case: they were significantly more inclined to agree that Sally should retract the modal claim, "Joe might be in Boston", than they were to agree that what Sally said was false. No such difference was observed in the nonmodal case, where people almost equally agree that Sally should take back her assertion "Joe is in Boston" and that it is false. This lends support for another type of Divergence:

- **Divergence 2:** When presented with cases like BOSTON, ordinary speakers' intuitions about the falsity of a BEP and the appropriateness of retraction come apart: they are not inclined to assess the BEP as false, but are inclined to regard retracting the BEP as appropriate.

These two forms of divergence present at least a *prima facie* problem to the relativists, whose objections against canonical contextualism are often built upon cases of rejection and retraction. They contend that such cases show contextualism makes wrong prediction about the truth-value of BEPs, but their arguments cannot be vindicated *merely* by invoking our intuitions about the appropriateness of rejection/retraction: if **Divergence** and **Divergence 2** are correct, the appropriateness of rejecting/retracting a BEP does not warrant its falsity.

This is not the place to adjudicate the contextualism-relativism debate. After all, each theory comes in many varieties and it is certainly possible that some versions of contextualism or relativism can, with some tinkering, accommodate these linguistic data. But I am inclined to think that the linguistic data themselves raise a more fundamental and independently interesting questions about our intuitive linguistic judgments: why do our intuitions about truth/falsity come apart from our intuitions about the appropriateness of rejection/rejections? What do these distinct and separable intuitions track? What do (or can) they tell us about the semantics of epistemic modals? It is hard to see how to choose among different semantic theories unless these prior questions about these intuitive judgments, which serve as the key evidence for our semantic theorizing, have been answered. In the next section, I will develop an account of BEP that will shed light on these questions.

4.3 The Double Lives of BEP

On a simple picture about the use of "true/false" in conversations, saying "P is true" is just expressing one's agreement with P or acceptance of P, and saying "P is false" is just expressing one's disagreement with P or rejection of P. The studies reviewed in the last section, however, suggest that this view is too simple, as least when P is a BEP. Moreover, as Khoo notes in his paper, even for simple, nonmodal statements, this picture is perhaps not entirely correct either. Someone agnostic about the existence of God disagrees with the believer's claim "God exists", but he does not thereby (unlike atheists) regard the claim as false.

Our linguistic practice of rejection, then, is more complicated that what might appear at first glance. Future theoretical and empirical research will likely reveal more unnoticed phenomena. This is as it should be. I do not intend my proposal to exhaust all the complexity and subtlety surrounding our use of BEPs. Instead, I want to start with the simple question: what are BEPs for? What communicative purpose do they serve? My hope is that by approaching BEPs from this angle, we will have a better grasp on our intuitive judgments about BEPs.

So what are we doing when we utter a BEP in a discourse context? Two things, I think. First, when I tell you that it might be the case that *p*, I put *p* into the "public space" as a live possibility that is worth noting in our inquiry, given our common theoretical (e.g., know the truth) and practical interests (e.g., find the keys). Second, through the epistemic use of "might", I indicate that although p is not (yet) established as a fact, it is at least left open by the relevant evidence or information that I have (If I believe that *p* is a fact, I would probably just assert *p* instead.) So besides putting *p* to the front, I also indirectly implicate something about myself. It is indirect because the sentence does not explicitly mention me or my information (unlike: "I think that p"). Still, it is my information that serves as the ground for my suggestion that p is an open possibility (for both of us).

Both points call for elaboration. Here is a slightly fancier way to expand on the first point. Let's say a proposition *p* is presupposed by a speaker S (at a particular time of conversation) iff S presumes that all participants of the conversation (at that time) take *p* for granted. The common ground of a conversation is the set of propositions mutually presupposed by the participants of that conversation. The common ground can be represented by a set of possible worlds (called the context set): the set of possible worlds in which all the propositions in the common ground are true. This is, of course, essentially the Stalnakerian model of communication (Stalnaker 1978, 2014). On this model, when a speaker asserts that *p*, she is in effect proposing to eliminate all the worlds in which p is false, and thereby to alter the common ground. Consequently, if other participants accept her assertion, then the updated common ground is no longer compatible with not-p.

We can extend the basics of the Stalnakerian theory to the communicative contributions of BEPs. If S asserts a BEP "it might be the case that p", she is in effect proposing that the common ground should be "friendly" to (at least some) p-worlds. That is, it should not only include not-p worlds. The exact intended impact of the assertion depends on the configuration of the common ground before the BEP is asserted: if originally it already had some p-worlds, then the suggestion is that they should not be all eliminated (unless opposing evidence is uncovered); if originally it included no p-worlds (e.g., it was mutually presupposed that p is false, but then S finds some new evidence indicating that p still might be true), then the suggestion is that the common ground should be expanded to include some p-worlds,[4] so that the updated common ground is at least compatible with p.

It is worth noting that this facet of BEP is primarily concerned with the prejacent: the main function of saying "it might be the case that p" is to suggest that p shouldn't be ruled out by the participants of the conversation, and that it is potentially significant for the aim of the conversation. In making this suggestion, the speaker is not concerned with her informational states *per se*, but with p itself. The primary life, then, that BEPs lead is an impersonal life, so to speak. So for example, when I claim that "The keys might be in the car", I'm conveying information with regard to the whereabouts of the keys, not myself or my information.

However, beneath the surface, BEPs have a personal life as well. As I noted above, our linguistic practice of uttering BEPs has a second facet: in typical cases, I

[4] Actually the story is a little more complicated. Asserting "It might be the case that p" is not just about retaining or incorporating p-worlds as such, but it is also to indicate, for Gricean reasons, that these worlds have a special theoretical or practical relevance to the aim of the conversation.

bring up a worth-noting possibility to you because I think it is at least left open by my evidence or information. It is a reasonable and potentially significant suggestion, given the evidence available to me. In this sense, the utterance is dependent upon my perspective. Through the use of the epistemic "might", I in effect–for the want of a better term–"signal" to you this perspective-dependency of my BEP utterance. Furthermore, it seems that, by implicating myself or my perspective in this way, I have also admitted my potential limitations that come with the perspective-dependency: I might be ignorant of certain important facts, I might be biased, or perhaps I simply have not reasoned carefully. In short, I might be wrong. Perhaps you know better. After all, this is why I choose to utter the BEP instead of its prejacent. But still, given that I am sincere and serious in suggesting that p is a (potentially significant) possibility that we should recognize, I believe p is a reasonable supposition for me to make.

So a BEP's personal life is not a *private* life: the utterer indicates, albeit indirectly, to her audience that she is speaking from her perspective, and that her perspective could turn out to be misguided or partial. If her audience is collaborative and sensible enough, they will recognize these indirect messages behind the assertion. So even when they refuse to accept the suggestion that p is a live option, because, say, they have some privileged information, they might still acknowledge, or least not deny, the reasonableness of her suggestion.

We can find some evidence for this personal life of BEP by considering what happens when things go awry. Here is a variant of KEYS:

> **NO-RETRACT.** Alex is trying to help Billy to find his keys.
> Alex: They might be in the drawer.
> Billy: (looks in the drawer, agitated) They are not. Why did you say that?
> Alex: Look, I didn't say they were in the drawer. I said they might be there–and they might have been. Sheesh. (Von Fintel and Gillies 2008, 81)

As far as I know, the folk intuition regarding this case has not been systemically tested; but Alex's response here sounds entirely unproblematic. In the literature, this example is mainly used to undermine the relativist arguments based on retraction: contrary to what the relativists predict, in some cases it is appropriate for the speaker to stick to her guns and not to retract her BEP in the face of new, opposing evidence. But few commentators have tried to explain why sometimes (such as **KEY**) it is appropriate, perhaps even required, to retract, while sometimes it is appropriate not to retract. I think my two-part story sketched above promises an answer to this puzzle.

In her response to Billy, Alex underscores the fact that her earlier claim is might-claim. Why this fact? It clearly has something to do with the way Billy dismisses her claim: he reacts as if what she suggests is totally unreasonable or even ridiculous ("The keys are simply not in the drawer. Why did you suggest otherwise?"); it seems he is even implicitly questioning her intention and her status as a collaborative interlocutor ("Are you trying to trick me by suggesting something false?") Bill's reaction shows that in some sense he fails, to use a technical term, to *get* what Alex tried to convey earlier. According to my analysis, what he misses, and thus what Alex feels need to be stressed, is the perspective-dependency of her claim. In her response, Alex aims to defend the reasonableness of her claim by emphasizing this aspect of her claim: it is a reasonable suggestion *for her* to make, from her (earlier) perspective, *given* the information available to her back then.

