How Opacity Spawns Transparency
A Theory of Self-knowledge of Beliefs

Peters, Uwe

Awarding institution:
King's College London

How Opacity Spawns Transparency – A Theory of Self-knowledge of Beliefs

Uwe Peters
King's College London

PhD Thesis in Philosophy
September 2015

Primary supervisor
Professor Nicholas Shea

Secondary supervisor
Professor David Papineau

Dedicated to Charlotte and Maja

CONTENTS

ABSTRACT

This thesis develops a theory of self-knowledge of beliefs that suggests that when the same resources employed to identify other people's beliefs are applied to oneself then non-interpretive and non-inferential self-knowledge of beliefs results. Like many other accounts, the proposal is based on the *transparency* of beliefs, i.e. the phenomenon that we can determine whether we believe *p* by determining whether *p*. However, it is superior to extant accounts of this kind because it makes the transparency of beliefs intelligible without presupposing any representation of one's own mind. The account does so by proposing that the transparency of beliefs relies on a combination of practical and theoretical reasoning that, as soon as it has been executed once, remains subsequently implicit or unarticulated in the transition from the result of one's determining whether *p* to one's self-ascription of a belief about *p*. The transition at issue then leads to non-inferential self-knowledge. I argue that this account of self-knowledge, the *implicit dual-reasoning* (IDR) theory, is not threatened by the psychological data often taken to undermine the existence of non-interpretive and non-inferential self-knowledge of attitudes and can be integrated with a well-supported interpretation-based account of self-knowledge of attitudes in general. On the resulting hybrid view, self-knowledge of attitudes is typically interpretive but beliefs retrievable in conscious thinking remain knowable non-inferentially. Toward the end, I supplement this view with an account of the function of self-knowledge of beliefs. The overall theory of self-knowledge that emerges from the thesis supports a hitherto unexplored picture of the relation between the nature of self-knowledge of beliefs and the nature of other-knowledge of them. It suggests that the existence of transparency-based, non-inferential self-knowledge of beliefs is grounded in the *prima facie* entirely unrelated fact that other people's beliefs are opaque, i.e. only interpretively accessible, to us.

ACKNOWLEDGMENTS

I would like to thank Nick Shea for his excellent supervision, patience, and support throughout. Many thanks also to David Papineau for his supervision in the earlier stages of the research. Finally, I would like to express my gratitude to the Department of Philosophy at King's for the AHRC PhD scholarship.

CHAPTER 1

Introduction

In this chapter, I introduce the topic of the thesis and the three main questions that the thesis aims to answer. I mention the problems that need to be solved to answer the three questions, and say how and where these problems will be dealt with in the different chapters of the thesis. I also clarify some of the key terms that will be used.

1. The topic

This thesis is about self-knowledge of beliefs. It is often thought that self-knowledge of beliefs has two features that are particularly explanatorily problematic and philosophically interesting (e.g., Byrne 2005, 2011; Boghossian 2008; Fernández 2013).

The first feature is that one's judgments about one's own beliefs seem to be more strongly justified and more likely to amount to knowledge than other people's judgments about them. For instance, suppose you and I disagree on whether I believe *p*. Suppose also that you hold that I believe *p* but I think that I don't believe *p*. Typically, when such disagreement occurs then the matter will be resolved by deferring to *my* view on the issue. The assumption is that I'm in a better position to know what my beliefs are than anyone else. This indicates that I'm more strongly justified in my self-ascriptions of beliefs than others are in their ascriptions of beliefs to me. Self-knowledge of beliefs that has this feature is what I shall call *authoritative* in nature.

There is a second aspect of self-knowledge of beliefs that poses an explanatory problem. For instance, when I want to find out about your beliefs, I need to observe, interpret, or draw inferences about you to determine your beliefs. In contrast, typically, I seem to be able to know my own beliefs without any kind of observation, interpretation, inference or evidence pertaining to my having the beliefs. Self-knowledge of beliefs that has this feature is what I shall call *immediate* in nature.

Authoritative and immediate self-knowledge of beliefs puts one in a privileged position with respect to one's own beliefs compared to anyone else. It is what I shall call

privileged self-knowledge of beliefs. Privileged self-knowledge of beliefs is the main topic of this thesis.

2. The questions

Three questions come to mind when thinking about privileged self-knowledge of beliefs: whether there is such a thing, if there is, how it can be explained, and why it exists at all. I shall refer to these as the *whether*-, the *how*-, and the *why*-questions. What motivates asking them?

Consider the *whether*-question. It is widely accepted among philosophers that we do at least sometimes have privileged self-knowledge of beliefs (e.g., Searle 1983; Davidson 1994; Shoemaker 1996, 2012; Moran 2001, 2012; Nichols and Stich 2003; Bilgrami 2006; Goldman 2006; Boghossian 2008). In recent years, however, a number of theorists have argued that this is mistaken. They maintain that we in fact come to know our own attitudes only via a swift and normally unconscious process of interpretation or, more generally, via an inference from evidence (e.g., Dennett 1992; Gopnik 1993; Gazzaniga 1995; 2000; Stephen and Graham 2000; Wilson 2002; Dretske 2003; Cooper 2007; Lycan 2008; Lawlor 2009; Williams and Happé 2010; Carruthers 2009a, 2011; Mandelbaum 2014; Cassam 2014, 2015). The opposition of these two views leads naturally to the question of whether or not we do have privileged self-knowledge of beliefs. In this thesis, I will argue for an affirmative answer. I will do so via developing a particular answer to the *how*-question.

Suppose we take the intuition that there is privileged self-knowledge of beliefs at face value. How can this kind of knowledge be explained? Consider, for instance, the immediacy of privileged self-knowledge of beliefs. It is often noted that this feature of self-knowledge is

> puzzling because knowledge of our own thoughts and beliefs is surely knowledge of contingent facts, and the usual presumption is that knowledge of contingent matters must be based on observation, evidence or inference. So if it is just a contingent fact about me that I believe that *p*, and yet I know without observation, inference or evidence that I have this belief, then it needs to be explained how this is possible. (Cassam 2011: 1-2)

One approach to explaining the immediacy of self-knowledge of beliefs appeals to the

transparency of beliefs, i.e. the apparent fact that one's own belief *p* can be self-known simply by determining whether *p* (Moran 2001, 2012; Byrne 2005; Boyle 2011). As Evans (1982) puts it in an often-cited passage:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward – upon the world. If someone asks me 'Do you think there is going to be a third world war?' I must attend, in answering him, to precisely the same outward phenomena, as I would attend to if I were answering the question 'Will there be a third world war?' I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*. (225)[1]

The claim is that if I want to find out whether I believe *p*, then I don't need to turn 'inward', i.e. I don't need to attend to and represent some aspect of my own mind. I just need to attend to the world and those features of it that allow me to settle whether *p*. If upon reflection on them or upon swiftly recalling that *p* I conclude that *p* then I may on the basis of the outcome of my determining whether *p* judge that I believe *p*. I shall call this procedure, i.e. one's settling whether one believes *p* by determining whether *p*, the *transparency method* (TM).

TM is often taken to help explain the immediacy of self-knowledge of beliefs. The underlying thought is the following. To determine whether *p*, I might need to observe states of affairs, draw inferences, and, in general, rely on evidence that *p*. But, by assumption, the observations, inferences or evidence that I then rely on only support or speak against *p,* which is a state of affairs of the non-mental, external world. They don't speak to the matter as to whether I *believe* that *p*. Since that is so, if, as TM suggests, I can work out whether I believe *p* by determining whether *p* alone then I can find out about my belief without observation, inference, or evidence because whatever observation, inference or evidence might be involved in determining whether *p*, it does not pertain to mental affairs. And when it is claimed that the acquisition of self-knowledge of a belief *p* does not involve any observation, inference or evidence, then the claim is only that acquiring such knowledge doesn't involve any kind of observation, inference or evidence *that one believes* that *p*, i.e. that supports or speaks

[1] Evans' claim that to find out whether one believes *p*, one *must* attend to the phenomena that one attends to in order to answer the question of whether *p* is perhaps too strong. Sometimes we find out about our beliefs by self-observation or being told about them by others, e.g., in therapeutic contexts.

against one's believing that *p*. So if I can come to know my own belief *p* simply by determining whether *p* alone, then I can come to know it immediately. That I can know my own belief *p* immediately can then be explained by holding that this is because for me the question as to whether I believe *p* is answered in the same way as the question as to whether *p*.

However, this line of thought can clearly only be a first step toward an explanation of the immediacy and privileged nature of self-knowledge of beliefs. For it swiftly leads to the next question of how exactly I can possibly find out about whether I believe *p* via TM, i.e. simply by determining whether *p* alone.

Before considering this issue further, it is worth noting that the viability of TM as a method for coming to know one's own beliefs deserves to be explored. This is because TM has the following three attractive features.

(1) Suppose I want to find out whether I believe the dodo is extinct. To answer the question, it doesn't help me to sit back and wait until the answer pops up in my mind. Rather, it seems that I need to *do* something. Finding an answer to the question takes some effort on my part (Fernández 2013). My coming to know whether I believe the dodo is extinct seems to be a *cognitive achievement*.[2] It isn't something that just happens to me. TM does justice to this point. For it suggests that in order to determine whether I believe that the dodo is extinct, in the context of TM, I need to move from asking 'Do I believe that the dodo is extinct?' to asking myself 'Is the dodo extinct?' Further, even if I know about dodos and immediately recall that they are indeed extinct, the answer to the question doesn't then just suddenly appear in my mind. I rather need to attend to the matter and think about it. All this requires cognitive effort. TM has thus the advantage of allowing us to explain why the acquisition of self-knowledge of beliefs is something a subject does or achieves.

(2) Furthermore, TM appears to capture something real. For when I'm asked whether I believe *p*, it seems that I do then, just as TM suggests, start determining

[2] Many philosophers have made this point and used the term 'cognitive achievement' in their work on self-knowledge, see, e.g., Boghossian (2008: 154), Fernández (2013: 28), and Cassam (2014: 136). I borrow the term from them.

whether *p* to answer the question. Of course, it remains to be seen whether this kind of procedure is *de facto* a way of forming second-order beliefs and acquiring self-knowledge of beliefs. Still, the intuitive plausibility of the view that it *is* offers a good ground for exploring TM further.

(3) Finally, *prima facie* TM suggests that the privileged nature of self-knowledge of beliefs can be explained without postulating any mysterious 'inner sense' or other special mechanism for awareness of one's own mind in addition to general first-order reasoning capacities, the belief-concept, and the self-concept. It thus promises to allow for the formulation of an attractively explanatorily and ontologically economical account of self-knowledge of beliefs.

Points (1)-(3) motivate me in this thesis to investigate whether a viable TM-based account of privileged self-knowledge of beliefs can be developed.

Note that what I shall be concerned with is whether TM is a method that helps explain privileged self-knowledge of beliefs by telling us how such knowledge is *acquired*. It isn't obvious that privileged self-knowledge of beliefs should be explained by appealing to how it is acquired, or indeed that TM is a method for acquiring self-knowledge. For instance, one might hold that privileged self-knowledge of beliefs should be accounted for by appealing to the metaphysics of beliefs: some philosophers writing about TM claim that there is a constitutive link between first-order beliefs and second-order beliefs about them, and argue that TM is merely a means for bringing the second-order beliefs into consciousness rather than forming them in the first place (e.g., Boyle 2011; Shoemaker 2012).

Below I shall say why I think this proposal is unsatisfactory. For now I just note that in this thesis, I shan't make any assumptions about a constitutive connection between beliefs. Instead I want to try to find an answer to the question of how to explain privileged self-knowledge of beliefs by searching for an answer to the question of how it is acquired. My strategy will be to take TM as a starting point in the theorising on the matter.

3. The problems and claims

If TM is to explain privileged self-knowledge of beliefs by telling us how we acquire

such knowledge, and, more generally, if a TM-based account of privileged self-knowledge of beliefs is to provide answers not only to the *how*-question but also to the *whether*-, and the *why*-questions, then a number of issues need to be addressed. The following four problems come to mind.

The first problem

A TM-based account of privileged self-knowledge of beliefs needs to resolve two closely related issues pertaining to first-order thinking.

(1) TM suggests that we can find out whether we believe *p* via reflection on and a judgment about whether *p*. Reflecting on and forming a judgment about whether *p* is a matter of conscious first-order thinking.[3] As it happens, it has recently been argued that in conscious first-order thinking we have only *indirect* access to our own judgments via intermediaries (e.g., visual imagery or inner-speech tokens) that express them and first need to be interpreted and their underlying attitudes self-ascribed in order to acquire attitude-like roles (e.g., Frankish 2009, 2012; Carruthers 2014, 2015). If this view is right then the conscious first-order thinking involved in TM will similarly already require interpretative processing and self-ascriptions of attitudes. That is, it will already require turning 'inward' and representing aspects of one's own mind. In my view, this is at odds with TM. For, as noted, the method suggests that in order to determine whether one believes *p*, one only needs to attend to and represent worldly states of affairs pertaining to whether *p*. It might be suggested that TM only pertains to *conscious* thinking and is compatible with there being unconscious representations of aspects of one's own mind when one is in the context of the method settling whether *p*. While this is an option, I take it that it isn't what advocates of TM have in mind. In any case, I here assume that TM is a method that allows one to come to know one's own belief *p* without, *consciously* or *un*consciously, turning inward and representing aspects of one's own mind prior to the formation of the self-ascription of the belief *p*. With this in mind, for a TM-based account of privileged self-knowledge of beliefs to be viable, it needs to be shown that the view that we only have indirect access to our own judgments in conscious first-order thinking is false and that we have direct access to them.

[3] I will specify what I mean by 'conscious thinking' in the next section. The same holds for the terms 'belief', 'judgment' and 'self-ascription'.

(2) Furthermore and relatedly, TM suggests that we can find out whether we believe *p* on the basis of a judgment *p*. However, a proposition *p* might be judged, supposed, doubted etc. Since that is so, it needs to be explained how we can tell apart these different attitudes when they figure in our conscious thinking and how we can detect specifically the *judgment* that *p*. For in applications of TM, we are thought to form a self-ascription of a belief *p* on the basis of the judgment *p* rather than the supposition or doubt *p*. If a tenable TM-based account of privileged self-knowledge of beliefs is to be developed, then it needs to be shown how, in the context of TM, the judgment *p* can be detected without already being represented.

Since one central aim of the thesis is to try to develop a TM-based account of privileged self-knowledge of beliefs, the issues captured in (1) and (2) need to be resolved. This will happen in Chapter 2, in the first step of the overall argument in the thesis. I show that the recent case for the view that we only have indirect access to judgments in conscious thinking is unconvincing and doesn't threaten the alternative, i.e. the direct access proposal. I also offer positive support for the view that we have direct access to our own judgments in conscious first-order thinking and that this access involves the operation of a sub-personal mechanism that detects and differentiates the attitude types of representations without representing the attitudes themselves. This then resolves the first problem that stands in the way of a TM-based explanation of privileged self-knowledge of beliefs.