Let me take stock. In this section, I have argued that BEPs have a more salient impersonal life as well as a less salient but easily discernible (for ordinary speakers at least; Billy is an exception) personal life. More precisely, my contention is that when a speaker utters a BEP, she is doing two things: first and foremost, she is suggesting or proposing that some objective fact should be mutually recognized as a live and significant possibility; second and less directly, she is indicating or implicating the perspective-dependency of her suggestion or proposal.

One natural consequence of this two-level story is that the audience could respond to either aspect of uttered BEP. With regard to the first and primary aim of the speaker's assertion "It might be the case that *p*", one could accept the suggestion (that is, allow that the common ground be compatible with *p*); or reject it (that is, insist that p-worlds should be ruled out from the common ground). Since making this sort of suggestion is the main aim of uttering a BEP, it is also the target of typical responses to BEPs. Indeed, the earlier examples of rejection we have seen, sentences (2) and (5), are exactly this kind of response.

Occasionally, though, the audience's rejection/acceptance can be directed at the perspective-dependent aspect of the assertion. Presumably, this happens when the reasonableness or appropriateness of the suggestion is at issue. Billy in NO-RETRACTION might not be sensitive to the perspective-dependency of Alex's suggestion, his response to it nevertheless is perceived as questioning its reasonableness, and that is what Alex seeks to defend in her response. A more straightforward case in which the audience's acceptance/rejection targets the reasonableness is harder to come by, since whether the suggestion is reasonable for the speaker to make is not the main point of such communication. But consider the following example:

> **CONSPIRACY.** Charlie and Diane are friends. Charlie is a kind, hard-working, and, for the most part, pleasant man who believes in a few conspiracy theories: the 9/11 attack was an inside job, LBJ plotted the assassination of JFK etc. (He does online "research" in his spare time, and has been to many conspiracy conventions and all that) Diane, on the other hand, believed none of these, but from time to time she enjoyed talking with Charlie about them, if just for fun. Today the topic is 9/11, and United Flight 93 in particular. Charlie believes that the plane was not hijacked, but was shot down by a U.S fighter jet. "What about the phone calls?" asks Diane, referring to the phone calls made by some passengers before the crash, in which they told their families or friends that the plane was hijacked. Charlie thinks that they were faked. He has heard that there is some program that could record and edit people's voices.

> But as it happens Diane is a specialist on voice morphing technology, so she explains to Charlie how the technology actually works, and why it is practically impossible to do the kind of editing required by Charlie's hypothesis. Charlie by and large believes that what Diane says, but it has not alleviated his doubts (perhaps the FBI just have better secret technology?). So he says,
>
> (7) "Okay, but still, the phone calls might be faked."
> Frustrated, Diane responds,
>
> (8) "No, that's not true!"

It seems to me that Charlie's assertion (7) is false, and that Diane's response is entirely appropriate. If I am right, then this example stands in an interesting contrast with FAT TONY, where people are much more inclined to judge that the relevant modal claim is true (although they agree it should be rejected).

There are some reasons to think that Diane, in uttering (8), is not (merely) denying that the calls' being faked is a live possibility. She has made her position abundantly clear in the previous discourse. Moreover, she has offered, to the best of her knowledge, strong evidence and considerations to the contrary, in the hope that Charlie would accept what they overwhelmingly suggest. In doing so, she is not just trying to explain her stance: what she does is not so much about justifying her rejection than getting Charlie to see why he should abandon the conspiracy theory. She wants to show, in other words, that given the evidence and considerations, Charlie (or indeed any reasonable human being) should rule out the possibility that the calls were faked. Charlie, however, is not entirely convinced: what he means by (7) is that despite all the evidence and considerations that Diane has just told him, the calls' being faked is still a serious, genuine possibility. And in saying (8), Diane is responding to the reasonableness of this claim. It is as if saying "Now that I've told you all the things about voice morphing, it is not reasonable for you to hold onto that conspiracy anymore, not even as a genuine possibility."

Suppose my analysis of **CONSPIRACY** is on the right track. Where does this lead us? Recall that in the previous section, I reviewed some evidence for **Divergence** and **Divergence 2**: our intuitions about truth/falsity of a BEP and the appropriateness of rejecting/retracting a BEP come apart. In this section, I argue that a BEP has two distinct communicative functions. Wouldn't it be nice, you might wonder, if these two observations line up with each other?

Indeed. I think observations like **Divergence 1** and **Divergence2** provide yet another piece of evidence for a two-level account of BEP. In **CONSPIRACY**, a truth-value judgment[5], (8) is used to deny the reasonableness of a BEP. A similar story can be told about cases like **Fat Tony** and **Boston**. The basic idea is this: on the one hand, our intuitions about the appropriateness of rejecting/retracting a BEP are concerned with the impersonal life of the BEP, as whether one rejects or retracts a BEP is primarily determined by whether one takes a putative fact to be a real, live possibility. On the other hand, in cases like **FAT TONY** and **BOSTON**, the intuitions about the truth of a BEP track the personal life of the BEP: whether one regards a BEP as true or false is primarily determined by whether one regards the BEP as reasonable, given the perspective of the speaker of that BEP. For example, Expert B's assertion "Fat Tony might be dead", is rejected because the suggestion should not be accepted as a live possibility, for the purpose of the investigation; however, it is nevertheless regarded as true because it is (or so it seems to us) a reasonable hypothesis, given his perspective and the evidence available to him.

Note that I am not saying that ordinary judgments of truth and falsity of a BEP *only* respond to its secondary, perspective-dependent facet. Things are unlikely to be black or white for abstract terms like "true", "false". In ordinary discourse, they may well be ambiguous in that they can serve to express different kinds of positive

[5] You might think that "truth"/"falsity" in compositional semantics are technical terms that need not match onto ordinary speakers' use of the terms, so we should not assume that (7) expresses the kind of truth-value judgment relevant to compositional semantics. Good point. I admit that "true" in ordinary language could admit different uses, but in this particular example I don't see why we should not take it at face value.

or negative responses. It is certainly possible to use them to express acceptance or denial of BEPs. However, my goal in this section is not so much to characterize the functions of ordinary judgments of truth/falsity as to distinguish two aspects of the communicative imports of BEP. It thus suffices for my purpose to show that the otherwise surprising patterns of folk truth-value judgments, shown in the previous section, can be easily accommodated in this framework. In the appendix I will examine an alternative semantic framework, proposed by Von Fintel and Gillies (2011), that also postulates a two-level structure. By revealing some of its problems, I aim to show the strength of my account, as well as its distinctive features.

So far I have been focused on what the speaker is trying to convey with her utterance of BEP, but it is worth emphasizing that the two-level structure of BEP is not only manifested on the speaker's side. As noted earlier, the hearer can assess a BEP along these two dimensions as well: she can either reject or accept the suggestion made by the BEP, and she can also respond to the perspective-dependent reasonableness of the BEP. I have argued that, although the first facet of BEP figures more prominently in ordinary communication, we are, or at least are expected to be, sensitive to the second facet as well (and there is some *prima facie* evidence that, at least in some contexts, our judgments of truth/falsity of a BEP reflect our evaluations of its perspective-dependent reasonableness.) This means that the ability of perspective-taking is critically implicated in the assessment of and response to BEPs: to assess whether an uttered BEP is reasonable from the speaker's perspective, we need to first try to take on her perspective and see things from her angle. Hence, in the remainder of this chapter, I will take a brief tour to the relevant psychological literature (section 4) and examine its implication for my proposal (section 5).

4.4 The Psychology of Perspective Taking

Sometimes I find myself tapping unconsciously the rhythm of a boy band ballad. Once I realize it, I immediately stop tapping and start to feel embarrassed, fearing that the people around will think that I have a bad taste in music. After all, I think, it is relatively easy to figure out from the rhythm what the song is, as it is petty well-known.

My fear is probably unwarranted. In a very comforting study (Newton, 1990), a group of participants were asked to tap the rhythms of some well-known songs, while another group was asked to listen and identify the songs on the basis of the rhythms. As it turned out, the listeners' actual probability of correct identification was extremely low (about 0.025). On the other hand, I am in good company too: on average, the tappers in the experiment greatly overestimated the probability of correction identification (about 0.5) The reason for this large difference, according to some theorists (Griffin and Ross, 1991), is that when they tapped the rhythms, the subjects rehearsed the music renditions in their heads; but they (perhaps at the sub-personal level) failed to recognize the differences in the subjective experiences between them and the listeners, and thus overestimated the likelihood that the listeners would be able to identify the songs.