The second problem

As noted, TM suggests that we can find out whether we believe *p* by determining whether *p*. However, determining whether, say, a shrimp's heart is in its head pertains only to shrimps. It has *prima facie* little to do with anyone believing anything: a shrimp has its heart in its head even if no one believes it, and one could believe that a shrimp's heart is in its head even if it weren't. Given this, how can, as TM suggests, determining whether a shrimp's heart is in its head provide one with an insight into one's *belief* about whether a shrimp's heart is in its head? Why "should the evidence that the subject has about how the world is have any bearing on what beliefs a particular person has?" (Martin 1998: 110) Finding answers to these questions is the second problem that stands in the way of providing a satisfactory TM-based account of self-knowledge of beliefs.

The problem has two aspects. First, given that *p* has little to do with whether or not one

believes *p*, why even begin to determine whether *p* in order to find out whether one believes *p*? The issue here concerns the *intelligibility* of TM from the point of view of the user of the method. Second, since *p* might be the case even if no one believes *p*, how can TM lead to a *justified* judgment or belief about one's own belief and so possibly to self-knowledge of the belief? The question pertains to the *knowledge*-status of the self-ascription that is thought to result from an application of TM.

Different theories have been proposed to deal with this twofold problem. I shall refer to these accounts as *transparency theories*. In Chapter 3, in the second step of the overall argument in this thesis, I aim to show that a wide range of extant transparency theories doesn't adequately resolve the twofold problem.

In Chapter 4, I then develop a new kind of transparency theory that avoids this and other shortcomings of the existing accounts. The proposal is the heart of the thesis. I call it the *implicit dual-reasoning* (IDR) theory of self-knowledge of beliefs, for it solves the problems of the intelligibility of TM and the knowledge-status of the output of the method by proposing that TM is grounded in a combination of a particular episode of practical and theoretical reasoning. Furthermore, according to the IDR account, most components of this dual reasoning remain, when it has been executed once, in subsequent applications of TM implicit or unarticulated in the transition from the result of one's determining whether *p* to one's self-ascription of a belief about *p*, and the transition at issue then leads to non-inferential, privileged self-knowledge of beliefs. The IDR theory hence also answers the *how*-question.

The third problem

There are, however, accounts of self-knowledge of beliefs that don't invoke TM, call them *TM-independent* accounts. They might offer explanations of self-knowledge of beliefs that are preferable to the one that the IDR theory proposes. In addition, a number of theorists hold that empirical findings, e.g., on self-interpretation and confabulation, indicate that we lack privileged self-knowledge of attitudes including beliefs. These two issues form the third problem that needs to be addressed if a TM-based account such as the IDR theory is to be a credible view of self-knowledge of beliefs that also answers the *whether*-question.

I will deal with them in Chapter 5, in the third step of the overall argument in the thesis. I examine the two most common empirically supported TM-independent accounts of

self-knowledge of attitudes, what I call the *inner-scanner theory* – which claims that we know our own attitudes non-inferentially via a simple detection mechanism (e.g., Nichols and Stich 2003; Goldman 2006) – and the *interpretive sensory-access theory* – which holds that we come to know our own attitudes by employing the same interpretive faculty as the one we use to work out other people's mental states, namely the mindreading system (Carruthers 2009a, 2011). I argue that the IDR theory is preferable to these two proposals when it comes to self-knowledge of beliefs that we can retrieve in conscious thinking.

The argument against the interpretive sensory-access theory and for the IDR account will bring out that the empirical data that is often taken to threaten the existence of privileged self-knowledge of attitudes doesn't in fact undermine the existence of privileged self-knowledge acquired in the way the IDR theory proposes. Indeed, I contend that once the way in which the IDR account explains the acquisition of self-knowledge of beliefs is combined with psychological evidence on the processing constraints of judgment-forming systems, an argument that supports the existence of privileged self-knowledge of beliefs results. That is, the IDR account helps to support an affirmative answer to the *whether*-question via a particular answer to the *how*-question. This resolves the third problem for the project of providing a TM-based account of privileged self-knowledge of beliefs.

As it stands, however, the IDR theory is only a belief-specific view of self-knowledge. Since that is so, I then consider, also in Chapter 5, whether it can be integrated into another, more general account of attitudes. I argue for a positive answer and propose a hybrid view that combines the IDR theory with a revised version of the interpretive sensory-access account. According to the hybrid view, the acquisition of self-knowledge of attitudes requires turning onto oneself the interpretive mechanism that evolved for the acquisition of other-knowledge of attitudes, but although the mechanism typically operates interpretively even when applied to oneself, in applications of TM, it also produces privileged self-knowledge of beliefs.

The fourth problem

One question that the hybrid view of self-knowledge of attitudes leaves open is why the ability to ascribe attitudes to other people was, as the view suggests, turned inward onto oneself to begin with so as to give rise to self-knowledge of attitudes. More specifically,

what remains unexplained is why there is such a thing as privileged self-knowledge of beliefs at all. A full account of self-knowledge of beliefs shouldn't leave it a mystery as to why there is such a thing to begin with. The provision of an explanation is what I take to be the fourth problem that a satisfactory TM-based account of self-knowledge of beliefs should solve.

I will deal with it in Chapter 6, in the fourth and final step of the overall argument in this thesis. I take the question as to why there is such a thing as privileged self-knowledge of beliefs to be a query about the evolutionary function of such knowledge. An answer to the question then involves saying what adaptive or reproductive advantage privileged self-knowledge of beliefs confers onto subjects who have it. After dismissing three common extant answers, I develop my own account. It combines a recent cognitive-scientific proposal on the function of conscious metacognition with the upshots of the preceding discussion about TM. On the resulting view, self-knowledge of beliefs, in general, has the function to enable us to report our own beliefs to others. This leads in cooperative environments to improved joint decision-making and better coordination with advantages for both self and others. The function of TM-based privileged self-knowledge of beliefs, in particular, is then to enable the most efficient and reliable verbal communication of our beliefs. This answers the *why*-question and solves the fourth and final problem for a satisfactory TM-based theory of self-knowledge of beliefs.

Taken together, the answers to the *why*-, the *how*-, and the *whether*-questions that will be developed in the thesis will lead to an overall theory of self-knowledge of beliefs that integrates philosophical work on TM with empirical research from across the cognitive sciences. In the last chapter of the thesis, in Chapter 7, I assemble the theory from its four components, which were developed in the different chapters. I sum up the answers that it provides to the three questions and contend that its last component, which pertains to the function of self-knowledge of beliefs, ties all the answers together and yields a final evolutionary argument for the resulting overall account.

The four-component theory of self-knowledge of beliefs that is put in place at the end of the thesis supports a so far unexplored view on the relation between self- and other-knowledge. In the literature on self-knowledge, it is typically assumed that if self-knowledge of beliefs resulted from applying to oneself the same resources via which we

know other people's thoughts then self-knowledge of beliefs would be interpretive and inferential in nature. The view of self-knowledge of belief developed in the thesis suggests that when the same resources that we deploy to determine other people's beliefs are applied to oneself then, somewhat surprisingly, in virtue of the way the human cognitive system is built and operates, non-interpretive and non-inferential self-knowledge results. More generally, the overall theory of self-knowledge of beliefs that emerges from the thesis suggests that the phenomenon that other people's beliefs are opaque,[4] i.e. only interpretively accessible, gives rise to the phenomenon that our own beliefs are non-inferentially self-knowable transparently, i.e. by determining the truth of their contents. That is, the opacity of other people's beliefs engenders the transparency of our own beliefs.

4. Key terms

Before starting the discussion, I will now briefly specify some of the central terms that will be used. These are 'conscious thinking', 'beliefs', 'judgments', and 'self-ascriptions'.

Conscious thinking

According to transparency theories, self-knowledge of a belief *p* is acquired via reflecting on whether *p*. Reflecting on whether *p* is a matter of *conscious thinking*. *Thinking*, broadly construed, is the use of representations for the purpose of reasoning, action guidance, or report. *Conscious* thinking, in particular, could be different things depending on what notion of consciousness is at issue.

There are at least two different notions. Following Block (2002), a representation might be *phenomenally* conscious, which involves it having experiential properties (a 'what-

[4] The term 'opaque' here has little to do with the same term as it is used since Quine (1953) in the philosophy of language. In the philosophy of language, the term 'opaque' refers to linguistic contexts, created by, e.g., 'believe that' constructions, in which the substitution of a singular term referring to x with a different term also referring to x may change the truth-value of the containing sentence. I here adopt the use of the opacity-metaphor from Carruthers (2011) and Cassam (2014).

it's-like-ness'),[5] or it might be *access* conscious.

Access consciousness is typically specified by appeal to the "global workspace" theory (Baars 1988; Dehaene and Naccache 2001) and "working memory" (Baddeley 2006, 2010). Since I will focus on access consciousness only, a brief introduction to the global workspace account and working memory will be useful.

The global workspace

For a representation to be access conscious it needn't necessarily have experiential properties. It only needs to be widely available for reasoning, action guidance, and report. This is typically taken to mean that the representation is in a central or global workspace in the mind, an area across which the different specialist sensory and conceptual systems that are assumed to make up the mind can communicate and broadcast their outputs (Tye 1995; Block 1995; Dennett 2001; Carruthers 2015). Any content that is broadcast in the workspace is automatically accessible to all systems that are connected to the workspace.

To see what is meant by contents being 'in' the workspace, it helps to consider the implementation of the workspace in the brain. At the implementation level, the workspace is taken to be a distributed neural network with long-distance connectivity that can link various specialised brain areas. The exact "contours of the workspace fluctuate as different brain circuits are temporarily mobilised, then demobilised", where, at any given time, only a fraction of the neurons that have the appropriate long-distance and widespread connectivity are activated and constitute the mobilised workspace (Dehaene and Naccache 2001: 13). Given this, for *R* to be 'in' the workspace is for the brain area that encodes *R* to be part of the neural network that presently constitutes the mobilised workspace.

Working memory

When *R* is not only broadcast but also currently used by the systems connected to the global workspace then *R* is thought to figure in working memory, which is a component

[5] This refers to the subject tokening the representation. As Nagel (1974) puts it "an organism has conscious mental states if and only if there is something it's like to be that organism – something it's like for the organism" (436).

of the global workspace.[6] It is a workspace with limited capacity in which information is frequently activated and maintained to support on-going tasks. It is involved when we, for instance, mentally rehearse an action, keep in mind a phone number while searching for a pen, do mental arithmetic, or follow directions we have just been given.

The activation and maintenance of information in working memory for on-going tasks happens via attentional processes that involve the operation of a central executive system that directs and makes use of various sensory systems,[7] and an "episodic buffer", where information from these systems can be integrated with information from semantic and episodic memory (Baddeley 2006, 2010).

Returning with this to conscious thinking, in what follows, I will only be concerned with *access* conscious thinking (henceforth 'conscious thinking'), i.e. thinking that involves the use of representations that are broadcast in the workspace. That is, I will focus on thinking that implicates working memory.

Beliefs

As noted, conscious thinking, e.g., reflecting on whether *p*, is according to transparency theories required for acquiring knowledge of one's own belief *p*. When reflecting on whether *p*, one often draws on one's *beliefs* pertaining to the question of whether *p*.

There are different views on what beliefs are. In what follows, I assume a standard representational account (e.g., Fodor 1975; 1990; Dretske 1988; Millikan 1993; Cummins 1996). That is, I assume that a belief *p* is a mental representation that has the propositional content *p* and plays a particular functional role. More specifically, for S to believe *p* is for her to be disposed to treat *p* as a true premise in reasoning, for action-guidance, and report, where this involves her committing herself to the truth of *p*.

[6] Different theorists understand different things by the term 'working memory' (see Cowan 2008). Some equate the global workspace with working memory, e.g., Block (2007). I here follow Carruthers' (2011: 48f) view.

[7] Baddeley (1986, 2006, 2010) holds that the central system uses only the auditory system ("phonological loop") and the visual system ("visual-spatial scratchpad"), but this view has been criticised by a number of theorists and it is now widely accepted that the central system can make use of all sensory systems (Carruthers 2014).

I assume furthermore that beliefs are standing states. They don't currently figure in thinking but are stored in memory. They can be drawn upon and become activated in conscious thinking for the purpose of reasoning, action guidance, or report. But when this happens then they themselves don't figure in conscious thinking. For conscious thinking is, just like consciousness in general, an occurrence or event. It can come and go. In contrast, one's belief *p* is a state, not an event. It persists over time in that it doesn't, e.g., disappear when one attends to something else, falls asleep, or becomes unconscious. Since consciousness is an occurrence or event but beliefs are states, beliefs can't themselves be conscious, because they can't be both states and events (Crane 2013).

Still, as noted, beliefs can become activated in conscious thinking. When they are activated or drawn upon in one's conscious thinking then what figures in consciousness isn't beliefs but *judgments* with the same content.

Judgments

Judgments are the mental events of affirming and committing oneself to the truth of a proposition. They too are mental representations that can play a role in reasoning, action-guidance, or report. Different types of judgments can be distinguished. I shall distinguish two.

Type 1

Consider the following episode of theoretical reasoning. Suppose I reflect on whether whales are mammals or fish. Suppose too that I then recall that (i) whales lactate, and that (ii) only mammals lactate. On the basis of (i) and (ii), I then conclude that (iii) whales are mammals.

Since the contents involved in this piece of reasoning are globally available for inferences, action-guidance, and report, this is an episode of conscious thinking. Furthermore, since I arrive at (iii) via retrieving (i) and (ii), and since (i) and (ii) are what I believe, it follows that in the episode of thinking I draw on my beliefs about (i) and (ii) to settle whether whales are mammals. That is, I activate previously formed beliefs with the contents (i) and (ii).

The activation and retrieval of the beliefs might result from effortful processing (e.g.,

when I'm trying to remember whether *p*) or happen spontaneously without much effort (e.g., when I immediately recall that *p*). In either case, the events that then occur in consciousness are *judgments* about (i) and (ii). These events, in which existing beliefs are activated and their contents currently entertained for an on-going activity, are the first kind of judgments that I shall distinguish. They correspond to what are sometimes called 'occurrent' beliefs. Since beliefs are states, this is arguably a misnomer. In any case, I shall refer to them as judgments.

Type 2

The events in which pre-existing beliefs are activated aren't the only kind of judgments. When upon reflection on whether whales are mammals, I conclude that whales indeed are mammals, or when, without prior reasoning, e.g., spontaneously, it occurs to me that they are, then the events that at that moment occur in consciousness are judgments also. They are the events of newly settling[8] whether *p*, and form the second kind of judgments that I shall distinguish.