This example reflects a general tendency of human cognition: in assessing or predicting others' actions or decisions, we tend to overimpute our privileged information (i.e., information only accessible to us) onto them. This creates a great challenge for perspective-taking, which has been dubbed "the curse of knowledge". (Birch, 2005), or "epistemic egocentrism" (Royzman et al., 2003). For example, the difficulty in setting aside one's private information, even when one knows that the information to be unattainable to the other party, is largely responsible for the fact that "one's prediction of another's perspective becomes skewed toward one's own privileged viewpoint." (Royzman et al., 2003, 38).

Epistemic egocentrism is a common feature of human cognition, and it is well documented in various branches of contemporary psychology. Here I will briefly mention two more examples for the purpose of illustration, one from developmental psychology, and the other from social psychology, but it is worth bearing in mind that the prevalence of EE is supported by a wide range of studies across different paradigms (for review, see Nickerson 1999; Royzman et al. 2003)

In the "Theory of Mind" literature, psychologists found that young children seem unable to understand that other individuals have a different perspective from their own (or at least they seem unable to use that information to predict their behaviors in verbal tasks[6].) For example, in the classic version of the false belief task (Wimmer and Perner, 1983), a child is shown two dolls, Sally and Anne, playing with a marble. Sally puts the marble in a basket and leaves the scene. Then Anne moves the marble from the basket and hides it in a box. The child watches all these steps. Lastly, Sally returns to the scene, and the child is asked of the question: "Where will Sally look for the marble?" Children under 4 typically fail the task: most of them said that Sally would look for the marble in the box, where it actually is. Apparently the children's own privileged perspective have interfered with their predictions. That is, they failed to set aside their knowledge and take on the perspective of Sally: as far as she is concerned, the marble should be in the basket, since she does not know that the marble was moved. Hence, this failure could be seen as an instance of epistemic egocentrism.[7]

[6] In fact this is a bit oversimplified. For one thing, on my use of the term, an individuals "perspective" include her perceptions, intentions, beliefs, knowledge, emotions, all of which are partially depend on her interaction with the world. But it is very unlikely that children's understanding of them come in one package. For another, since Onishi and Baillargeon (2005) and many works inspired by it, many developmental psychologists now agree that younger children (1 or 2 year-olds) manifest, in non-verbal tasks, at least some capacity to think and reason about other individuals' mind. For a recent review of this literature, see Baillargeon et al. (2016)

[7] Another example: older children often exhibit epistemic egocentrism as well. Taylor et al. (1994) found that when told a novel piece of new information, or a new skill, , they were likely to judge

Normal adults, of course, have no problem understanding that other people might have different beliefs and preferences. But even for them, it is difficult to inhibit their privileged information when they assess or predict others' behaviors. For example, we tend to overestimate how likely others are to detect that we are lying (Gilovich et al., 1998). Or consider a study more relevant to our topic: linguistic understanding. In Keysar's (1994) study, subjects read the following vignette: Mark asked his office mate, June, to recommend a restaurant to which he could take his visiting parents to dinner. June recommended a new Italian restaurant called Venezia: "I just had dinner there last night and it was marvelous. Let me know how you all enjoy it" (173). So Mark and his parents ate at Venezia later that day. The participants were given different information about the food and service at that occasion. The first group of participants was told that both the food and service were just mediocre; the second group was told that they were superb. In both conditions, it was made clear that the information was not available to June (so it is privileged information for the participants). In the end of all these scenarios, Mark wrote to June: "You wanted to know about the restaurant, well, marvelous, just marvelous." (173)

The participants could interpret this comment itself as either sarcastic or sincere, depending on the version of story they have read. The task for them was to judge whether the uninformed addressee, June, would perceive the comment as sarcastic. Presumably, to perform this task properly, they need to discount their privileged information: they need to take on June's perspective and determine whether she would believe it to be sarcastic. Their own knowledge about Mark's experience was simply irrelevant for June's perception. However, the results suggest that the privileged information was not easily discounted: while only 3% of participants in the second group (who learned that Mark and his parents' dining experience was positive) judged that

that other children would know it too, and that they themselves have known it all along. Here they impute their current knowledge not only to others, but also to their earlier selves.

June would regard the comment as sarcastic, 27% of the participants in the first group (who learned that Mark and his parents' dining experience was negative) judged that she would regard it as sarcastic.This seems to be another manifestation of cognitive effects of epistemic egocentrism: those who knew the comment was sarcastic were more likely to judge that the uninformed addressee would think the same (even when they knew that the addressee did not share their information).

The next thing I want to discuss is the psychological *mechanisms* or *processes* of perspective-taking. Here I want to focus on an influential account of perspective-taking that is sufficiently general and promises to explain EE: the so-called "egocentric anchoring and adjustment heuristics". This account is articulated and defended in Epley et al. (2004), but its basic ideas can be traced to Tversky and Kahneman (1974), and are relatively well-known in the "heuristics and biases" tradition in social psychology. According to this account, people adopt others' perspectives by serially adjusting from one's own (i.e., the anchor). That is, they use their own perspective as a starting point, and "only subsequently, serially, and effortfully account for differences between themselves and others until a plausible estimate is reached." (Epley et al., 2004, 383).

To further explain this model and its explanation of EE, let us consider one of their studies (Study 5), which can be seen as a more complicated version of Keysar's (1994) experiment. In this study, participants read that the protagonist, Tom, went to a comedy show his friend strongly recommended, and made a comment about it in his email. Tom appeared to be praising the comedian in the email, but again, the comment can be read as sarcastic or sincere. In the positive information condition, participants were told that Tom had a great time at the show; in the negative information condition, participants were told that Tom really hated the comedian. They were then asked to imagine that "100 people were asked to read Tom's email message", who knew nothing about Tom's experience at the show, and to estimate the "percentage of these people who would interpret the message as sincere versus sarcastic." (335)

The pattern of the results was similar to the Keysar study. Participants in the negative information condition estimated that a smaller percentage (M=60.78%) of those who read Tom's email would interpret the message as sincere than those in the positive information condition (M=78.11%). More interestingly, Epley et.al used another manipulation, in which they asked another two groups of participants to estimate the range of the percentage of people who read the message would interpret it as sincere. The average range provided by participants who were in the negative information condition was from 59.26% to 73.16%; the average range provided by participants who in the positive information condition was from 64.10% to 78.32%.

These result seem to suggest that there is a *stop-rule* for perspective-adjustment: when subjects in the negative information condition were asked to estimate the percentage of the people who would interpret the message as sincere, they started with their own judgment (i.e., it was not sincere), and skewed their estimation toward the opposite direction to accommodate the difference between themselves and the interpreters (who were not aware of Tom's experience at the show), and stopped right after the estimation reached the lower boundary of the apparent plausible range (that is, plausible from their perspective); likewise, subjects in the positive information started from their own judgment (i.e., it was sincere) to skewed their estimation toward the sarcastic end, and stopped right after it reached the upper boundary of the apparent plausible range.

Therefore, in both cases, the resulting estimations were closer to the respective starting points, i.e., the egocentric anchors. This model thus neatly explains why the subjects' judgments exhibit epistemic egocentrism. As stated, the model is very sketchy, and there are crucial questions that a full-developed version should address (e.g., how does the adjustment work?). Nevertheless, I think this model has enough intuitive content to enable us to make some concrete predictions about our evaluations of BEP.

4.5 Predictions of the Two-Level Account

My hypothesis, introduced in section 3, states that whether one regards a BEP as true or false is primarily determined by whether one regards the BEP as reasonable, given the perspective of the speaker. This means that while considering the (perspective-dependent) reasonableness of a BEP, we need to take into account the perspective of the speaker. Now *if* our perspective-taking is subject to EE (i.e., epistemic egocentrism), and in particular if perspective-taking works in the way that "anchoring and adjustment" model prescribes, then it is natural to expect that our intuitive judgments that track the perspective-dependent reasonableness of BEPs would, to some extent at least, exhibit EE as well.

We thus have a straightforward way to test my hypothesis. The general prediction is this: the more difficult it is for the hearer to adjust to the speaker?s perspective, the more likely it is for her to unreflectively impute her privileged knowledge onto the speaker, and consequently, the more likely it is for her to assess the reasonableness of the speaker's assertion on the basis of her own (the hearer's) knowledge. In such conditions, if the assertion is reasonable from her own perspective, she is more likely to regard it as true; if the assertion is unreasonable from her own perspective, she is more likely to regard it as false.

To put my proposal to test, then, we would need a scenario where such conditions obtain. Perhaps the most obvious candidate is one that involves a kind of *information asymmetry* between the participants of conversation: the hearer has all the relevant information, but falsely assumes that the speaker has it too. Given epistemic egocentrism, it is then likely that the hearer would projects her knowledge onto the speaker. Consider, for example, a variant of **FAT TONY**. In this variant, the "evidence" secretly planted by Tony in order to mislead the police were in fact very messy and uncompelling. Expert A, after carefully examining the evidence, quickly dismissed them; moreover, he then found other evidence that overwhelmingly suggest that Tony was alive. So he thought the case was now concluded and filed a report. The next day, however, Expert A ran into Expert B, who happened to be investigating the same case. Expert B brought up the case and said:

(5) "Fat Tony might be dead. "

How would Expert A respond to this? Presumably, he would reject it. And more importantly, it seems appropriate (if somewhat unnatural) for him to reject it by saying: "That's false!" Likewise, it seems that we, as extra-contextual evaluators, would probably agree as well that what Expert B said is false. At the very least, compared to the original version of **FAT TONY**, I am more inclined to agree that what he said in this situation is false.

But now suppose this is only half of the story. The rest unfolds as follows: like Expert A, Expert B was an experienced, smart investigator. However, as it turned out, the evidence available to him was secretly planted by Fat Tony's son, Slim, who was much more careful and shrewd than his father. That evidence was apparently very compelling, and thus no wonder that Expert B thought that Fat Tony might be dead.

Now that I have told the complete story, doesn't the intuition that (4) is false seem to have weakened somewhat? Expert B's side of the story is *de facto* identical with the original **FAT TONY**. In both cases, Expert B had some apparently compelling, but ultimately misleading, evidence, and the conclusion he arrived at, i.e., that Fat Tony might be, is reasonable and cautious given that evidence. Furthermore, we can make the cases more similar by supposing that the evidence that Slim planted was so strong that it overrides Expert A's evidence (So after some thought, even Expert A himself conceded, albeit reluctantly, that Expert B was right.) Assuming that the results of (Knobe and Yalcin, 2014) and (Khoo, 2015) are generally reliable, we should expect that most would be strongly inclined to agree that what Expert B said is *true* when they are presented with this version of the case.[8]

At any rate, if my argument is on the right track, the hypothesis I sketched earlier would have a ready explanation of these intuitions. When we are told the first half of the story, which only gives information about Expert A's investigation, we strongly feel that Expert B's assertion (4) is false, and it is appropriate for Expert A (who didn't know, at this point, about Expert B's evidence) to reject that assertion by saying that it is false. According to my hypothesis, we are sensitive to the perspective-dependent reasonableness of BEPS, and, in some cases at least, we use truth-value judgments to express our assessment of this aspect of the uttered BEPs. Now since epistemic egocentrism is a common feature of human cognition, it is perfectly natural for Expert A to evaluate Expert B's assertion (4) on the basis of his own information, especially when he did not know, and had no reason to assume, that Expert B had strong opposing evidence. For the same reason, we, the readers of the story, were similarly inclined, even though we were ignorant of Expert B's epistemic situation. Things change when additional background information about Expert B's investigation is revealed. Just like in the original case, we are now less inclined to judge that Expert B's assertion to be false (and more inclined to judge it to be true). This is because, on my hypothesis, the additional information enables us to better adjust to the perspective of Expert B than before. While we know that Fat Tony is alive, the new information helps us to better appreciate his epistemic situation and

[8] Of course, what their actual responses would be is up for actual empirical tests, and so far I have only presented little more than some thought experiments and speculations based on the available data. Moreover, there could be other complexities or confounding factors I have ignored in framing the case. But folk intuitions about these variants are clearly testable, and we only need minimal modifications in the original design.

limitation. From this adjusted perspective, it does seem that his overall judgment about the situation is reasonable and (thus) blameless.

Lastly, let me briefly mention two other types of cases that might be relevant. First, in general our moral intuitions can affect our modal thinking (Phillips and Knobe, 2018), and with respect to perspective-taking in particular, it seems sometimes we fail to take on the perspective of others due to our concerns about morality. For example, if the speaker himself, or what he said, is morally repugnant, the hearer might be disinclined to adopt the speaker's perspective (even if there is no information asymmetry).[9] Second, since perspective-shifting takes time and sometimes deliberate effort, time pressure can impede the process. So when the hearer is asked to evaluate the speaker's modal claim in a very limited time, she is more likely to rely on her private information.

[9] Suppose that Finn, a greedy businessman, wants to build on a beautiful river a factory that manufactures dangerous chemicals. The chemical waste, once spilled into the river, will seriously pollute the water and the surrounding environment. Gardner warns Finn of the non-trivial likelihood of the disaster, especially given the history of Finn's factories (suppose most of his factories and chemical plants have had such accidents, some very serious). Finn knows all the empirical facts, but he couldn't care less about the environment, so he shrugs, "Still, this one might be safe". In cases like these, if the hearer is less inclined to take on Finn's perspective for moral reasons, she will be, according my account, more inclined to judge that Finn's assertion to be false. I expect this inclination to be less strong in a morally neutral (and otherwise analogous) case, where no moral intuitions are activated to affect perspective taking. If such contrast is borne out by empirical evidence, this will provide additional support for my hypothesis.)

APPENDIX

VON FINTEL & GILLES' NONSTANDARD CONTEXTUALISM

Kai von Fintel and Anthony Gillies (henceforth vF & G) have recently proposed a new theory of epistemic modals that aims to meet the relativist challenges, while preserving some of the core elements of canonical contextualism (Von Fintel and Gillies, 2011). Their proposal is rich in ideas and innovations, but, as we will see, it also faces problems of its own. Nevertheless, since their theory bears some structural similarities with my two-level account, I believe that a close examination of their theory is not only of independent interest, but also will help to clarify various aspects of my account.

Like the canonical contextualist, von Fintel and Gillies maintain that an epistemic modal quantifies over the information available to a group of contextually salient agents. However, and here comes the key innovation of their brand of contextualism, they contend that when many participants are involved in the conversation, the context does not determine a *unique* group whose information is relevant; instead, there are in principle many legitimate ways to draw the boundaries. For example, the relevant group could be just the speaker, or just the hearer, or the group that include the speaker and the hearer. In other words, there are (or at least could be) different contextually admissible ways to specify the relevant information that a modal quantify over, and as a result, the BEP utterance is contextually ambiguous, and essentially so. In the words of vF&G: "There is no such thing as "the context", only the contexts admissible or compatible with the facts as they are. The context of the conversation really does not provide a determinate resolution and . . . there is a cloud of contexts at the given point of the conversation.” (Von Fintel and Gillies 2011, 119-119)

The most striking feature of vF & G’s account is perhaps the systematic and ubiquitous ambiguity of epistemic modals. But what interests me most is the underlying two-level structure of BEP utterance (not unlike mine, as we will see) they postulate. According to vF & G, each of the contexts in the “cloud” provides a resolution for the BEP, and thus an uttered BEP has many meanings. However, this does not imply that by a single utterance, the speaker somehow manages to assert many different propositions. They instead posit a distinct kind of speech act, the effect of which is to “put into play” or “float” some propositions that are not asserted. The upshot is that, when the speaker utters a BEP, she puts into play multiple propositions (corresponding to the different ways of disambiguations), while only asserting one proposition.

Now I have to say that I am not entirely clear what “putting into play” or its communicative import exactly is. But at the very least, given their characterization, its force seems to be in some sense weaker than assertion: when the speaker makes multiple propositions available for uptake and denial, she does not have to be in position to assert each and every one of them. It might help to consider how their account works in a concrete example, e.g., **KEYS**. Since only two participants, Alex and Billy, are involved in that conversation, there are three possible groups whose information could be relevant for the interpretation of Alex?s utterance, “You might have left them in the car.”: {Alex}, {Billy}, and {Alex, Billy}. According to vF & G, three readings are put into play: a solipsistic or speaker-centric proposition (i.e., that it is compatible with Alex’s information that Billy have left the keys in the car), a hearer-centric proposition (i.e, that it is compatible with Billy’s information that Billy have left the keys in the car), and a group proposition (i.e., that it is compatible with information of Alex and Billy, qua a group, that Billy have left the keys in the car). But Alex is only justifiable in making claim about her own evidence, so only one proposition, the solipsistic reading, is asserted.