Typically, judgments of this kind lead to standing beliefs with the same content. But not all of them do, for belief formation might be affected by non-doxastic factors, e.g., biases, fear, etc., that may prevent a judgment *p* from leading one to acquire the behavioural dispositions characteristic of believing *p*. For instance, a subject S might judge that BA degrees from foreign countries are of an equal standard to those of her own country while her decisions when, e.g., hiring people or making recommendations reveal that she doesn't really hold the corresponding belief (Peacocke 2003). Arguably, what happens in such a case is that a non-doxastic factor, a bias, prevents the judgment involved from having the causal effects on thought and behaviour that are stereotypical[9] of a belief with the same content. Still, it is widely accepted that normally judgments (of type 2) do have these effects and do issue in beliefs with the same content.

In sum, then, judgments, as I shall understand them in this thesis, are the events of

[8] Judgments of type 1 also settle a matter for the subject. For instance, when S recalls that *p* then the matter as to whether *p* is settled for her. But it isn't then *newly* settled.

[9] When judgments don't lead to beliefs, they might nonetheless give rise to belief-like states, e.g., states that dispose one to assert that *p* and self-ascribe the corresponding belief that *p* but lack other properties of a paradigmatic belief. There is some debate on whether or not such states are beliefs proper (see, e.g., Bayne and Hattiangaddi 2013). For present purposes, I reserve the term 'belief' for paradigmatic beliefs.

affirming and committing oneself to the truth of a proposition. They are involved in activating previously formed beliefs or in newly settling a theoretical matter, which typically amounts to the formation of a belief.

Two more distinctions with respect to judgments will be relevant below. There are *perceptual* and *non-perceptual* judgments. For instance, when I see the face of the woman walking in front of me suddenly as that of my girlfriend, or when I suddenly hear a muffled tune as my phone ringing, then these are perceptual judgments. They are tied to my immediate perceptual experience. In contrast, when I conclude that whales are mammals, or that 17 is a prime number, these are non-perceptual judgments. They might carry abstract content and are not tied to my immediate perceptual experience. In what follows, I will only focus on non-perceptual judgments and whenever the term 'judgment' is used then these kinds of judgments are meant.

Further, judgments might be *conscious* or *unconscious*. Conscious judgments are judgments that occur in the workspace. I will mostly be concerned with conscious judgments. In general, when judgments figure in consciousness and become effective in, e.g., my reasoning above from (i) to (iii), then I'm *accessing* them.

Note that the term 'accessing' here does not imply representing the judgments themselves in addition to their contents. In general, accessing representations in conscious thinking doesn't necessarily involve representing these representations because representations can be broadcast in the workspace and be used in first-order reasoning even when the particular system responsible for forming meta-representations, i.e. mental representations that represent one's own mental representations as such, isn't then consuming them. Meta-representation isn't a condition for representations to be broadcast and figure in working memory (see Prinz 2011).[10]

Self-ascriptions

Among the meta-representations are self-ascriptions of beliefs. I take a self-ascription of a belief *p* to be a conscious second-order judgment with the content *I believe p*, not a linguistic expression. A self-ascription of a belief *p* is required for self-knowledge of the belief but it isn't sufficient, for in order to count as knowing that *p*, it isn't sufficient that one judges or believe *p*. There are different views on what exactly is required for

[10] There will be more discussion of and support for this point in Chapter 2.

knowledge that *p* (see Ichikawa and Steup 2012). I shall assume that knowledge is justified and true judgment or belief. This view isn't without some well-known problems (see, e.g., Gettier 1963). However, I shall not worry about these problems here but will instead only focus on justification and truth. In general, while the thesis is about self-knowledge of beliefs, most parts of it pertain more specifically to self-ascriptions of beliefs (i.e., to second-order judgments). The issue of justification and truth of these self-ascriptions will only briefly figure in the thesis, in Chapters 3 and 4, where I discuss TM-based accounts of self-knowledge and develop the IDR theory.

CHAPTER 2

The accessibility of judgments in conscious thinking

While my main focus in this thesis is on self-ascriptions and self-knowledge of beliefs, which is a matter of second-order thinking, in the present chapter, I will take a step back and consider the kind of access we have to our own judgments in conscious first-order thinking.

It is a common and often tacit assumption that when upon reflection on whether *p*, we judge *p*, then the judgment is directly accessible to us in conscious thinking for further reasoning, action-guidance, or report. That is, it is assumed that we can access our own judgments without inferring them from some intermediary representations that express them. I shall call this the *direct access* (DA) *view*.

The DA view has recently come under attack. It has been argued that at the conscious level, we have only *indirect* access to our own judgments via intermediaries, i.e. other representations such as visual imagery or inner-speech tokens that express them and first need to be interpreted and their underlying attitudes self-ascribed in order to acquire attitude-like causal roles (Carruthers 2011, 2014, 2015; Frankish 2009, 2012). I shall call this the *indirect access* (IA) *view.*

Which one of the two views is correct?

In this chapter, I contend that the case for the IA view doesn't in fact threaten the DA proposal, that the IA view is arguably false, and that the DA alternative is correct. After that, I develop the DA view further by specifying the mechanism that underlies our direct access to attitudes in first-order reasoning. I argue that the mechanism at issue operates at the sub-personal level and detects and differentiates the attitude types of the representations that figure in first-order reasoning without representing the attitudes themselves. I shall propose that it is able to do so by tracking the vehicle properties of the representations.

In section 1, I introduce the case against the DA and for the IA view. It has been put forward by Peter Carruthers (2011, 2014, 2015). After that, I briefly turn to Keith Frankish's (2009, 2012) version of the IA view and relate it to Carruthers' proposal. In sections 2, I then argue against the IA view and for the DA alternative. Section 3

elaborates the DA view by *inter alia* casting light on the mechanism involved in one's direct access to attitudes in first-order reasoning.

1. The case for the indirect access view

Carruthers (2011, 2014, 2015) claims that we only have indirect access to our own attitudes such as judgments and decisions in conscious thinking. His argument relies on empirical data and theoretical considerations pertaining to the contents of the global workspace and working memory. I have already introduced the global workspace and working memory in Chapter 1 and will now proceed directly to an exposition of Carruthers' argument.

As noted, the representations involved in conscious thinking are those that figure in working memory. This mean that if subjects had in conscious thinking direct access to their own attitudes, e.g., judgments, decisions, etc., then these attitudes would need to be able to enter working memory. With this in mind, in his case for the IA view, Carruthers turns to the question of what kind of representations can enter working memory.

He argues that imaging studies show that with the use of working memory, mid-level sensory areas are always co-activated and damage to or impeding the function of these areas by, e.g., transcranial magnetic stimulation (TMS), negatively affects performance in working memory tasks that involve *inter alia* abstract thoughts (2014: 154). Carruthers holds that these and other findings and theoretical considerations, which I shall not rehearse here, suggest that attention directed at mid-level sensory areas is the "main causal determinant of" and a "necessary condition for global broadcasting" and "entry of a given […] representation into working memory" (2014: 153).

He argues further that attention directed at mid-level sensory areas leads to the production of sensory or (in the off-line activation case) imagistic representations (e.g., visual imagery, inner speech etc.). Assuming that attention directed at mid-level sensory areas is a necessary condition for the broadcast of representations, it then follows that working memory processing always involves sensory or imagistic (henceforth 'sensory-imagistic') representations, Carruthers holds (2014: 148f, 2015: 92f).

He contends that this doesn't mean that only sensory-imagistic contents can enter the workspace. Conceptual content can do so too provided that it becomes "bound into" sensory-imagistic representations, which happens when the representations that carry that content are co-activated with sensory-imagistic ones (Carruthers 2014: 148, 2015: 64f).

In addition, two kinds of attitudes can enter working memory also: (i) "sensorily embedded judgments" (e.g., when one suddenly sees something as a friend's face), and (ii) affective or "momentary felt desires", for, in general, sensory and affective representations can be broadcast, Carruthers claims (2015: 80). However, he insists that there is "no evidence" that other attitudes can enter the workspace, or that there is a separate "purely amodal workspace", which could "operate independently of sensory representations" (2014: 155f).

Carruthers argues further that if attitudes that are not of type (i) or (ii) could enter working memory then reporting their occurrence should be "trivially easy", and "self-knowledge" of them "should not need to be *interpretive*",[11] because they would be available to whatever systems are responsible for the formation of self-ascriptions and report (2015: 113). Empirical studies show, however, Carruthers continues, that subjects often unknowingly confabulate or interpretively self-ascribe their own attitudes, which suggests that they are poor at accurately identifying them (ibid). These findings are "best explained" by "an account that does not postulate the global broadcasting of propositional attitude events in working memory", Carruthers claims (2014: 155).

He grants that some sensory-imagistic representations, e.g., inner-speech tokens, might *seem* to be attitudes, e.g., judgments or decisions. But, in his view, they aren't, because they lack the right functional profiles. For instance, judgments settle reasoning on an issue (e.g., via the activation of an existing belief or the formation of a new one) and are immediately available for action-guidance. Imagery, however, e.g., an inner-speech utterance 'It will rain today' doesn't have these properties, for I might say to myself in inner speech 'It will rain today' while remaining agnostic about the matter. Further reasoning is required to settle an issue and lead to action, Carruthers holds (2011: 103-105; 2014: 156). He argues that what gives any kind of "image" say a judgment or "belief-like rather than [say] a supposition-like causal role will depend on one's

[11] All italics in the quotes of this paper are original.

interpretation of its nature", which typically happens unconsciously and is performed by the "mindreading system", i.e. the inferential capacity that enables a subject to determine and predict people's mental states and events on the basis of people's behaviour, the context, or circumstances (Carruthers 2014: 152, 2011).

Carruthers' argument against the DA view and for the IA alternative can be summarised thus.

(1) The representations that subjects are able to directly access in conscious thinking are those that can enter working memory.
(2) Empirical and theoretical considerations on working memory tell us that the latter is sensory-based in that only sensory-imagistic and affective representations (plus conceptual contents embedded in them) can figure in the workspace.
(3) Apart from (a) sensorily embedded judgments and (b) affective attitudes such as felt desires, these representations aren't themselves attitudes, for they don't play the right functional roles.
(4) Apart from (a) and (b), no attitudes are broadcast in working memory.
(5) Thus, apart from (a) and (b) (I will drop this qualification in what follows and take it as read), subjects have no direct access to their own attitudes in conscious thinking.

On Carruthers' view, when we retrieve attitudes, e.g., beliefs, and bring them to bear in conscious thinking then this works as follows. An unconscious

> belief or non-sensory judgment might issue in an episode of inner speech with the same or sufficiently similar content. When interpreted, the latter can assume a causal role somewhat like that of a judgment […], and its content will be accessible to all of the systems that consume global broadcasts. This might activate additional stored information or relevant goals, issuing in yet further episodes of inner speech or in the evocation of suitable visual or other imagery. As a result, it is true that any belief or other attitude can be brought to bear in the evaluation of any belief or decision [but this happens only] indirectly, through the effects that attitudes can have on the contents of the global workspace. (Carruthers 2014: 158, 2015: 15)

The view isn't idiosyncratic. Frankish (2012) advocates a very similar proposal. He

specifies his account in terms of the "dual-system theory" of reasoning (ibid).

According to the dual-system theory, humans possess two reasoning systems, often called 'System 1' or '(S1)' and 'system 2' or '(S2)'. System 1 (S1) is typically taken to be a collection of autonomous sub-systems, whose operations are fast, automatic, non-conscious, parallel, and independent of working memory. In contrast, system 2 (S2) is thought to process information in a slow, controlled, conscious, and serial manner that is demanding of working memory. While many researchers have assumed that the two systems are entirely distinct, this assumption has been challenged. On the currently most common view, the defining feature of S2 is that it utilises working memory while S1[12] doesn't do so (Evans and Stanovich 2013).

Both Frankish and Carruthers endorse this kind of dual-systems picture. Furthermore, they both claim that S2 reasoning processes are sensory-based. They also agree that in order to count as an attitude such as a judgment or decision, any S2 event will need to be able to settle reasoning (Frankish 2012: 41; Carruthers 2015: 194f). The subtle difference between them is that, unlike Carruthers, Frankish claims that sensory-imagistic S2 events *can* settle reasoning and hence qualify as attitudes themselves.[13] He holds that

> we should think of S1 decisions and judgments as terminating reasoning at the S1 level, and S2 decisions and judgments as terminating reasoning at the S2 level. […] [O]n this reading, suitable rehearsed utterances *would* count as decisions and judgments. For they would settle the matter *as far as S2 reasoning is concerned*. (Frankish 2012: 46)

But Frankish too thinks that imagery such as a "rehearsed utterance assumes the causal powers of a decision or judgment only when it has been interpreted and has generated a suitable higher-order S1 belief" (Frankish 2012: 46). Hence, imagistic representations are, on his view, only "virtual" attitudes: they are "realised by a combination of the rehearsed utterance and the resulting higher-order S1 belief" (Frankish 2012: 46-47). So, while for Carruthers sensory-imagistic representations are mere intermediaries of

[12] Some theorists take S1 to be a multitude of systems. I ignore this here, as it won't matter for my argument.

[13] See Frankish (2012) for his motivation for holding that there are S2 attitudes also.

attitudes, for Frankish, they aren't just intermediaries but can become 'virtual' attitudes themselves when they are realised in higher-order S1 attitudes. Given this, on Frankish's view, one could directly access one's own, e.g., 'virtual' judgments via accessing one's own sensory-imagistic representations. However, these representations will on his view only qualify as judgments if they are interpreted first. Since that is so, I shall treat Frankish's proposal as a version of the IA view. For I take it that if we had direct access to our own attitudes, then no interpretation of imagery, not even an unconscious one, would be needed.

2. The case for the direct access view

In this section, I want to show that the argument for the IA view doesn't in fact undermine the DA view. After that, I offer two points to the effect that the IA view is false. The second one will also provide positive support for the DA view.

Consider again Carruthers' claim that conceptual information can enter the workspace when it is "bound into" sensory-imagistic states (2014: 146). There are two different ways of understanding the locution 'bound into'. On the first one, which is Carruthers', "conceptual information that is activated by interactions between mid-level areas and the association areas gets bound into the content of attended perceptual states" *prior* to entering the workspace (2014: 148). On this view, the conceptual information can't directly enter the workspace. On the second reading, however, proposed by, e.g., Wu (2014), the "binding" at issue would involve a

> coactivation of two sorts of representations where their informational contents both continue down the computational chain into cognition. For example, I hear sounds as having a certain meaning because of the co-activation of sensory, syntactic, and semantic representations that all feed forward. There is in this sense binding of conceptual and non-conceptual contents, but it is consistent with the dynamic model [of the workspace, according to which the "contours of the workspace fluctuate as different brain circuits are temporarily mobilised, then demobilised" (Dehaene and Naccache 2001: 13),] and assumes an architecture that allows conceptual areas direct access to downstream processing relevant to central cognition. (172)

Wu (2014) argues that both proposals on how to understand the notion of conceptual information being 'bound into' sensory-imagistic representations are in line with the empirical studies that suggest that there is a co-activation of sensory areas in every working memory task. Importantly, however, unlike on the first, i.e. Carruthers' view, on the second proposal, beliefs and other attitudes can themselves enter the workspace: they can travel down the 'computational chain' into the workspace alongside sensory-imagistic representations. The co-activation data that Carruthers cites doesn't undermine this. It hence doesn't in any way threaten the DA view.

Carruthers responds that if, as the DA view implies, non-sensory and non-affective attitudes could enter the workspace themselves, then subjects should be able to detect and reliably report them. But, he continues, social psychological studies on, e.g., confabulation and self-interpretation[14] suggest that subjects often fail to correctly self-ascribe their own attitudes. Carruthers takes this to show that they aren't broadcast in the workspace (2015: 113).

However, this point is unconvincing. For the formation of self-ascriptions of attitudes requires disengaging one's first-order reasoning and considering a different topic, namely one's own mental states. Since that is so, it might be that in one's first-order reasoning, one *does* have direct access to one's own attitudes and these attitudes are in the workspace, but as soon as one shifts the topic of one's current thinking to one's own mental states, one's first-order attitudes are no longer broadcast, forcing the system responsible for producing self-ascriptions to operate on the contents that remain in workspace and are currently available (including, e.g., representations about one's current overt and covert behavior). Given this possibility, the confabulation and self-interpretation studies that Carruthers mentions don't undermine the view that judgments can, in conscious first-order thinking, enter the workspace alongside sensory representations and are then separate from them directly accessible to the subject. Carruthers' case for the IA view thus doesn't in fact threaten the DA view.

In the next sub-section, I shall take this further and argue that the IA view is false. I shall make two points. The first one pertains to the question of how the IA view explains the formation of judgments in conscious theoretical reasoning. The second one relates to empirical findings on autism.

[14] I will say more on these studies in Chapter 5.

2.1 Conscious theoretical reasoning

Suppose I reflect on whether whales are mammals. I find that (i) whales lactate, and that (ii) only mammals lactate. Based on (i) and (ii) I then judge that (iii) whales are mammals.

This is an episode of conscious theoretical reasoning in which I form a particular judgment about whales that I wouldn't have formed unless the contents (i) and (ii) had occurred to me in consciousness, i.e. unless these contents had been broadcast in the workspace. How would advocates of the IA view explain this piece of reasoning and judgment-formation?

According to the IA view, only affective and sensory-imagistic representations are directly accessible at the conscious level. Suppose then that (i) and (ii) are contents carried by sensory-imagistic representations, e.g., inner-speech tokens, that are broadcast to the various judgment- and decision-making systems connected to the workspace.

On the IA view, the judgment-forming system producing the judgment that whales are mammals can't take these representations straight away as input and in response produce the judgment that whales are mammals as output. This is because, on the IA view, for all that the systems connected to the workspace can tell, (i) and (ii) might be judged, supposed, doubted, hoped, feared etc. And when the judgment-forming system can't determine whether (i) and (ii) are judged (or believed) as opposed to supposed, doubted, hoped, etc., it can't produce the judgment that whales are mammals anymore than it can produce the judgment that it is noon on the basis of the *supposition* that the clock says it is noon and if the clock says this then it is noon. On the IA view, the representations about (i) and (ii) will need to be interpreted first because what gives an "image a [say] belief-like rather than [say] a supposition-like causal role, will depend on one's interpretation of its nature", and hence on "mindreading" (Carruthers 2014: 152; Frankish 2012: 46).

Suppose this is right, what happens after the interpretation? How do I arrive at the conclusion that whales are mammals? Carruthers (2014) claims that the "best account that we have of how interpreted images achieve their attitude-like effects is that the

latter result from the intervention of higher-order beliefs and goals" that themselves "remain unconscious" (155f).

To see how this is meant to work, it helps to first consider the IA explanation of how imagistic representations attain decision-like roles before returning to judgments. It is captured in the following passage:

> Suppose, for example, that the sentence, 'I shall go to the bank', is tokened in inner speech. Under interpretation by the language faculty working together with the mindreading system, this might be heard as expressing a decision to go to the bank. […] If I have a standing desire to do what I have decided, […] then the belief that I have made a decision may evoke a higher-order motive to execute that decision. These taken together might then cause me to walk over to the bank, or to form the intention of visiting the bank. […] [T]he underlying practical reasoning will look something like this: I have decided to go to the bank; I want to do what I have decided; so I shall go to the bank. The final [unconscious] event in this sequence is *itself* a decision to go to the bank. (Carruthers 2014: 156)

Advocates of the IA view hold that a similar line of thought applies to judgments:

> After evaluating various hypotheses about the future of the economy, I may rehearse the utterance, 'So there will be a recession'. And when broadcast, this may give rise to an [unconscious higher-order] belief that I am committed to the view that there will be a recession, or that I have judged that there will be, and given a desire to execute my commitments, or to act in line with my judgments, this will lead to behavior appropriate to a judgment that there will be a recession. (Frankish 2012: 45)

The claim is that when one "interprets oneself as judging that *p*", then the "behavior appropriate to the judgment" will, given one's desire to act in line with one's judgments, involve not only "acting in ways required by the truth of *p*, but also […] reasoning in ways required by it", i.e. "taking *p* as a premise, dismissing hypotheses that conflict with *p*, and so on" (ibid; see also Carruthers 2011: 104f).

Why assume that one has a desire to act in line with one's judgments? The thought is that one has this desire because one wants to avoid inconsistencies between one's

actions, mental or overt, and one's attitudes (ibid).

With this in mind, how would advocates of the IA view explain my arriving at the conclusion that whales are mammals? Given the points just mentioned, they would propose that this happens via my engaging in the following unconscious piece of reasoning:

(a) I have judged that (i) whales lactate and that (ii) only mammals lactate. [This is a higher-order judgment resulting from my mindreading system's interpretation of imagery.]
(b) I'm committed to acting and reasoning in ways required by the truth of (i) and (ii). [This is because I understand that I'm *judging* (i) and (ii).]
(c) Act in line with this commitment/judgment. [This is a higher-order desire, for I wish to avoid attitude-action inconsistencies.]
(d) Doing so requires accepting the implications of (i) and (ii).
(e) It is an implication of (i) and (ii) that whales are mammals.
(f) Accept that whales are mammals. [This is a decision.]
(g) Whales are mammals.

The problem with this is perhaps not hard to see. (g) is formed on the basis of my *wanting* to act in line with my judgments (because I wish to avoid attitude-action inconsistencies) and on evidence that doesn't in fact support the truth of the judgment that whales are mammals. At best, only evidence for my *judging* that whales lactate and that only mammals lactate is brought to bear. And this is quite irrelevant to whether or not whales are as a matter of fact mammals. Since I can't come to *judge* that whales are mammals merely on the basis that I want to act in line with my judgments (i) and (ii), and since I also can't judge that whales are mammals on the basis of evidence that is irrelevant to the truth of whales' being mammals, (g) can't be a judgment. If it is claimed that the unconscious system producing (g) *can* produce a judgment that whales are mammals on the basis of (a)-(f), then decisions would come to play the role of judgments and hence themselves be judgments, which is absurd.

This isn't to deny that one is able to *indirectly* bring it about that one has a certain attitude. For instance, over time by exposing oneself only to particular pieces of information, by acting as if one believes *p* etc., one might come to believe *p*. This might well work. What seems absurd is rather that with the decision and commitment to act as

if one has judged *p*, one automatically makes the self-ascription of the judgment *p* true. This, however, is what advocates of the IA view tend to claim. Here, for instance, is Carruthers (2010):

> If I interpret myself as having formed a judgment that *p,* or as having decided to do *q*, then I thereby commit myself to thinking and acting in the future on the assumption that I believe that *p,* or that I intend to do *q*. (And note this will be true even if the initial interpretations had been inaccurate in themselves.) My self-attributions thereby become self-fulfilling. (106)

In another place:

> the data available to introspection (inner speech and the like) might give rise to a higher-order belief that I have just formed a belief in the permissibility of capital punishment. And then this combined with my desire for consistency will explain my future patterns of thinking and acting. Hence my self-attribution of belief becomes self-fulfilling, even though it may originally have been formed from a confabulated interpretation of introspective and other data. (Carruthers 2010: 102)

Applied to judgments, the thought is that even when I have not formed the first-order judgment *p* and when the self-ascription of the judgment is false, the self-ascription that I judged *p* will in conjunction with "my desire for consistency" nonetheless bring about the *same* effects as the judgment *p*: "self-attributions become self-fulfilling" (Carruthers 2010: 106).

Advocates of the IA view indeed need to make the odd claim that a decision to act in line with the judgment *p* even brings about a commitment to the truth of *p*. For if the decision to act in line with the judgment *p* does not also (just as a judgment *p*) produce a commitment to the truth of *p* then I can't form the (unconditional) judgment that whales are mammals on the basis of (i) and (ii), which I *do*. I can only form the (unconditional) judgment that whales are mammals on the basis of (i) and (ii) if I commit to the truth of (i) and (ii). If I merely supposed (i) and (ii) are true, or took it that (i) and (ii) are true only under the condition that some other state of affairs obtains then I won't be able to judge that whales are in fact mammals. So in order to be able to explain how I can in my reasoning about whether whales are mammals come to form (on the basis of (i) and (ii)) the judgment that whales are mammals, the advocate of the IA view can't help but hold that decisions can somehow be judgments.

Since the preceding is based on what in Carruthers' own words is the "best account that we have" of how indirect access to attitudes in conscious thinking works (2014: 155f), and since the account lead to an absurd conclusion, for judgments are clearly *not* a matter of direct decision, we now have a *reductio* of Carruthers' (2014), and Frankish's (2012) IA view.

It is worth considering Carruthers' (2015) more recent proposal on how judgments and new beliefs are formed via conscious theoretical reasoning. He holds that in an episode of "discursive reasoning" that leads to a new belief, inner speech, e.g., an inner-speech token '*p*' is "processed by the language and mindreading system, with the result that one hears" the sentence as "*asserting*" or as "expressing the judgment" that *p*:

> If no inconsistencies with one's existing beliefs are detected and if we assume that one's own assertions (and judgments) are treated as reliable by default, then one will […] form the belief that […] [*p*]. In effect, one believes one's own assertions, much as if they were statements from a reliable informant. (Carruthers 2015: 188)

On this proposal, inner-speech tokens still need to be interpreted by the mindreading system to attain a judgment-like role, but there is now no longer an unconscious piece of practical reasoning involved in which the higher-order judgment produced by the mindreading system is combined with a higher-order desire for consistency to form a decision to reason and act in line with the judgment. Rather, the proposal is that the inner-speech token occurs, is interpreted as an assertion and then gives rise to a belief with the same content that ensures one's reasoning and acting in line with what one asserted in inner speech.

To see the problem with this view, it again helps to consider how it would explain my above inference pertaining to whales. Presumably, the account would be this. In my reflection on whether whales are mammals, at some point, I say to myself in inner speech 'whales lactate' and 'only mammals lactate'.[15] The imagery is subsequently interpreted by the mindreading system, and I hear them as "assertions" (Carruthers 2015: 188). Since I take my assertions to be true, I come to form the belief that (i) whales lactate, and the belief that (ii) only mammals lactate. Since judgments and

[15] Arguably, my saying these things will be grounded in my believing them. This means that oddly, on Carruthers' view, I would come to form all the beliefs I retrieve in conscious thinking twice. For presumably my retrieving my beliefs in conscious thinking will then involve the formation of assertions about their content in inner speech.

beliefs themselves aren't, on the IA view, broadcast, the inference to the conclusion that whales are mammals is then drawn outside the workspace on the basis of my first-order beliefs (i) and (ii). The subsequent judgment that whales are mammals leads to the production of a further inner-speech sentence 'whales are mammals', which is again accessible at the conscious level, and the episode of reasoning is over: I will hear 'whales are mammals' again as an assertion of mine and so have attained an answer to my initial question of whether whales are mammals.

The problem with all this is the following. My reasoning started with me wondering if whales are mammals. According to the proposal at hand, I then heard myself asserting that whales lactate and that only mammals lactate. Since hearing myself asserting that *p* requires, on Carruthers' view (2015: 81), that I represent myself at the conscious level as asserting *p*, at the conscious level, upon wondering whether whales are mammals, I will represent something along the lines of *I'm asserting that whales lactate*, and *I'm asserting that only mammals lactate*.[16] If this were right, however, then at the end of the reasoning, at the conscious level, the only support I would have for the view that whales are mammals would be that I *assert* that whales lactate, and that I *assert* that only mammals lactate. But this fails to capture my reasoning: when I conclude that whales are mammals then I don't take this conclusion to be supported by what is for me just *assertions* of mine. I will rather find it supported by the worldly facts that whales lactate and that only mammals lactate. It is these facts that sustain my conclusion and settle for me the matter of whether whales are mammals. They are my basis for deductively inferring that whales are mammals.

On Carruthers' view, however, I can't draw the inference at the conscious level. For, on his view, I'm at that level of processing only presented with the inner-speech tokens 'whales lactate' and 'only mammals lactate' – which by themselves can't serve as a basis for the inference at issue – and the second-order contents *I'm asserting that whales lactate*, and *I'm asserting that only mammals lactate* – which don't support the conclusion that whales lactate either. Carruthers' view thus seems to preclude conscious deductive inferences. Since that is so, it arguably is mistaken, for there *are* conscious deductive inferences. This is evidenced by the fact that if I were subsequently asked why whales are mammals, I would say that this is because whales lactate, and only mammals lactate. The fact that for me whales are mammals because whales lactate and

[16] Carruthers (2015: 81) puts the content of the second-order representation slightly differently. But he insists that one represents one's asserting that *p*, and that is all that is required for my point here to hold.

only mammals lactate indicates that the first-order contents *whales lactate*, and *only mammals lactate* are presented to me in a way that for me settles how things are. There is hence reason to believe that the representations about these contents are conscious judgments that I can directly access in my conscious thinking. This contradicts the IA view.

2.2 Autism and logical reasoning

There might be other versions of the IA view than the ones just discussed. But in its most general form, the IA view holds that in conscious thinking, we can access our own judgments only via intermediary representations from which we need to infer (unconsciously) the underlying attitude. On any such view, a representation that identifies the attitude underlying the intermediary representation will at some point come into play because the result of the inference from the intermediary representation will be a (conscious or unconscious) representation that classifies the attitude. If some version of the IA view were correct, then, one would expect that subjects who have difficulties representing judgments also have difficulties in tasks requiring conscious thinking that involves the accessing of judgments, e.g., for the purpose of theoretical reasoning.