So, on the speaker's side, Alex asserts the solipsistic proposition while putting into play three distinct propositions. But on the hearer's side, which proposition is Billy responding to when he rejects Alex's utterance? We have seen that Billy has no reason to deny the solipsistic proposition, which, after all, is only about Alex's own information. The group reading seems to be a more plausible target of his rejection, and vF & G agree. The moral is that the hearer need not to respond (only) to asserted propositions; vF & G hold that it is sometimes more appropriate to respond to the unasserted propositions that are put into play. More generally, they argue that the hearer can confirm (deny) the BEP if, among the propositions that the speaker puts into play, the strongest and most informative proposition that the hearer "can reasonable have an opinion about is such that [the hearer] thinks it is true(false)." (121) In **KEYS**, the most informative proposition that Billy reasonably has an opinion about is the group reading, which concerns the information available to them as a group. So that is what Billy denies: since he knows that the keys are not in the car, and given what Alex says, he can reasonable infer that it is not compatible with the information available to them that the he have left them in the car. VF&G summarizes their analysis of the exchange as follows:

> . . . the BEM itself has three meanings, but Alex acts as if the solipsistic reading were the one that matters, while the hearer acts as if the hearer-centric or even the group reading were the active ones. This asymmetry is what gives BEMs their quasi-magical properties: a speaker can utter them based on just her own evidence but it serves as a probe or test or trial balloon into the hearer's evidence. When things go well and a hearer takes up a BEM, this fact becomes common ground between speaker and the hearer and thus it follows that it is common belief between them that the prejacent is compatible with the information they qua group have. (123)

It is perhaps not too difficult to notice the structural and substantive similarity between vF & G's proposal and mine. First and foremost, both proposals recognize two distinct communicative contributions of BEP utterance, a more objective one and a more subjective one. On my account, when a speaker utters a BEP, she is 1) suggesting that some objective fact should be mutually recognized as a live and significant possibility, 2) indicating the perspective-dependency of her suggestion or proposal. On vF&G's proposal, Alex puts into play a group-proposition of the BEP while asserting the solipsistic proposition about her own information. Moreover, of the two kinds of communicative functions, the more objective one is more important for the main purpose of the conversation, and is also what the hearer responding to.

For this reason, it might seem that these are just slightly different way to implement the same idea. However, interesting differences would emerge once we start to dig into the details. According to vF & G, a single BEP utterance has multiple meanings, one of which is asserted, and others are merely put into play. For example, when Alex brings out the possibility of the keys' being in the car, she puts into play a set of three propositions, although she is only in position to assert the solipsistic proposition about her own information. Two general questions suggest themselves: first, which propositions are merely put into play? Second, and relatedly, why is the speaker entitled to put into play a proposition that she does not (and is not entitled to) assert?

It is not clear to me how they would address the first question. They contend that the context, or more precisely, the cloud of contexts, provides various resolutions of the relevant parameter, and thereby makes available variable propositions to be floated or put into play. But they have said little with regard to how the boundary of the "cloud" is drawn, which context is admissible, etc. They seem to suggest that, among these floated propositions, the speaker is only justified in asserting the ones for which she has the sufficient evidence. For example, for Alex's utterance (1), there are three groups, {Alex}, {Billy},{Alex and Billy}, whose information could be relevant for the interpretation, and Alex is justified in asserting only the solipsistic reading. But what makes a proposition available to be put into play? Judging by their discussion of the cases, it seems what propositions can be put into play is partly a function of the number of participants. If more people are involved in the conversation, then more readings might be available. But certainly factors other than number are crucial too, otherwise it would lead to almost boundless proliferation of the propositions put into play in many ordinary contexts.[1] It is not clear, though, what those factors might be, as vF & G have offered no principled explanations about which factors, linguistic and otherwise, figures in the determination of the set of propositions that are put into play or "floated".

We have encountered the second question earlier, albeit in a slightly different from. Recall that, in their discussion of canonical contextualism, vF & G question the entitlement of Alex's entitlement to speak on behalf Billy's and thus the group's information. The same question can be raised against their proposal too: given that Alex is not justified in asserting the group proposition, what entitles her to put into play the proposition? Again, it is unclear what vF & G might say about this. At one point, vF & G propose the following pragmatic principle to connect what is asserted and what is put into play:

- **ASSERT:** Suppose an utterance of might(B)(φ) by S puts into lay the propositions P1, P2, Then S must have been in position to flat out assert one of Pi's. (120)

The idea is that we are entitled to put into play a set of propositions only when we are entitled to assert one proposition in the set. However, this principle only gives a necessary condition for putting into play a set of propositions, not a sufficient one.

[1] If, say, the speaker is talking to a group of 10 people, it is hard to imagine how she could put into play so many propositions, including Addressee1-reading, Addressee2-and-3-reading, Addressee-2-5-7-8-9-reading, by a single BEP utterance, and more importantly, how these proposition could be relevant for the purpose of conversation.

So it does not follow from the fact that the speaker is entitled to assert one of Pi that she is thereby entitled to put into play the whole set. Hence, **ASSERT** fails short as an explanation of, say, Alex's entitlement to put into play the group-reading. Furthermore, even if the condition is both necessary and sufficient, we would need to know which proposition are in the set includes in the first place, otherwise it is impossible know *which* proposition can be justifiably put into play. So it seems the solution to the second problem presuppose a solution to the first.

There are other difficulties if being justified in asserting one proposition is sufficient for putting into play the whole set (assuming, of course, that we know what the set includes). Suppose that, when Alex says to Billy "The keys might be in the car." she is justified in thinking that Billy's evidence does not rule out prejacent, then, presumably, she is justified in asserting the modal proposition that makes a claim about Billy's information, i.e., the Billy-reading of the uttered sentence. Now if being justified in asserting any proposition of the set suffices for putting into play the whole set, then by that utterance, Alex would also put into play the solipsistic reading and group reading (assuming they are members the set). But what it is to put into play a solipsistic proposition about one's own information?

A lot of the problem, I think, arise because of the vagueness of vF & G's key notion, "put into play". As a general notion, it is often explicated metaphorically, ("float", "travel"), and it is very hard to pin down exactly what kind of speech act it is supposed to be, or what communicative purpose it serves. By contrast, speech acts like suggesting or proposing are much more easier to understand, and can be made more precise on the standard Stalanakerian model of conversation.

We might hope that vF & G's use of the notion in specific examples, such as **KEYS**, can provide more clues about what they have in mind. Of the example, vF&G says "When Alex puts the three propositions into play, the other readings, the hearer-centric reading and the strong group reading, are floated–It is as if she is conjecturing that the B-reading and the A+B reading are true or asking whether they are true." (121; my emphasis). It is telling that they analogize the communicative function of propositions put into play to conjecturing and asking. Conjecturing and asking about a possibility are tentative in a way that makes it different from asserting that something is compatible with one's evidence, and more like suggesting/proposing that the possibility is taken as live and significant (given the interests of the conversants). They are tentative in nature because they relevant possibility is not taken as an established fact, but a reasonable and informed guess. Moreover, the speaker does not seek to hide the tentative character; instead, she exploits it, and uses the BEP "as a probe or test or trial balloon into the hearer's evidence"(123), as vF & G put it.

Hence, I think perhaps the considerations that motivates vF & G's idea of "put into play" is very close to mine. Both accounts attempt to capture nonassertive and tentative character of part of the communicative import of BEP. Indeed, I am inclined to think that best way to interpret what they have in mind when they talk about "put into play" is along the lines that I have sketched in the previous section, namely, as a kind of suggesting or proposing. However, as they proceed to flesh out the notion, it departs from my account in various ways.

One salient difference is that, on my account, when a speaker utters a BEP, she is only suggesting one possibility. This is more economic than what vF & G's account delivers: on their account, the speaker is putting into play a whole set of propositions, and, depending on features of the context (e.g., the number of people involved in the conversation), the set could be extremely large. But the more complex the set becomes, the harder it is to make sense of communicative import of all the propositions that are allegedly put into play—that is, the harder it is see which proposition are put into play and why. Moreover, even in simpler examples, some floated proposition might be redundant. Recall that, in **KEYS**, Alex is said to put into play both the Billy-reading and the group-reading. However, if what matters is the group's search for the keys, as they vF&G concedes, it is hard see what additional work is done by the floated Billy-reading, over and above the group-reading, as for as the communication is concerned.

To make things worse, it questionable whether the group reading *per se* is needed at all. We can see traces of this problem in vF & G's own discussion of **KEYS**. On their analysis of the exchange, what Billy rejects is the floated group proposition, not the asserted solipsistic proposition. However, when they try to explain the underlying rationale, things become a little complex. According to vF&G, rejecting the group proposition is the more appropriate and cooperative thing for Billy to do because the goal of the conversation is to find out where the keys are, not "to find out whether the speaker's evidence or the group's evidence at the time of the conversation rules out the keys are in the car"(121; my emphasis). This explanation, however, is puzzling in that it is in apparent tension with their official story, because now it seems that the group reading of the BEP, i.e., the proposition about the information available to the group, does not really play any substantive role in the communication. Instead, what matters for conversation participants is the objective facts in the world (e.g., where the keys are), not the information of the group (i.e., what the group information rules out). Consequently, what the hearer rejects or accepts should not be a claim about the group's evidence either.

So, it seems, their analysis of this particular case does not sit well with the official doctrines of their theory. Hence, while I am sympathetic with their diagnosis of the case, especially the statement, quoted above, about the goal of the conversation, I believe my proposal offers a more accurate and straightforward way to capture the basic idea: when the speaker utters a BEP, what she suggested is not about the evidence or information of the group, but about things in world. On (the first half of) my account, it is as if saying that the possibility that a certain fact obtains is a reasonable, significant possibility for the group to consider. Importantly, it is not a claim about the group's information. I think this marks perhaps the deepest difference between vF & G's account and mine. Both accounts recognize a more subjective dimension of BEP and a more objective one, and thus both posit a two-level structure that reflects this. However, objectivity is construed differently. On their account, objectivity is manifested in inclusiveness: the group reading concerns the information available to a larger set of people that includes the speaker. On my account, however, objectivity is understood in terms of perspective-independency: the primary function of BEP is making a suggestion, but the suggestion is about objective facts in the world, not about any participant's (and thus the group's) information. In other words, when a speaker utters *Might p*, the content of the suggestion is *p*. The seems to better capture the point of the modal conversation, if we consider the more concrete cases: participants are primarily interested in objective facts in the world (e.g., where the keys are), not in what is excluded from their information. My account is thus not only more parsimonious, it also offers a more adequate explanation of the actual uses of BEP in typical contexts. In fact, nothing in vF & G's discussion of cases gives us any reason to prefer the first way of construing objectivity; if anything, it is the opposite, as their analysis of the Alex-Billy exchange seems to suggest.

To sum up: in this appendix, I have compared and contrasted my account of BEP with vF & G's structurally similar account. Their account is both intriguing and innovative, but it is often too schematic and, as mentioned above, leaves some serious questions unanswered. To be fair, I think their discussions of concrete examples are often very plausible, but many of their insights can be easily (and better) accommodated by my account. Moreover, my two-level account does not have the complex and apparently redundant features their cloudy contextualism seems to require.

BIBLIOGRAPHY

Allen, T. A. and Fortin, N. J. (2013). The evolution of episodic memory. *Proceedings of the National Academy of Sciences*, 110(Supplement 2):10379–10386.

Amsterdam, B. (1972). Mirror self-image reactions before age two. *Developmental psychobiology*, 5(4):297–305.

Anderson, J. R. and Gallup Jr, G. G. (1999). Self-recognition in nonhuman primates: past and future challenges.

Anscombe, G. (1975). The first person. In *Mind and Language*, pages 45–65. Oxford: Clarendon Press.

Apperly, I. A. and Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4):953.

Baillargeon, R., Scott, R. M., and Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology*, 67:159–186.

Baillargeon, R., Scott, R. M., and He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, 14(3):110–118.

Barth, J., Povinelli, D., and Cant, J. (2004). Bodily origins of self. *The self and memory*, pages 11–43.

Barth, J., Reaux, J. E., and Povinelli, D. J. (2005). Chimpanzees'(pan troglodytes) use of gaze cues in object-choice tasks: different methods yield different results. *Animal cognition*, 8(2):84–92.

Bauer, P. J. (2007). *Remembering the times of our lives: Memory in infancy and beyond*. Psychology Press.

Bauer, P. J., Larkina, M., and Deocampo, J. (2010). Early memory development. In Goswami, U., editor, *The Wiley-Blackwell handbook of childhood cognitive development*, volume 2, pages 153–179. Wiley-Blackwell, Oxford, UK.

Bauer, P. J., Wenner, J. A., Dropik, P. L., Wewerka, S. S., and Howe, M. L. (2000). Parameters of remembering and forgetting in the transition from infancy to early childhood. *Monographs of the Society for Research in Child Development*, pages i–213.

Birch, S. A. (2005). When knowledge is a curse: Children's and adults' reasoning about mental states. *Current Directions in Psychological Science*, 14(1):25–29.

Bloom, P. (2000). *How children learn the meanings of words.* The MIT Press.

Breen, N., Caine, D., Coltheart, M., et al. (2001). Mirrored-self misidentification: Two cases of focal onset dementia. *Neurocase*, 7(3):239–254.

Breen, N., Caine, D., Coltheart, M., Hendy, J., and Roberts, C. (2000). Towards an understanding of delusions of misidentification: Four case studies. *Mind & Language*, 15(1):74–110.

Burgess, N. (2006). Spatial memory: how egocentric and allocentric combine. *Trends in cognitive sciences*, 10(12):551–557.

Callaghan, T. and Corbit, J. (2014). The development of symbolic representation. In *Handbook of Child Psychology and Developmental Science*, pages 451–535. New York: Wiley.

Campbell, J. (1995). *Past, space, and self.* MIT Press.

Cappelen, H. and Dever, J. (2013). *The inessential indexical: On the philosophical insignificance of perspective and the first person.* OUP Oxford.

Carey, S. (2009). *The origin of concepts.* Oxford University Press.

Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, 28(2):141–172.

Castañeda, H.-N. (1966). 'he' : A study in the logic of self-consciousness. *Ratio*, 8(130–157).

Chang, L., Fang, Q., Zhang, S., Poo, M.-m., and Gong, N. (2015). Mirror-induced self-directed behaviors in rhesus monkeys after visual-somatosensory training. *Current Biology*, 25(2):212–217.

Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition*, 23(2):149–178.

Chiandetti, C. and Vallortigara, G. (2008). Is there an innate geometric module? effects of experience with angular geometric cues on spatial re-orientation based on the shape of the environment. *Animal cognition*, 11(1):139–146.

Conway, M. A., Loveday, C., and Cole, S. N. (2016). The remembering–imagining system. *Memory Studies*, 9(3):256–265.

Courage, M. L., Edison, S. C., and Howe, M. L. (2004). Variability in the early development of visual self-recognition. *Infant Behavior and Development*, 27(4):509–532.

DeLoache, J. S. (1987). Rapid change in the symbolic functioning of very young children. *Science*, 238(4833):1556–1557.

DeLoache, J. S. (2000). Dual representation and young children's use of scale models. *Child development*, 71(2):329–338.

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends in cognitive sciences*, 8(2):66–70.

DeLoache, J. S. (2011). Early development of the understanding and use of symbolic artifacts. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, pages 312–336. Wiley-Blackwell Chichester, UK.

DeLoache, J. S. and Burns, N. M. (1994). Early understanding of the representational function of pictures. *Cognition*, 52(2):83–110.

Devitt, M. (2013). The myth of the problematic 'de se'. In Capone, A. and Feit, N., editors, *Attitudes De Se: Linguistics, epistemology, metaphysics*, pages 133–162. CSLI Publications, Stanford.

Doris, J. M. (2015). *Talking to our selves: Reflection, ignorance, and agency.* Oxford: Oxford University Press.

Dunphy-Lelii, S. and Wellman, H. M. (2012). Delayed self-recognition in autism: A unique difficulty? *Research in autism spectrum disorders*, 6(1):212–223.

Egan, A. (2007). Epistemic modals, relativism and assertion. *Philosophical Studies*, 133(1):1–22.

Epley, N., Keysar, B., Van Boven, L., and Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology*, 87(3):327.

Evans, G. (1982). *The Varieties of Reference.* Oxford University Press.