It is well known that people with autism tend to fail at representing their own and other people's minds including what they judge to be the case (Leslie and Thaiss 1992; Baron-Cohen 1995; Williams 2010).[17] If autistic subjects with an impaired ability to represent their own judgments performed normally in conscious thinking that involves the accessing of judgments then this would contradict the IA view and support the DA alternative.

Unfortunately, so far, there is no autism study in which subjects were tested in both self-related mentalising, i.e. thinking about mental events or states, and first-order reasoning. There are, however, autism studies in which subjects' *other*-related mentalising and first-order reasoning were explored (e.g., Scott and Baron-Cohen 1996; Scott et al. 1999). Furthermore, there are also studies that indicate that subjects with autism who exhibit a deficit in representing other people's judgments have the same

[17] This is not to say that all of them do. High-functioning autistic individuals might be able to ascribe beliefs correctly, but even they seem to use a different strategy to do so compared to neurotypical subjects (Frith and Happé 1999).

deficit in representing their own judgments (see, e.g., Williams and Happé 2009; Williams 2010).

With these points in mind, consider the following study on reasoning in autism. Scott and Baron-Cohen (1996) gave autistic and normal children two tests of abstract reasoning and a theory-of-mind (ToM) test. The ToM test was a standard mindreading task, the "false-belief" or "Sally-Anne" test (see Wimmer and Perner 1983). To pass it, one needs to understand that what another subject takes to be the case is different from what is in fact the case. That is, one needs to be able to represent the subject's judgment about what is the case. As for the two abstract reasoning tests that Scott and Baron-Cohen gave their subjects, these tests included a transitive inference task and an analogical reasoning task. In the transitive inference task, the children had to reason about relations between items, e.g., X→Y, Y→Z, so X→Z. In the analogical reasoning tasks, they had to reason about higher-order relations between items. For instance, they were presented with picture cards of different objects. Say, the first card was of a banana and the second one of a banana cut into pieces. On the third card was, say, a melon, and the fourth slot was left empty. To test their analogical reasoning ability, the children were asked to complete the sequence by choosing another card from five options, four of which were distractors. To succeed in the task, they had to choose the card with a melon cut into pieces.

Most autistic children failed the mindreading test, which suggests that they had an impaired ability to represent a subject's judgments. Interestingly, however, Scott and Baron-Cohen found that the children nonetheless performed comparably to the control groups on both the test of transitive inferential reasoning and the test of analogical reasoning. Furthermore, after each answer to one of the questions in the analogical reasoning test, Scott and Baron-Cohen also asked the children for a justification for their answer. This was to ensure that they were not merely guessing but were actually engaging in reasoning. Scott and Baron-Cohen found no significant difference between the justifications that the autistic children and the controls provided for their answers.

Note that to justify their judgments, the children had to mention the contents of the representations that sustained their conclusions, i.e. the contents of the representations that for them settled facts about abstract higher-order relations (e.g., A is to B as C is to

D). The representations about these abstract contents hence had to figure in working memory to provide the children with a justificatory basis for their response.

Returning with this to the initial issue, the IA view claims that the representations that enter working memory need to be decoded for their underlying attitudes and these attitudes need to be represented in order for the broadcast representations to play, e.g., judgment-like roles, i.e. to settle a matter for a subject and serve her as a justification in reasoning and for overt actions. If the IA view were right then the same should have been the case with respect to the representations that the children in Scott and Baron-Cohen's study had to access to support their conclusions. However, the autistic children exhibited in the mindreading test significant impairments in the ability to represent other people's judgments on what is the case. Since autistic subjects with deficits in representing other people's mental states and events typically have a corresponding deficit in representing their own mental states and events (see Williams and Happé 2009; Williams 2010), there is ground to hold that the representations that the autistic children accessed at the conscious level when performing the reasoning test didn't have to be inferentially decoded for their underlying attitudes in order to play attitude-like roles. That is, Scott and Baron-Cohen's data speak against the IA view, for the IA view holds that subjects can't directly access their own judgments[18] in conscious thinking but only representations whose underlying attitudes need to be inferred and represented first for them to attain attitude-like roles (Carruthers) or become "virtual" attitudes (Frankish).

Since the autistic children in the experiments didn't appear to require an ability to represent mental states or events to perform comparably to neurotypical children on the reasoning tests, presumably the neurotypical children participating in the experiment didn't need this either. For it seems unmotivated to assume that they had to employ a much more sophisticated cognitive procedure than the children with autism in order to be able to exhibit an equivalent performance. Thus, against the IA view, subjects can directly access their own judgments in conscious thinking.

[18] Note that the judgments that the children had to access carried abstract contents (pertaining to higher-order relations) and were not directly grounded in their perceptual experience. The judgments at issue hence can't plausibly be viewed as sensorily-embedded attitudes.

It might be argued that recent studies involving the violation-of-expectation paradigm and gaze tracking indicate that infants as young as 7 months are able to register other subjects' false beliefs (Onishi and Baillargeon 2005; Surian et al. 2007; Kovács et al. 2010). This ability is often referred to as "implicit theory of mind", and taken to be automatic and unconscious in nature (see, e.g., Low and Perner 2012; Schneider et al. 2015). Implicit theory of mind is operational long before children are able to pass the standard, verbal false-belief test, which is typically taken to probe explicit mindreading. With this in mind, in defense of the IA view, it might be maintained that, even though the autistic subjects in Scott and Baron-Cohen's study failed the standard verbal false-belief task, they could still have had a functioning *implicit* theory of mind, which would then have allowed them to indirectly access their own judgments (via mindreading inferences) for the purpose of first-order reasoning. If that is right, then Scott and Baron-Cohen's findings don't undermine the IA view.

However, the proposal that the autistic children could still have had a functional implicit theory of mind that enabled them to successfully perform the analogical reasoning is highly implausible. A recent review of both explicit and implicit ToM studies in autism found that with

> regard to implicit ToM processing, available eye tracking evidence consistently indicates that implicit ToM reasoning is impaired in ASD even in high-functioning adults with ASD, who pass experimental explicit ToM tasks. Individuals with ASD seem to lack a spontaneous sensitivity to other's mental states. (Sodian et al. 2015: 127)

It is thus unlikely that the autistic children in Scott and Baron-Cohen's study might still have had a fully functional implicit ToM. I conclude, then, that Scott and Baron-Cohen's finding that autistic children with theory-of-mind impairment still perform normally in logical reasoning tasks provides good support for the assumption that subjects can access their own judgments in conscious thinking without first having to infer them from some intermediary. That is, the data at issue suggest that the IA view is false and support the DA alternative.

3. Attitude-type detection without meta-representation

Still, there is a grain of truth in the IA view that the DA view needs to accommodate. The IA view holds that conscious first-order reasoning involves a mechanism that detects the attitudes of the representations that one accesses. By revisiting some of the points made earlier, I shall now argue that such a mechanism is indeed operative in both theoretical and practical reasoning and underlies our direct access to attitudes in conscious thinking. Unlike the IA view claims, however, the mechanism doesn't produce meta-representations but tracks the attitude types of representations via the vehicle properties of the latter.

To see the support for these claims, consider again theoretical reasoning. Suppose I wonder whether *p*, and then reason: *q*→*p*, *q*, so *p*. I might in principle hold many different attitudes towards the proposition *q*. I might, e.g., judge, suppose, doubt, etc. *q*. Furthermore, the judgment-forming system that produces the judgment *p* will presumably (as one of the systems connected to the global workspace) have access to representations of all these different attitude types. After all, it will in some cases need to be able to form, e.g., conditional judgments on the basis of content that is merely supposed. The judgment-forming systems connected to the workspace will thus need to be able to use representations of different attitude types.

Since that is so, these systems will need to be able to detect and differentiate the attitude types of the different representations broadcast to them. Because if the judgment-forming system that on the basis of *q*→*p*, and *q* forms the judgment *p* were not able to detect judgments and keep them apart from suppositions, then it couldn't form the judgment *p* on the basis of the representations about *q*→*p*, and *q*. For *q*→*p*, and *q* could then for the system, for all it can tell, be only supposed rather than judged. And it can't infer from, e.g., the supposition *q* and the judgment *q*→*p* that (in fact) *p*, because it doesn't follow from the supposition *q* and the judgment *q*→*p* that (in fact) *p.* To be able to produce the judgment that (in fact) *p* on the basis of the representations *q*→*p*, and *q*, the judgment-forming system needs to detect the attitude type of the representations involved in first-order reasoning.

However, to do so, the system can't employ a mechanism that produces as outputs representations of the attitude types of the representations at issue themselves. For suppose it did. Suppose the judgment-forming system were presented with the contents *I believe q*→*p*, and *I believe q*. It could then no longer infer from the representations

that it has available that *p*, for it doesn't follow from *I believe* $q \rightarrow p$, and *I believe q*, that *p*. In order to be able to form via deductive inference the judgment that *p*, the judgment-forming system needs to operate on the first-order representations $q \rightarrow p$, and *q*, and be able to track their attitude-types. That is, the judgment-forming system needs to have a mechanism that allows it to detect the attitude types of these representations without representing the attitudes themselves.

The point can also be illustrated with respect to practical reasoning. Practical reasoning typically leads to decisions. It settles what to do. A decision-making system normally combines a particular desire with a belief about what has to be the case in order for the desire to be satisfied. The combination of these two attitudes leads the system to produce a decision. For instance, suppose I want to get a drink and believe that there is a drink in the fridge. If all goes well, a decision-making system will combine the desire at issue with the belief and produce the decision (for me) to go to the fridge to get the drink. For the decision-making system to be able to form the decision in this way, it needs to have a mechanism available that allows it to keep track of the attitude roles of the two representations, i.e. of what is wanted and what is the case. Otherwise, for the decision-making system, I might desire rather than believe that there is a drink in the fridge, and I might believe that I get a drink rather than desire that I get a drink. If that were so, then the system couldn't on the basis of the representations involved produce the decision (for me) to go to the fridge to get a drink. To be able to produce this decision on the basis of the mentioned attitudes, the decision-making system thus needs to keep apart what content is believed and what content is desired. That is, it needs to be able to detect and differentiate the attitude types of the representations it consumes.

But as before, the attitude-type detection mechanism required for this to happen can't produce as its outputs meta-representations, i.e. representations of the desire to get a drink and of the belief that there is a drink in the fridge. Because if it did, then the belief-representation available to the decision-making system would no longer specify what needs to be the case for the desire to be satisfied. For the belief *I believe there is a drink in the fridge* doesn't tell the decision-making system where the drink actually is. It only pertains to a fact about my mind. Unless the information about the location of the drink is passed on to the system, it can't form a decision on how to act to satisfy the desire. Hence, practical reasoning too involves the operation of a mechanism that

detects and differentiates the attitude types of representations without representing the attitudes themselves.

In response it might be argued that perhaps at the sub-personal level all that is available is the belief *I believe there is a drink in the fridge* and not the belief *there is a drink in the fridge*. Maybe our cognitive system evolved so as to use, at the sub-personal level, the higher-order belief as a basis for the inference to the first-order conclusion.

But note that any such proposal will still be contradicted by the fact that subjects who have an impaired ability to (implicitly or explicitly) form representations about mental events or states, e.g., autistic individuals, or neonates (Block 2007), can nonetheless clearly engage in practical and theoretical reasoning to form decisions and judgments (see Scott and Baron-Cohen 1996, and Torralva et al. 2013 for relevant data). This casts serious doubt on the plausibility of the above response.

Suppose, then, that first-order reasoning does indeed involve a mechanism that tracks the attitude types of the representations involved without producing meta-representations. How does this work?

To begin with, the mechanism that the systems connected to the workspace employ for detecting and differentiating the attitude types of representations without representing the attitudes themselves needn't be viewed as *sui generis*. I want to suggest that it is functionally of the same kind as the one operative in what is often called 'interoception', i.e. the sensing of internal bodily changes, e.g., heartbeat, pain, visceral or muscular sensations, temperature, etc. (Goldman 2006; Craig 2015). The neural mechanism involved in interoception is the lamina I spinothalamocortical system which transmits activation from primary afferents that represents the physiological condition of the body to a thalamocortical relay nucleus, the posterior part of the ventromedial nucleus, or VMpo (ibid). In the VMpo are cortical representations of different bodily sensations, e.g., pain, visceral or muscular sensations, temperature etc. that are then via a homeostatic neural route mapped onto the internal bodily states. I propose that such a sub-personal mechanism for the tracking of internal bodily states is also employed by the judgment- and decision-making systems connected to the workspace. It allows these systems to detect and discriminate the attitude types of the representations that they access in theoretical or practical reasoning by tracking distinct neural activation patterns

that occur when the representations are tokened.[19] That is, the mechanism tracks attitude types via the intrinsic properties of the vehicles of the representations accessed in judgment- and decision-making.

This proposal illustrates how it can be that, even though the outputs of the mechanism at issue aren't meta-representations, they nonetheless allow the judgment- and decision-making systems involved in first-order reasoning to detect, differentiate, and correctly combine representations of different attitude types. What remains unresolved is how the attitude-type detection mechanism just introduced is able to treat specific properties of the vehicle of the representations accessed in first-order reasoning as indicative of, e.g., a judgment being instantiated rather than, say, a supposition. An answer to this question will need to be part of a full account of first-order reasoning. First-order reasoning is, however, not my primary focus in the thesis. I shall thus leave the issue open. What matters for my purposes here and in what follows is only that there exists a sub-personal mechanism that detects and differentiates attitude types without representing them.

4. Conclusion

In this chapter, I considered the nature of the access that we have to our own judgments in conscious thinking. I introduced a recent case for the view that we have only indirect access to our own judgments in conscious thinking via sensory-imagistic and interpretation-dependent expressions of them. I contended that the case for the IA view doesn't undermine the DA alternative and offered two points to the effect that the IA view is in fact false: the IA view (a) leads to an untenable picture of conscious theoretical reasoning, and (b) is contradicted by findings on autism. I argued that the second point doesn't only suggest that the IA view is mistaken but also provides positive support for the DA view. I then developed the DA view further by proposing that our direct access to our own attitudes in conscious first-order thinking involves the operation of a sub-personal mechanism that detects and differentiates the attitude types of representations without representing the attitudes themselves. I ended by suggesting

[19] Goldman (2006) also proposes a mechanism that tracks the attitude types of representations via neural properties. But he holds that it produces meta-representations as outputs. I will introduce and critique his view below.

that the mechanism is able to do so by tracking the neural properties of the representations that figure in judgment- and decision-making.[20]

[20] Before leaving the issue of access to one's own judgments in conscious thinking, it might be worth noting that if we assume that the empirical data show that workspace processing always implicates sensory-imagistic representations, then the DA view allows for three different versions. (1) It might be that the judgments accessed in conscious thinking are sensory-imagistic representations themselves. (2) It might be that they are not sensory-imagistic representations but rather separate amodal representations that accompany the latter into the workspace, carry the content, and play the attitude role of the judgments. (3) Or it might be that they are representations that integrate components of both sensory-imagistic and amodal ones, where the components of the latter kind of representation account for the judgment-effects of the integrated representations. To settle which one of these proposals is correct would *inter alia* require delving into the "imagery debate" (e.g., Block 1983; Tye 2000; Pylyshyn; 2003; Kosslyn et al. 2006). Here isn't the place to do so. What matters for my present purposes and for the following discussion is only that the case for the IA view doesn't undermine the DA view, and that there is reason to believe that subjects do have direct access to their own judgments in conscious thinking. These points hold independently of which one of (1)-(3) turns out to be correct.

CHAPTER 3

Extant transparency theories – A critique

In the preceding chapter, I argued that we have direct access to our own judgments in conscious first-order thinking. I now want to turn to the question of the relation between conscious first-order thinking and self-ascriptions and knowledge of beliefs.

It is a common assumption in philosophy that if one has a concept of belief and a self-concept, then to find out whether one believes *p*, conscious first-order thinking to settle whether *p* is sufficient (e.g., Evans 1982; Moran 2001; Byrne 2005, 2011; Fernández 2013; Cassam 2014). The idea is that if, e.g., I want to determine whether I believe *p*, I don't need to attend to and represent some aspect of my mind. I just need to attend to the world and those features of it that allow me to settle whether *p*. If upon reflection or swiftly recalling that *p*, I conclude that *p*, then I may, on the basis of the outcome of my query, judge that I believe *p*. Above, I called this the *transparency method* (TM). Accounts of self-knowledge of beliefs that invoke TM to explain the acquisition of self-knowledge of beliefs are what I called *transparency theories*.

In this chapter, my aim is to assess a wide range of currently available transparency theories to see whether they offer a satisfactory explanation as to how self-knowledge of beliefs is acquired. I shall show that this is not the case.

In the discussion, I will put aside the more specific question of whether the accounts are able to explain the privileged nature of self-knowledge of beliefs. I won't take the issue up until the middle of the next chapter. What concerns me here and until then is the question of whether extant transparency theories manage to satisfactorily explain the acquisition of self-knowledge of beliefs more generally.

In section 1, I introduce what I take to be the two puzzles that TM poses for anyone who wishes to explain the acquisition of self-knowledge of beliefs by appeal to the method. The discussion of these puzzles will lead me to the formulation of a set of conditions that any adequate transparency theory should meet. In section 2, I mention the two different approaches to TM that have been adopted, before, in sections 3 and 4, showing that the extant theories that belong to the two approaches don't manage to meet the set of adequacy conditions introduced in section 1.

1. Two puzzles

TM poses two central explanatory problems, what I shall call the *intelligibility puzzle* and the *knowledge puzzle*.

The intelligibility puzzle

TM suggests that one can find out whether one believes *p* by determining whether *p*. Many philosophers have noted that this leads to a problem, for *prima facie* the belief-related question has little to do with the world-related question (Martin 1998; Moran 2001; Byrne 2011; Boyle 2011; Roessler 2013; Barnett 2015). For instance, *p* might be the case[21] even if one doesn't believe *p*. After all, one isn't omniscient. Indeed, given the huge number of facts, what is the case will vastly outstrip what one believes. There is thus not even a reliable correlation between what is the case and what one believes. Conversely, one might believe *p* even if not *p*. I might believe that bats are birds even if they aren't. Given these points, why "should the evidence that the subject has about how the world is have any bearing on what beliefs a particular person has?" (Martin 1998: 110)

The problem here can be pressed further. For normal adult subjects with the concept of belief will arguably understand the points just mentioned. They will understand that given the vast number of facts, it doesn't follow that when *p*, then they believe *p*. They will acknowledge that it is in fact more likely that when *p*, then they do *not* believe *p*. Similarly, they will understand that they might have various beliefs that don't correspond to what is the case. Thus, they will grant that whether or not *p*, this is no indication as to whether or not they believe *p*. Since that is so, how can a normal adult subject S, who understands that whether or not *p*, this won't tell her whether or not she believes *p*, nonetheless proceed to determine whether *p* to find out whether she believes *p*? This is what I call the *intelligibility puzzle*. The puzzle pertains to the issue of how it can make sense to S to deliberately start determining whether *p* when she wants to know whether she believes *p* even though she understands that whether or not *p*, this won't tell her whether or not she believes *p*.

Theorists, both internalists and externalists alike (see below), who wish to hold that TM is a procedure for acquiring self-knowledge of beliefs need to offer an answer because

[21] I use the expressions '*p* is the case', '*p*', and '*p* is true' interchangeably to refer to *p* only.

otherwise their opponents might propose the following objection.

> TM is *not* a procedure for acquiring self-knowledge of beliefs. For any subject S who has the concept of belief, which is a prerequisite for forming self-ascriptions of belief, will understand that given the vast number of facts, it is much more likely that when *p*, then she does *not* believe *p*, because there are clearly a great many more facts than she has beliefs. This will prevent her from proceeding from *p* to the self-ascription *I believe p*. It will prevent her from using TM to work out her own beliefs.
>
> Granted, if asked 'Do you believe *p*?' (or 'Do I believe *p*?'), S might answer the question in the same way as the outward-directed probe as to whether *p*, and that might motivate the assumption that TM is a method for working out one's own beliefs. However, this phenomenon doesn't support the assumption, for it might simply be due to the fact that S hears the question 'Do you believe *p*?' (or 'Do I believe *p*?') as a way of asking the first-order question 'Is *p* the case?'

Unless the intelligibility puzzle is solved, i.e., unless it is explained how it can make sense to S to determine whether *p* in order to find out whether she believes *p* even though she acknowledges that whether or not *p*, this won't tell her whether she believes *p*, opponents of TM can use the just-mentioned argument to insist that TM is not in fact a method to work out one's own beliefs at all.

Advocates of TM might propose different solutions to the intelligibility puzzle. However, in their suggested solutions, they should not presuppose that the subject using TM is already able to represent her own mental processes or states without explaining how this meta-representation comes to be. Otherwise, their proposal would be at odds with TM because the method, as it is typically understood, suggests that in order to determine whether one believes *p*, one only needs to represent worldly states of affairs pertaining to whether *p*. Furthermore, their proposal would be incomplete at best and already assume what it set out to explain at worst.

The preceding points can be summarised in the following first adequacy condition:

> (AC1) Any adequate account of self-knowledge of beliefs on which TM provides one with self-knowledge needs to solve the intelligibility puzzle, i.e. it needs to explain how a subject S can come to determine whether *p* in order to determine whether she believes *p* even though she understands that whether or not *p*, this

won't tell her whether she believes *p*. Furthermore, the account shouldn't presuppose that S is already able to represent her own mental processes or states without explaining how she is able to do so.

The knowledge puzzle

Suppose that the intelligibility puzzle has been solved. There is then a second issue that needs to be dealt with if TM is to be a method for acquiring self-knowledge of beliefs. It pertains specifically to the question of how TM can lead to *knowledge*.

There are different views on what makes a judgment or belief that *p* knowledge (see Ichikawa and Steup 2012). As I already said in Chapter 1, in this thesis, I assume that knowledge is justified and true judgment or belief. I shall not worry about the well-known problems with this view (see, e.g., Gettier 1963) but will instead only focus on justification and truth.

Given this, since in the context of TM the transition from *p* to the second-order judgment *I believe p* is "neither deductively valid nor inductively strong" (Byrne 2011: 204), how can the second-order belief be justified and true? This is what I shall call the *knowledge puzzle*.

It is distinct from the intelligibility puzzle. For one might propose an account that shows that in the context of TM S's second-order judgment *I believe p* is justified and true even when she herself doesn't have a grasp of the intelligibility of the transition that she engages in when forming the second-order judgment. That is, a solution to the knowledge puzzle doesn't necessarily amount to a solution to the intelligibility puzzle, for one might endorse, e.g., an externalist account of epistemic justification. Before saying more on the different accounts of justification, it will be good to summarise the point just made in the following second adequacy condition:

(AC2) Any adequate account of self-knowledge of beliefs on which TM provides one with self-knowledge needs to solve the knowledge puzzle, i.e. it needs to explain how the self-ascription *I believe p* that is thought to result from applications of TM can be justified and true.

2. Two approaches

The two just introduced adequacy conditions will become useful below when assessing the currently available transparency theories and developing an alternative. Before considering extant transparency accounts, it will help structure the following discussion to distinguish between the two different approaches that have been adopted to explain TM and to deal with the puzzles just mentioned. There is what I shall call an *epistemic* approach and a *non-epistemic* approach to TM.

The epistemic approach

On the epistemic approach, it is assumed that TM is a procedure for acquiring self-knowledge of beliefs that involves the detection of a pre-existing condition of oneself, i.e. of one's believing or not believing *p* via evidence pertaining to the obtaining of that condition. Furthermore, the thought is that the judgment that self-ascribes the belief *p* is causally linked to the self-ascribed belief *p*. The theories that belong to the epistemic approach are then intended to explain how the link at issue can ensure that the self-ascription qualifies as knowledge.

As noted, to count as knowledge, the self-ascription *I believe p* that is thought to result from an application of TM needs to be justified. There are two kinds of theories on epistemic justification: internalism and externalism.[22] According to internalism, for S's belief *p* to be justified, S must be aware of or at least immediately capable of being aware of the justifier(s) of the belief, i.e. the fact(s) according to which she is justified in having the belief. In contrast, according to externalism, S may not always be aware of or able to be aware of the fact(s) in virtue of which she is justified in having a belief. For instance, according to reliabilism, which is the most common form of externalism, S is justified in believing *p* iff her belief that *p* is produced by a cognitive process *C* and *C* is reliable, i.e. it tends to produce a high proportion of true beliefs. On this view, for S's belief to be justified, she needn't be able to be aware of *C*, or of whether *C* is reliable. Indeed, she may have no idea as to how her belief was formed. As long as the process that led to the belief is reliable, the belief will be justified.

[22] There are different versions of each, see, e.g., Kornblith (2001). The following characterisation captures the basic difference between them.

If we apply the two views of epistemic justification to the second-order judgment that is assumed to result from an application of TM, then, for the internalist, in order for S's self-ascription *I believe p* to be justified S must be aware of or at least immediately capable of being aware of the justifier(s) of the self-ascription. In contrast, for the externalists, S's self-ascription might still be justified even in cases when she herself isn't or can't be aware of the justifier(s). For instance, as long as the self-ascription *I believe p* is the product of a reliable process, it may still count as justified. To deal with the two puzzles that TM poses both internalist and externalist accounts might be proposed. But one might also adopt an entirely different approach.

The non-epistemic approach

What both kinds of accounts just introduced have in common and what makes them instances of the epistemic approach to TM, is the assumption that TM is a means for acquiring knowledge of a pre-existing condition of oneself, i.e. of one's believing or not believing *p*. Advocates of the *non-epistemic* approach reject this assumption. They insist that the phenomenon that we determine whether *p* in order to answer whether we believe *p* needs to be explained differently. Some advocates of the non-epistemic approach appeal to the metaphysics of beliefs and hold that one does not acquire self-knowledge of beliefs when using TM but merely brings the self-knowledge that one already has and that is constitutive of one's first-order beliefs into consciousness, e.g., via reflection on whether *p* (see, e.g., Boyle 2011; Shoemaker 2012; Valaris 2011, 2014b). Other advocates of the non-epistemic approach hold that in applications of TM, one *does* obtain self-knowledge of one's belief *p*, but not via discovering the pre-existing fact that one believes *p*. Rather, they propose that this happens by committing oneself to the truth of *p* and therewith making it the case that one believes *p* (see, e.g., McGeer 1996, 2008; Moran 2001, 2012).

In the next two sections, I will critically discuss the currently available theories that fall within the epistemic approach, and the non-epistemic approach. As will become clear, each of them faces a problem with respect to meeting the adequacy conditions introduced above. Indeed, they already fail to meet (AC1) and there will be little need to mention (AC2) until the next chapter.

3. The epistemic approach

As noted, the epistemic approach comprises internalist and externalist theories. Before going into the discussion of examples of the two kinds of accounts, a clarification is in order. There is much debate in epistemology on whether internalism or externalism is the correct view of justification and knowledge (see, e.g., Kornblith 2001; Conee and Feldman 2001; Bonjour and Sosa 2003; Goldman 2009). If one has misgivings about internalism or about externalism in general, then one might reject a particular account of TM merely on the basis that it is, e.g., internalist or externalist in nature. In what follow, I shall remain neutral on whether internalism or externalism is true. I prefer to assess the accounts of each kind on their own merits.

3.1 Internalist theories

Internalist theories are proposals that assume that in order for S's self-ascription *I believe p* to be justified, S must be aware of or able to become aware of the justifier(s) of her judgment. Among these theories are what I shall call the *rationality view*, the *consciousness-based view*, and the *mental-action view*.

3.1.1 The rationality view

A number of philosophers hold that the transition involved in TM is explained by facts about rationality. Two different proposals can be distinguished. One appeals to Moore-paradoxical statements such as '*p*, but I don't believe that *p*' (Moore 1942). The other rests on the assumption that rational subjects tend to believe what they ought to believe. Gordon (2007), for instance, can be viewed as adopting the first one. He construes TM in terms of an "ascent routine" (ibid).[23] On his view, in applications of the method, we

[23] Gordon's ascent routine view is in fact only concerned with linguistic expressions such as 'I believe it is raining' that are formed on the basis of first-order expressions 'It is raining'. He writes that once the prefix 'I believe' has been attached to 'it is raining', one's "major remaining need now is to construe the sentence" 'I believe that it is raining' "as a self-ascription" (2007: 162). For ascent routines themselves "cannot make us self-ascribers. They are only a first step in […] a gradual developmental process" (2007: 164). Gordon confesses that his own "suggestions regarding the developmental hurdles beyond ascent routines remain mere conjecture" (ibid). Nonetheless, his consideration on Moore-paradoxical statements might be viewed as capturing one way in which TM can lead to self-ascriptions. That is why I included his proposal here.

> step *up* a semantic level from an assertion that *p* to a self-ascription of a belief that *p*. We allow ourselves to move from […] *expressing* the belief that *p* to *self-ascribing* the belief that *p*. Thus, we may move from an assertion about the weather, 'It's raining,' to an assertion about ourselves, 'I believe it's raining,' from a weather report to a self-report. (2007: 154)

Gordon ties this transition to rationality. He holds that the

> permissibility of the move from asserting that *p* to affirming that one believes that *p* is closely related to the *im*permissibility of asserting that *p* and denying that one believes that *p*. […] [E]ven though statements of the form, '*p*, but I don't believe that *p*' are non-contradictory, we reject them as absurd and, in a way, inconsistent. (2007: 154)

On Gordon's view, reasoning in accord with TM, i.e. affirming that one believes *p* on the basis of *p*, is required for avoiding absurd assertions and thoughts, which is part of being rational.