Fivush, R. (2011). The development of autobiographical memory. *Annual review of psychology*, 62:559–582.

Fivush, R. and Nelson, K. (2006). Parent–child reminiscing locates the self in the past. *British Journal of Developmental Psychology*, 24(1):235–251.

Friedman, W. J. (2004). Time in autobiographical memory. *Social Cognition*, 22(5: Special issue):591–605.

Gallup, G. G. (1970). Chimpanzees: self-recognition. *Science*, 167(3914):86–87.

Gallup, G. G. (1998). Self-awareness and the evolution of social intelligence. *Behavioural Processes*, 42(2):239–247.

Gallup, G. G., Anderson, J. R., and Shillito, D. J. (2002). The mirror test. *The cognitive animal: Empirical and theoretical perspectives on animal cognition*, pages 325–333.

Gilovich, T., Savitsky, K., and Medvec, V. H. (1998). The illusion of transparency: biased assessments of others' ability to read one's emotional states. *Journal of personality and social psychology*, 75(2):332.

Goddard, L., Dritschel, B., Robinson, S., and Howlin, P. (2014). Development of autobiographical memory in children with autism spectrum disorders: Deficits, gains, and predictors of performance. *Development and Psychopathology*, 26(1):215–228.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading.* Oxford University Press.

Griffin, D. W. and Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. In *Advances in experimental social psychology*, volume 24, pages 319–359. Elsevier.

Grush, R. (2000). Self, world and space: The meaning and mechanismsof ego-and allocentric spatial representation. *Brain and Mind*, 1(1):59–92.

Haden, C. A., Ornstein, P. A., Eckerman, C. O., and Didow, S. M. (2001). Mother–child conversational interactions as events unfold: Linkages to subsequent remembering. *Child Development*, 72(4):1016–1031.

Hall, D. G. (1999). Semantics and the acquisition of proper names. In *Language, logic, and concepts: Essays in memory of John Macnamara*, pages 337–372. The MIT Press.

Harley, K. and Reese, E. (1999). Origins of autobiographical memory. *Developmental Psychology*, 35(5):1338.

Hermer, L. and Spelke, E. S. (1994). A geometric process for spatial reorientation in young children. *Nature*, 370(6484):57.

Higginbotham, J. (2003). Remembering, imagining, and the first person. In Barber, A., editor, *Epistemology of language*, pages 496–533. Oxford: Oxford University Press.

Hobson, R. P. (1995). *Autism and the development of mind.* Psychology Press.

Holton, R. (2015). Primitive self-ascription: Lewis on de se. In Lower, B. and Schaffer, J., editors, *A Companion to David Lewis*, pages 399–410. John Wiley & Sons.

Hudson, J. A. (1990). The emergence of autobiographical memory in mother-child conversation. In Hudson, R. F. . J. A., editor, *Knowing and remembering in young children*, pages 166–196. Cambridge University Press.

James, W. (1890). *The Principles of Psychology.* New York: Holt.

Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology*, 26(2):165–208.

Khoo, J. (2015). Modal disagreements. *Inquiry*, 58(5):511–534.

Klatzky, R. L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial cognition*, pages 1–17. Springer.

Klein, S. B. and Gangi, C. E. (2010). The multiplicity of self: neuropsychological evidence and its implications for the self as a construct in psychological research. *Annals of the New York Academy of Sciences*, 1191(1):1–15.

Knobe, J. and Yalcin, S. (2014). Epistemic modals and context: Experimental data. *Semantics and Pragmatics*, 7:10–1.

Kratzer, A. (1991). Modality. In von Stechow, A. and Wunderlich, D., editors, *Semantics: An international handbook of contemporary research*, pages 639–650. de Gruyter.

Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives*, volume 36. Oxford University Press.

Lagattuta, K. H. and Wellman, H. M. (2001). Thinking about the past: Early knowledge about links between prior experience, thinking, and emotion. *Child Development*, 72(1):82–102.

Leslie, A. M. (1987). Pretense and representation: The origins of" theory of mind.". *Psychological review*, 94(4):412.

Lewis, D. (1979). Attitudes de dicto and de se. *The Philosophical Review*, 88(4):513–543.

Lewis, M. (2003). The emergence of consciousness and its role in human development. *Annals of the New York Academy of Sciences*, 1001(1):104–133.

Lewis, M. and Brooks-Gunn, J. (1979). *Social Cognition and the Acquisition of the Self.* New York: Plenum Press.

Lind, S. E. and Bowler, D. M. (2009). Delayed self-recognition in children with autism spectrum disorder. *Journal of autism and developmental disorders*, 39(4):643–650.

MacFarlane, J. (2011). Episdemic modals are assessment-sensitive. In Egan, A. and Weatherson, B., editors, *Epistemic Modality*, pages 144–178. Oxford University Press.

MacFarlane, J. (2014). *Assessment sensitivity: Relative truth and its applications.* OUP Oxford.

Macnamara, J. (1982). *Names for things: A study of human learning.* MIT Press.

Magidor, O. (2015). The myth of the de se. *Philosophical Perspectives*, 29(1):249–283.

Markowitsch, H. J. and Staniloiu, A. (2011). Memory, autonoetic consciousness, and the self. *Consciousness and cognition*, 20(1):16–39.

McCormack, T. (2001). Attributing episodic memory to animals and children. *Time and memory: Issues in philosophy and psychology*, pages 285–314.

McCormack, T. (2015). The development of temporal cognition. In Lerner, R. M., editor, *Handbook of child psychology and developmental science*. Wiley.

McCormack, T. and Hoerl, C. (2005). Children's reasoning about the causal significance of the temporal order of events. *Developmental Psychology*, 41(1):54.

McCormack, T. and Hoerl, C. (2007). Young children's reasoning about the order of past events. *Journal of experimental child psychology*, 98(3):168–183.

Millikan, R. G. (2012). Are there mental indexicals and demonstratives? *Philosophical Perspectives*, 26(1):217–234.

Mitchell, R. W. (1993). Mental models of mirror-self-recognition: Two theories. *New ideas in Psychology*, 11(3):295–325.

Mitchell, R. W. (1997a). A comparison of the self-awareness and kinesthetic–visual matching theories of self-recognition: Autistic children and others. *Annals of the New York Academy of Sciences*, 818(1):39–62.

Mitchell, R. W. (1997b). Kinesthetic-visual matching and the self-concept as explanations of mirror-self-recognition. *Journal for the theory of social behaviour*, 27(1):17–39.

Moore, C. (2006). *The development of commonsense psychology*. Psychology Press.

Morin, A. (2011). Self-recognition, theory-of-mind, and self-awareness: What side are you on? *Laterality*, 16(3):367–383.

Mullen, M. K. and Yi, S. (1995). The cultural context of talk about the past: Implications for the development of autobiographical memory. *Cognitive Development*, 10(3):407–419.

Neisser, U. (1988). Five kinds of self-knowledge. *Philosophical psychology*, 1(1):35–59.

Nelson, K. (1993). The psychological and social origins of autobiographical memory. *Psychological science*, 4(1):7–14.

Nelson, K. (2009). *Young minds in social worlds: Experience, meaning, and memory.* Harvard University Press.

Nelson, K. and Fivush, R. (2004). The emergence of autobiographical memory: a social cultural developmental theory. *Psychological review*, 111(2):486–511.

Newton, E. (1990). Overconfidence in the communication of intent. Unpublished doctoral dissertation.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological bulletin*, 125(6):737.

Ninan, D. (2016). What is the problem of de se attitudes? In *About Oneself: De Se Thought and Communication*, pages 86–120. Oxford University Press.

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *science*, 308(5719):255–258.

Oshima-Takane, Y., TAKANE, Y., and Shultz, T. R. (1999). The learning of first and second person pronouns in english: network models and analysis. *Journal of Child Language*, 26(3):545–575.

Parker, S. T., Mitchell, R. W., and Boccia, M. L. (2006). *Self-awareness in animals and humans: Developmental perspectives*. Cambridge University Press.

Perner, J. (1991). *Understanding the representational mind*. The MIT Press.

Perner, J., Kloo, D., and Gornik, E. (2007). Episodic memory development: theory of mind is part of re-experiencing experienced events. *Infant and Child Development*, 16(5):471–490.

Perner, J., Mauer, M. C., and Hildenbrand, M. (2011). Identity: Key to children's understanding of belief. *Science*, 333(6041):474–477.

Perner, J. and Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in cognitive sciences*, 16(10):519–525.