The problem with this proposal is that from, e.g., my own perspective, when I'm employing TM, the point just mentioned only sanctions a move from *p* to the self-ascription *I believe p* if I already understand that when *p*, then I'm affirming *p* and affirming *p* amounts to believing *p* and hence is at odds with denying that I believe *p*. If I don't already understand this, then there is little basis for me to find it problematic to affirm the proposition *p, but I don't believe p*. Indeed, as a rational subject, I will arguably understand that *p* might be the case even if I don't believe *p*. For my denying this point would amount to my believing that I'm omniscient, which is absurd and so presumably not something a rational subject would believe. Hence, unless it is presupposed that I already somehow take it that when *p*, then I'm affirming *p*, then I won't find my thinking *p, but I don't believe p* absurd or inconsistent. Gordon's proposal relies on the assumption that I already have an insight into my own mind, into my affirming and so believing *p*. He doesn't, however, explain how it is acquired. His view thus doesn't meet (AC1).

Turning now to the second account of TM that appeals to facts about rationality, for instance, Finkelstein (2012) describes TM as follows: "The question of whether I believe that *p* is, for me, transparent to the question of what I ought rationally to believe

– i.e. to the question of whether the reasons require me to believe that *p*. I can answer the former question by answering the latter", where the latter can in turn be answered by determining whether *p*, because if *p*, then I ought to believe *p* (103). Cassam (2014: 121, 142) too grants that at least sometimes I can answer the question whether I believe *p* by answering the question of whether I ought to believe *p*. He writes that as long as my "attitudes are determined by my reasons, and I am also entitled to assume that they are so determined, I can determine what my attitudes are by determining what they rationally ought to be" which, given that this is supposed to be a construal of TM, is to say that I can do so by determining whether *p* (2014: 4). While Cassam goes on to mention various problems with this procedure, he concedes that the latter is nonetheless "one of a range of pathways" to self-knowledge of beliefs (2014: 101).

Smithies and Stoljar (2012) summarise the argument underlying this view of TM in the following way:[24]

> Let us assume that I am rationally entitled to believe (1) that if *p* is true, then I ought to believe that *p*. Now let us add [the] assumption that I am rationally entitled to believe (2) that if I ought to believe that *p*, then I believe that *p*. It follows (by means of a closure step) that I am rationally entitled to believe (3) that if *p* is true, then I believe that *p*. Therefore, I am rationally entitled to answer the question whether I believe that *p* by answering the question whether *p*, which solves the problem of transparency. (2012: 15)

The reasoning here is again supposed to provide the user of TM with an internalist justification for the transition involved in applications of TM.

However, the view that if *p* is true, then I ought to believe *p* is false, for there are infinitely more truths than a subject is able to believe. Further, since any conjunction of true propositions is a true proposition also, there are true propositions that are simply too complex to be believed by finite individuals such as me. Since 'ought' implies 'can', it is absurd that I ought to believe what is true. Given this, I will reject that if *p* is

[24] Smithies and Stoljar (2012) don't themselves advocate the argument but attribute it to Moran (2001, 2012). This rests on a misunderstanding of Moran's proposal on their part, however. For Moran claims quite explicitly that the transition involved in applications of TM is "not an inference" (2012: 233). Yet, it clearly is an inference on Smithies and Stoljar's construal of TM. I discuss Moran's view below.

true, then I ought to believe *p*. If that is so, however, then there is no way for me, on the proposal at hand, to proceed from *p* to the self-ascriptions *I believe p*. Finkelstein's, and Cassam's proposal hence either requires the user of TM to accept an absurd principle (i.e. if *p* is true, I ought to believe *p*) – which is an assumption that is hard to accept. Or it fails to explain how it can be intelligible for the user of TM to move on the basis of *p* to a self-ascription of the belief *p*. That is, it fails to meet (AC1).

3.1.2 The consciousness-based view

Some philosophers argue that when one employs TM and upon reflection on whether *p* concludes that *p* then one's conclusion is a conscious judgment *p*, which is in virtue of its conscious properties known to the subject and usable as a justificatory basis for the formation of the self-ascription *I believe p* (see, e.g., Peacocke 2003; Silins 2012; Smithies 2012). This is the consciousness-based view. I will focus on Peacocke's (2003), and Valaris' (2014a) versions of it.

Peacocke (2003) holds that when in applications of TM, I conclude that *p* then this will be a phenomenally conscious judgment that constitutes my basis for forming the self-ascription *I believe p* and coming to know my belief *p*. To motivate the assumption that conscious judgments are phenomenally conscious or experienced, he offers a set of examples. He writes, for instance, that when I try to recall someone's name and it then occurs to me that that person is called *X* then its so occurring to me contributes to the specification of what it's like for me experientially at that moment. It would be subjectively different for me, Peacocke maintains, if it occurred to me (falsely) that the person was *Y* instead of *X*, and subjectively different again if no name came to my mind. Peacocke holds that the same is true when it suddenly strikes me that *p*, or when upon reasoning on whether *p*, I conclude that *p*. When it strikes me that *p* or when I conclude that *p* then that "can be a partial specification of what it's like" for me at that moment (Peacocke 2003: 83).

Peacocke argues further that just as a pain experience can be a reason for me for judging that I'm in pain, the 'what-it's-like-ness' of a conscious judgment *p* can be a reason for me for judging that I believe *p,* where this judgment amounts to knowledge. For, typically, judgments are either initiations of beliefs or activations of existing beliefs. Hence, self-ascriptions of beliefs formed on their basis will typically be true. And since

the self-ascriptions of beliefs are via the phenomenally conscious judgments also internally justified, they qualify as knowledge (Peacocke 2003: 102).

But the proposal is not without problems. As noted, on Peacocke's account, it is in virtue of its phenomenology that my judgment *p* is my reason for the self-ascription *I believe p*. It isn't clear, however, whether there is such thing as a phenomenology of judgments to begin with. Peacocke provides examples to lend support to his claim but even if we grant that it is like something when it, e.g., suddenly strikes one that *p* or when upon reasoning on whether *p*, one concludes that *p*, why assume that one's experience at these moments is due to one's judging *p* rather than some affective or sensory-imagistic event that co-occurs with it yet might also occur without it?

Let's focus on the case when it strikes one that *p*. Consider dreams. Suppose I'm having a vivid nightmare. I'm dreaming that I'm being attacked by a chainsaw-wielding killer, call him 'Leatherface' (from the movie *The Texas Chainsaw Massacre*), and am frantically trying to escape. There is no reason to deny that when it strikes me in my dream that Leatherface is attacking me that the phenomenology is very much the same as when I'm awake and it strikes me that a dangerous lunatic with a chainsaw is attacking me. Indeed, in psychological studies on dreams, the physiological responses and affective involvement during intense, action-filled, or violent dreams, e.g., of "being chased or attacked by unfamiliar people, animals or insects" (Schenck and Mahowald 2002: 124), suggest that there "is nothing in the experience itself, in the actual qualitative character of the experience, that necessarily distinguishes the dream experience from a corresponding perceptual experience in the waking state" (Revonsuo 2006: 82). This view is also supported by the fact that I might at some point, when Leatherface is (in my dream) catching up with me and about to cut me into pieces, wake up covered in sweat and feel relieved that what I experienced was just a dream.[25]

[25] There is neuroscientific support for this view too. For instance, LaBerge (1990) writes that "serotonergic neurons" that are part of the "system that normally inhibits vivid images (hallucinations)" is "itself inhibited in REM sleep, allowing dreamed perceptions (i.e. images) to appear as vividly real as perceptions. In REM, also, sensory input is actively suppressed preventing competition from perceptual processes. Perhaps this explains in part why we are so inclined to mistake our dreams for reality: To the functional systems of neuronal activity that construct our experiential world (model), dreaming of perceiving or doing something is equivalent to actually perceiving or doing it. (126)

Now, it seems clear that when I'm *awake* and it strikes me that Leatherface is attacking me then I judge that he is attacking me. For this event will immediately lead to a belief that I'm in danger, and this belief in conjunction with my standing desire not to be harmed will cause me to be highly alert/awake to the situation. I will immediately form the decision to flee and, provided I'm not paralysed or by external factors prevented to move, swiftly implement the decision. I will flee. The causal effects of the event of it striking me that I'm being attacked suggest that I'm judging that I'm being attacked.
Does the same hold in the case when things strike me this way in my *dream*? In the dream-case, I will not become highly alert/awake. When it strikes me that Leatherface is attacking me, by assumption, my dream will continue. It will continue with me trying to escape from the danger. I will stay asleep and won't actually move away from where I am at that moment. It seems clear, then, that when it strikes me in my dream that Leatherface is attacking me, with this event, I'm not in fact judging that I'm being attacked For if I were, my remaining asleep would, given my desire not to be harmed and the fact that I'm not paralysed or in any other way prevented from moving, become hard to explain. There is thus reason to believe that the dream-event of it striking me that Leatherface is attacking me doesn't play the right functional role (i.e. it doesn't enter into the right relations to other mental states/events and action-guidance) to count as a judgment that I'm being attacked.

But recall that we also have strong reason to believe that the phenomenology is the same in both the dream-case and the awake-case. For, as noted, the physiological responses and affective involvement during dreams of the kind at issue indicate that there is little in the "experience itself, in the actual qualitative character of the experience, that necessarily distinguishes the dream experience from a corresponding perceptual experience in the waking state" (Revonsuo 2006: 82). Indeed, unless we assume that in dreams too we can experience what Kriegel calls the "most fundamental aspect of the experience of consciously judging that *p*", namely a "sense of committing to the truth of *p*" (2016: 32), it becomes difficult to account for the physiological responses to and affective involvement in dreams such as nightmares – e.g., the relief, upon suddenly waking up in terror, that what one experienced wasn't real but only a dream.

The best explanation of the bodily/emotional effects of vivid dreams such as the one under consideration is arguably that in these dreams, subjects experience *p* to be true

and their "sense of committing to the truth of *p*" (ibid) is the same as when they are awake and commit to the truth of *p*. It is just that the event that is then experienced isn't connected in the right way to various other mental states/events (e.g., desires, decisions etc.) and action control in order to have the paradigmatic effects of a judgment and so to qualify as a judgment itself. If this is right, then since the phenomenology in the example is the same in both the dream- and the awake-scenarios yet only in one case a judgment is present, it follows that the mental event that is experienced (i.e. it's striking one that *p*) can't, pace Peacocke, itself be a judgment.

Similar arguments can be proposed to deal with the other examples that Peacocke appeals to, e.g., cases when one recalls a name, or concludes that *p*, for phenomenally similar if not identical events can arguably also occur while one is dreaming. Vivid dreams hence help illustrate that the experiences that Peacocke refers to, and that according to Kriegel involve a "sense of committing to the truth of *p*" (2016: 32), aren't themselves judgments but at best accompany these attitudes and might occur without them. Peacocke's support for the claim that there is a phenomenology of judgments is thus undermined.

Independently of the point just made it is worth noting that despite the fact that there has recently been a flurry in work on "cognitive phenomenology", which is typically taken to be a distinctive, proprietary phenomenology of thought (Bayne and Montague 2011), the existence of such phenomenology is still controversial. The fact that this is so is itself telling. For if, as Peacocke and other advocates of the consciousness-based view (e.g., Silins 2012; Smithies 2012) suggest, conscious judgments had just like conscious perceptual states and sensations, e.g., pain experiences, a distinctive phenomenology that could serve a subject as a reason for forming self-ascriptions of mental states, then perhaps the existence of cognitive phenomenology should be as uncontroversial as the existence of the phenomenology of pain states. But this isn't so (see, e.g., Nichols and Stich 2003; Wilson 2003; Goldman 2006; Prinz 2012).

In response it might be suggested that perhaps conscious judgments are phenomenally *transparent*. Valaris (2014a), for instance, holds that conscious judgments are phenomenally transparent in that in consciously judging that the world is a certain way, one is only subjectively aware *that* the world is a certain way: the way specified by the content of one's judgment. One needn't have any awareness of one's judgment or belief

as such. Further, for Valaris, being aware of "something as a *fact about the world* can make a subjective phenomenal difference, or a difference to 'what it is like' for one, even if that state of awareness does not instantiate any qualitative properties that one is aware of" (2014a: 2). Valaris claims that this minimal kind of phenomenology is enough to ground an account of self-knowledge of beliefs. For when the subject is "rational and in possession of the concept of belief", she "must grasp the distinction between her own take on the facts and the facts themselves: she must know that her taking the world to be a certain way is a different matter from the world's really being that way" (Valaris 2014a: 7). This piece of knowledge, Valaris thinks, enables the subject to "*step back* from her awareness that *p*, and self-ascribe the belief that *p*" (2014a: 8). On this view, in order to "self-ascribe the belief that *p*, one simply has to consider whether *p* – whereupon it will occur to one that *p*, and one will be in a position to self-ascribe the belief *via* standing back" (ibid). Valaris claims that the transition involved in "stepping back from *p*" is not a case of "forming a belief *on the grounds that p*" but "a *sui generis*, and yet intuitively rational, transition", where the "rationality of this transition is explained in terms of both the phenomenology of occurrent conscious beliefs and the general knowledge that goes together with possession of the concept of belief" (2014a: 10).

The problem with Valaris' view is that when, e.g., you, upon finding that *p*, "self-ascribe a belief that *p* via *stepping back* from, or *bracketing*, your commitment to *p*," then this "involves recognising that *p* is part of your subjective take on the facts, as distinct from the facts themselves" (2014a: 9). Valaris overlooks that this recognition is in fact precisely what needs to be explained. He assumes that it belongs to having the concept of belief that one is able to recognise that *p* is part of one's subjective take on the facts (rather than just the fact *p*) without explaining how this recognition is possible in the first place. His view hence fails to meet (AC1).

It seems fair to say, then, that the consciousness-based view in general doesn't adequately explain how one can come to self-ascribe a belief *p* on the basis of the phenomenology of a conscious judgment *p*. I will now turn to the third kind of internalist account of TM that I mentioned above.

3.1.3 The mental-action view

Many philosophers argue that when we perform bodily actions, we tend to be aware of those actions. Some theorists claim that the same holds for mental actions and maintain that our awareness of mental actions can, e.g., in applications of TM, be the basis for self-knowledge of judgments and beliefs. This is the mental-action view. I will briefly discuss Peacocke's (2008), Soteriou's (2013), and Roessler's (2013, 2015) versions of it.