Perner, J. and Ruffman, T. (1995). Episodic memory and autonoetic conciousness: developmental evidence and a theory of childhood amnesia. *Journal of experimental child psychology*, 59(3):516–548.

Perry, J. (1977). Frege on demonstratives. *The philosophical review*, 86(4):474–497.

Perry, J. (1979). The problem of the essential indexical. *Noûs*, pages 3–21.

Perry, J. (1998). Myself and i. *Philosophie in synthetischer Absicht, Stuttgart*, pages 83–103.

Perry, J. (2006). Stalnaker and indexical belief. In *Content and modality: themes from the philosophy of Robert Stalnaker*, pages 204–221. Oxford: Clarendon Press.

Phillips, J. and Knobe, J. (2018). The psychological representation of modality. *Mind and Language*, pages 65–94.

Phillips, M. L. (1996). " mirror, mirror on the wall, who...?": Towards a model of visual self-recognition. *Cognitive Neuropsychiatry*, 1(2):153–164.

Plotnik, J. M., De Waal, F. B., and Reiss, D. (2006). Self-recognition in an asian elephant. *Proceedings of the National Academy of Sciences*, 103(45):17053–17057.

Posada, S. and Colell, M. (2007). Another gorilla (gorilla gorilla gorilla) recognizes himself in a mirror. *American Journal of Primatology*, 69(5):576–583.

Povinelli, D. J. (1995). The unduplicated self. In Rochat, P., editor, *The Self in Early Infancy*, pages 161–192. North-Holland/Elsevier Science Publishers.

Povinelli, D. J. (2001). The self: Elevated in consciousness and extended in time. In Moore, C. and Lemmon, K., editors, *The Self in Time: Developmental Perspectives*. Lawrence Erlbaum Associates Publishers.

Povinelli, D. J., Landau, K. R., and Perilloux, H. K. (1996). Self-recognition in young children using delayed versus live feedback: Evidence of a developmental asynchrony. *Child development*, 67(4):1540–1554.

Povinelli, D. J., Landry, A. M., Theall, L. A., Clark, B. R., and Castille, C. M. (1999). Development of young children's understanding that the recent past is causally bound to the present. *Developmental psychology*, 35(6):1426.

Povinelli, D. J. and Simon, B. B. (1998). Young children's understanding of briefly versus extremely delayed images of the self: emergence of the autobiographical stance. *Developmental Psychology*, 34(1):188.

Ramachandran, V., Altschuler, E., and Hillyer, S. (1997). Mirror agnosia. *Proceedings of the Royal Society of London B: Biological Sciences*, 264(1382):645–647.

Ramachandran, V. S. (2007). The neurology of self-awareness. *The Edge (online journal)*.

Recanati, F. (2007). *Perspectival Thought: A plea for (moderate) relativism*. Clarendon Press.

Recanati, F. (2012). *Mental files*. Oxford University Press.

Reese, E., Haden, C. A., and Fivush, R. (1993). Mother-child conversations about the past: Relationships of style and memory over time. *Cognitive development*, 8(4):403–430.

Reiss, D. and Marino, L. (2001). Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence. *Proceedings of the National Academy of Sciences*, 98(10):5937–5942.

Rochat, P. (2009). *Others in mind: Social origins of self-consciousness*. Cambridge University Press.

Rochat, P., Goubet, N., and Senders, S. J. (1999). To reach or not to reach? perception of body effectivities by young infants. *Infant and Child Development*, 8(3):129–148.

Roma, P. G., Silberberg, A., Huntsberry, M. E., Christensen, C. J., Ruggiero, A. M., and Suomi, S. J. (2007). Mark tests for mirror self-recognition in capuchin monkeys (cebus apella) trained to touch marks. *American Journal of Primatology*, 69(9):989–1000.

Royzman, E. B., Cassidy, K. W., and Baron, J. (2003). " i know, you know": Epistemic egocentrism in children and adults. *Review of General Psychology*, 7(1):38.

Schacter, D. L., Addis, D. R., and Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, 8(9):657.

Schellenberg, S. (2007). Action and self-location in perception. *Mind*, 116(463):603–632.

Schmitt, K. L. and Anderson, D. R. (2002). Television and reality: Toddlers' use of visual information from video to guide behavior. *Media Psychology*, 4(1):51–76.

Southgate, V., Chevallier, C., and Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental science*, 13(6):907–912.

Spelke, E., Lee, S. A., and Izard, V. (2010). Beyond core knowledge: Natural geometry. *Cognitive science*, 34(5):863–884.

Spelke, E. S. (2017). Core knowledge, language, and number. *Language Learning and Development*, 13(2):147–170.

Spencer, C. (2007). Is there a problem of the essential indexical? In O'Rourke, M. and Washington, C., editors, *Situating Semantics: Essays on the Philosophy of John Perry*, pages 179–198. Cambridge, MA: MIT Press.

Stalnaker, R. (1978). Assertion. In Cole, P., editor, *Syntax and Semantics*, volume 9, pages 315–332. New York: Academic.

Stalnaker, R. (2014). *Context*. OUP Oxford.

Stalnaker, R. C. (1981). Indexical belief. *Synthese*, 49(1):129–151.

Stephenson, T. (2007). Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy*, 30(4):487–525.

Suddendorf, T. and Butler, D. L. (2013). The nature of visual self-recognition. *Trends in Cognitive Sciences*, 17(3):121–127.

Suddendorf, T. and Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences*, 30(3):299–313.

Suddendorf, T. and Whiten, A. (2001). Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals. *Psychological bulletin*, 127(5):629.

Taylor, M., Esbensen, B. M., and Bennett, R. T. (1994). Children's understanding of knowledge acquisition: The tendency for children to report that they have always known what they have just learned. *Child development*, 65(6):1581–1604.

Tiffany, E. C. (2000). What is essential about indexicals? *Philosophical studies*, 100(1):35–50.

Troseth, G. L., Saylor, M. M., and Archer, A. H. (2006). Young children's use of video as a source of socially relevant information. *Child development*, 77(3):786–799.

Tulving, E. (1972). Episodic and semantic memory. In Tulving, E. and Donaldson, W., editors, *Organization of memory*, pages 381–403. New York. Academic Press.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual review of psychology*, 53(1):1–25.

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? *The missing link in cognition: Origins of self-reflective consciousness*, pages 3–56.

Tustin, K. and Hayne, H. (2010). Defining the boundary: age-related changes in childhood amnesia. *Developmental psychology*, 46(5):1049.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

Vallortigara, G. (2009). Animals as natural geometers. In Tommasi, L., Peterson, M. A., and Nadel, L., editors, *Cognitive biology: evolutionary and developmental perspectives on mind, brain and behaviour (eds Tommasi L., Nadel L., Peterson M.)*, pages 83–104. MIT Press.

Vargas, J. P., López, J. C., Salas, C., and Thinus-Blanc, C. (2004). Encoding of geometric and featural spatial information by goldfish (carassius auratus). *Journal of Comparative Psychology*, 118(2):206.

Von Fintel, K. and Gillies, A. S. (2008). Cia leaks. *Philosophical review*, 117(1):77–98.

Von Fintel, K. and Gillies, A. S. (2011). Might made right. In Egan, A. and Weatherson, B., editors, *Epistemic modality*, pages 108–130. Citeseer.

Wang, Q. (2001). Culture effects on adults' earliest childhood recollection and self-description: implications for the relation between memory and the self. *Journal of personality and social psychology*, 81(2):220.

Wang, Q. and Fivush, R. (2005). Mother–child conversations of emotionally salient events: exploring the functions of emotional reminiscing in european-american and chinese families. *Social Development*, 14(3):473–495.

Wang, R. F. and Spelke, E. S. (2002). Human spatial representation: Insights from animals. *Trends in cognitive sciences*, 6(9):376–382.

Weigle, T. W. and Bauer, P. J. (2000). Deaf and hearing adults' recollections of childhood and beyond. *Memory*, 8(5):293–309.

Welch-Ross, M. (2001). Personalizing the temporally extended self: Evaluative self-awareness and the development of autobiographical memory. In *Moore, Chris and Lemmon, Karen and Skene, Karen*, pages 97–120. Psychology Press.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Yonas, A. and Granrud, C. A. (1985). Reaching as a measure of infants' spatial perception. In Gottlieb, G. K. N. A., editor, *Measurement of audition and vision in the first year of postnatal life: A methodological overview*, pages 301–322. Ablex Publishing Corp.