Peacocke (2008) holds that when we perform a bodily action, we have knowledge of both the movements (e.g., 'My leg is going up') and the intentions underlying them ('I am lifting my leg'), where this knowledge is distinctive in that it doesn't require any observation of our own movements. For instance, I can know of my movements even when the relevant part of my body is anaesthetised: with my jaw anaesthetised for root canal treatment, I can still know, without needing to see myself in the mirror, that my mouth is open and I have just opened my mouth. Similarly, a patient whose re-afferent nerves in the arm were destroyed and who lost all feeling in the arm can still know, without having to observe herself, that her arm is going up and that she is raising it (Peacocke 2008: 247). When engaged in these bodily actions, subjects have self-knowledge of them, Peacocke claims.

He argues that the same holds for mental actions. For instance, mental actions such as "deciding, judging", and "reasoning" involve a "mental action-awareness", Peacocke thinks (2008: 250, 255). Furthermore, he claims that the

> distinctive way in which a subject comes to know of his own mental actions is by taking an apparent action-awareness at face value. You judge that it will rain. When so judging, you have an apparent action-awareness of your judging that it will rain. By taking this awareness at face value, you come to know that you judge that it will rain. (Peacocke 2008: 259)

To explain the mechanism producing action-awareness, Peacocke appeals to the "comparator model" account of motor control (Frith et al. 2000). According to the account, when, prior to a bodily action, the brain activates a motor schema and sends instructions to the muscles to initiate the movement, it simultaneously forms a

representation, an *efference copy*, of those instructions that is then passed on to an emulator system (or several) that contains a model of the body's movement capacities. This system turns the efference copy into a predictive model of the action, which is a representation of the somatosensory and other perceptual feedback that is expected if the action happens as planned. A further system, a *comparator*, then receives this model and compares it with incoming sensory feedback. Based on a match or mismatch, the system can then cause online corrections of the action if need be.[26]

Peacocke holds that it is via the efference copy or (as he calls it) the *corollary discharge* produced shortly before the bodily movement that one is aware of one's bodily action, e.g., that one's arm is going up and that one is raising it. He claims that this "applies quite generally, both in bodily and in mental cases": "awareness of mental actions is action-awareness of the same sort as occurs in bodily action awareness" and can be accounted for in the same "explanatory structure" (ibid).

The problem is that efference copies are representations of outgoing *motor* commands (Blakemore et al. 2002; Tsakiris and Frith 2009) but when a subject S is *judging*, say, that mosquitos carry plasmodia, it is hard to see what the efference copy of this mental action could possibly be representing. For there is arguably no *motor* program for the mental act of judging that mosquitos carry plasmodia, and there are no movement instructions passed on to S's muscles when she engages in the mental action at issue (Carruthers 2009b). It hence remains unclear how one can on Peacocke's account become action-aware of one's own judging, let alone come to know one's own beliefs.

Soteriou (2013) offers a different mental-action account. He starts with a consideration on decisions and one's awareness of bodily actions. When one decides to φ, he holds, one commits oneself to φ-ing. One's being so committed involves one's being disposed to remember to do certain things. One of these things is to impose on subsequent planning and reasoning the constraints of representing and treating as true the proposition that one will φ. Further, when the time arrives to φ, one's decision and commitment to φ will evolve into an intention-in-action with which one commits oneself both to now φ-ing and to accepting the proposition that one is now φ-ing. That is, the intention-in-action is accompanied by a belief about what one is presently doing. Soteriou argues that since there is a reliable concurrence of one's φ-ing and one's belief

[26] See Synofzik et al. (2013) for details and a recent update on the comparator model account.

about what one is presently doing, one's belief that one is φ-ing "embodies a form of knowledge of the action it concerns": it is "practical self-knowledge" of that action (Soteriou 2013: 313).

Soteriou then applies this line of thought to mental actions such as imagining *p*, reflecting on whether *p*, recollecting *p*, doing mental arithmetic etc. He holds that these are mental actions accompanied by intentions-in-action too. Since intentions-in-action give rise to practical self-knowledge of what one is doing, one also knows via these intentions what one is mentally doing when one is engaging in these mental actions, Soteriou claims (2013: 321). He argues further that the practical self-knowledge that one has of one's own mental actions can be used to explain how we find out about what we believe: "beliefs about what one is doing when one is consciously thinking ground one's beliefs about what one knows", and this serves as a "way of acquiring beliefs about what one believes" (Soteriou 2013: 354). The thought is the following.

> Suppose that one is engaged in the conscious activity of attempting to work out whether *p*, and suppose that one ends up concluding that *p*. One believes that what one set out to do was to work out whether *p*. If one believes that one has done what one set out to do, then one will believe that in concluding that *p* one has worked out that *p*. If one believes that one has worked out whether *p*, then one believes that one knows whether *p*. […] And on the assumption that the obtaining of the state of knowing that *p* entails the obtaining of the state of belief that *p*, this makes plausible the claim that in acquiring this belief about what has just happened, one has acquired the belief that one believes that *p*. (Soteriou 2013: 352)

Soteriou relates this to TM and claims that the line of thought he proposes

> should alleviate any worries that might be had about how a proposition about the world that one accepts, can justify a proposition about one's mind that one accepts – i.e. the worry about how the proposition *p* about the world that one accepts can justify the proposition that one believes that *p*. It is not that the proposition *p*, which one judges, justifies the proposition *I believe that p*, which one then subsequently judges. It is, rather, that when one concludes that *p*, one has beliefs about what one set out to do, what one did, and how one did it. One believes that one has worked out whether *p* and one has beliefs about how one did this – which

> amounts to having the belief that one knows that *p* and having beliefs about how one knows that *p*. (Soteriou 2013: 354)

One problem with Soteriou's proposal is that even if we grant that one knows that one set out to determine whether *p* because one has formed an intention-in-action and has hence practical knowledge of what one is doing, it isn't clear on his proposal how one could come to know that one has *done* what one set out to do. Perhaps one has practical self-knowledge of one's stopping to work out whether *p* because working out whether *p* and stopping to do so are plausibly viewed as mental actions accompanied by intentions-in-actions. Still, this knowledge won't allow one to infer that one has *concluded* one's reflection and formed a judgment on whether *p*, because there are many different reasons why one might intentionally stop one's reflection on whether *p*. That one stops when one has worked out whether *p* is just one of them. One might, for example, also do so when one is distracted, undecided on whether *p*, loses interest, etc.[27] One's practical knowledge of one's stopping to reflect on whether *p* will hence be insufficient to find out whether one has judged *p* or not. In order to come to know one's own belief *p* in the way Soteriou suggests, one would need to know that one has worked out that *p*, i.e. one would need to know one's judgment *p*. But judgments can't be intended and hence can't be accompanied by an intention-in-action, as Soteriou emphasises himself (see, e.g., 2005: 91, 95-96; 2013: 248, 341). So they can't be accompanied by practical self-knowledge. As a result, one's practical self-knowledge of what one is doing won't suffice to find out about one's belief *p*.[28] Soteriou's proposal thus doesn't manage to explain the transition involved in TM. It too fails to meet (AC1).

There is a third kind of mental-action account that seems to offer a way of solving the problem that Soteriou's proposal faces. It appeals to the linguistic expressions of one's judgment *p* and has been put forward by Roessler (2013, 2015). While philosophers working on TM typically take it that in the context of TM judging *p* provides one with an epistemic basis for a self-ascription of the belief *p*, Roessler argues for a "modest"

[27] This might happen as frequently and reliably as one's stopping one's reflection as a result of a judgment. So even if one is externalist, some argument is needed for assuming that there is a reliable connection between one's stopping one's reflection and one's judgment.

[28] In some place, Soteriou claims that "self-knowledge" of judgment is built into the ontology of conscious judgments: they are "necessarily accompanied by a distinctive form of self-knowledge" (2013: 340-341, 343). Unfortunately, he doesn't say how this knowledge is acquired in the first place.

interpretation of TM. He holds that S

> knows that she believes that *p* not on the basis of judging that *p* but *in* judging that *p*. […] In doing something intentionally, one generally knows what one is doing, under descriptions corresponding to the contents of at least some of one's intentions. Now sincerely asserting that *p* is an act informed (partly) by the intention to express one's belief that *p*. Hence, the subject will typically be aware of expressing her belief that *p*. (2013: 3)

Roessler argues that this point can be applied to cases of judgments that aren't overtly expressed:

> Judging that *p*, it is plausible to suppose, involves saying that *p* in 'inner speech' (if not in outer speech). […] [S]aying that *p* in 'inner speech' is often informed by the intention to express one's view that *p*, even if only for one's own benefit. Quite generally, then, judging that *p* involves being aware of expressing one's conviction that *p*. (2013: 3)

Roessler claims that in the context of TM "answering the question whether *p*" may then "be a 'way of gaining knowledge' of one's belief that *p*" because "answering the question whether *p* is a process that, if one does believe that *p*, and is not interrupted, can be expected to issue in one's judging that *p*; and in doing so, one will be aware of expressing one's view that *p*, hence aware of believing that *p*" (Roessler 2013: 4).

However, Roessler's proposal that a "basic way to know one's beliefs is provided by knowledge of what one is doing in expressing a belief" (2015: 157) fails to explain what needs explaining. For suppose we grant that in doing something intentionally, one knows what one is doing. Let's also grant that judging *p* involves saying that *p*, and saying that *p* is something one does intentionally. If that is right, then I might, when judging and saying that *p*, in virtue of my intention to say that *p*, know that I'm saying that *p*. However, when I know that I'm saying that *p*, I don't yet know that I have the "conviction that *p*", "belief", or "view that *p*" (Roessler 2013: 3). For I might say that *p* even if I'm, e.g., unconvinced that *p* or doubt that *p*. So even if, as part of my judgment *p*, I'm saying that *p*, and if my saying that *p* is intentionally done and therewith known to me, I still don't know my belief that *p* via my saying and judging that *p*.

Suppose now that, as part of my judging *p*, I'm again saying that *p*, but now my saying

that *p* is done with the "intention to express my belief that *p*" and not merely with the intention to express *p* (ibid). Clearly, to form the intention *to express my belief* that *p* in the first place, I will already need to know that I believe *p*. For otherwise I can't intend to express specifically my *belief* rather than doubt, hope, desire etc. *p*.

But then how do I know that I *believe* that *p*, which is required for forming the intention at issue? This is of course the question with which we started. Roessler's proposal doesn't offer an answer.[29] It presupposes that one can somehow self-ascribe the belief *p* without explaining how this works. It thus doesn't manage to meet (AC1).

To sum up the discussion so far, all the accounts that I have considered are internalist in nature. They assume that in the context of TM in order for one's self-ascription *I believe p* to be justified, one needs to be aware of (or with)[30] its justificatory basis. The discussion showed that the accounts remain unsatisfactory. I will now consider the second kind of epistemic theories of TM that I introduced above.

3.2 Externalist theories

Externalist theories are proposals that hold that one's self-ascription *I believe p* might be justified even if one isn't aware of or capable of being aware of the fact(s) in virtue of which one is justified in forming the self-ascription. Two kinds of externalist theories have been proposed, what I shall call the *world-mind inference view* and the *bypass view*.

3.2.1 The world-mind inference view

Byrne (2005, 2011) holds that in the context of TM when upon examining the evidence on whether *p*, I conclude that *p*, then the "next step involves an *inference from world to mind*: I infer that I believe that [*p*] […] from the single premise that [*p*]" (Byrne 2011: 203). On Byrne's proposal, i.e. the world-mind inference view, one forms self-ascriptions of beliefs by following the rule: "BEL: If *p*, believe that you believe that *p*" (Byrne 2005: 95).

[29] To be fair, he takes his proposal only to be one way in which one might specify the suggestion that TM involves coming to know one's own beliefs *in* judging rather than on the *basis* of a judgment. His main point is to introduce this alternative suggestion, not so much to commit to a particular specification of it.

[30] See Roessler's account.

BEL looks like a bad rule for arriving at justified true beliefs about one's own beliefs and hence at self-knowledge. For, as Byrne notes himself, that "*p* is the case doesn't even make it *likely* that one believes that it is the case" (2005: 95).
However, Byrne emphasises that BEL has a significant epistemic virtue: it is "self-verifying" in that if it is followed, the resulting self-ascription will be true. The self-ascription will be true because in order to be said to "follow" BEL at all, one must first "recognize that *p*" and "recognizing that *p* is *inter alia* coming to believe that *p*" (Byrne 2005: 96). Thus, upon following BEL, the resulting belief will be true, Byrne claims.

Indeed, he holds that BEL yields true beliefs even if one only "tries" to follow it, where "S *tries* to follow rule R iff S *believes* that *p* because S believes that conditions C obtain" (Byrne 2005: 97). S's following R entails that she tries to follow R but the converse doesn't hold. For instance, one is trying to follow BEL when one investigates whether *p*, mistakenly concludes that *p*, and then self-ascribes the belief that *p*. One's belief that one believes that *p* will then still be true.

Since that is so, BEL is "strongly self-verifying", and turns out to be a good rule for acquiring self-knowledge of beliefs after all (ibid). For the judgments it produces are clearly "safe", i.e. they couldn't easily be false and as a result plausibly qualify as knowledge, Byrne concludes (2005: 96-98).

His account provides an attractive externalist explanation of how the self-ascription *I believe p* that results from applications of TM can be knowledge: the transition from *p* to *I believe p* will, in virtue of its self-verification aspect, lead to a justified and reliably true self-ascription no matter whether one is aware of or able to become aware of the justifier of the self-ascription.

However, one of the problems with Byrne's proposal is that S's "recognition" or "conclusion" that *p*, which is a judgment, might in some cases fail to lead to the belief *p*. As noted above, a judgment *p* doesn't always result in a belief *p* because belief-formation is also affected by non-doxastic factors, e.g., biases, fear, etc. And these factors might prevent the judgment from having the causal effects on thoughts and behaviour that are required for it to lead to a belief. Since judgments don't necessarily result in beliefs, Byrne's claim that in general following BEL leads to a self-verifying self-ascription of a belief is too strong.

A more serious issue with his view is that it leaves the intelligibility puzzle unsolved and so fails to meet (AC1). The problem was that of explaining how it can be intelligible to S to deliberately determine whether *p* when she wants to find out whether she believes *p* even though she understands that whether or not *p*, this won't tell her whether she believes *p*.

The advocate of the world-mind inference view might suggest that S starts determining whether *p* to find out what she believes because she understands that using BEL leads to a self-verifying belief.

However, suppose that in the context of TM, S concludes that *p*. By assumption, she is then only aware of the worldly fact *p*. How is BEL supposed to make it intelligible to her to form the self-ascription *I believe p* on the basis of what for her is just the worldly fact *p*? Her knowledge that BEL is self-verifying won't help, for it requires her to view *p* as her judgment. Yet, she doesn't at that moment already have the insight that *p* is her judgment. If she did, then a self-ascription of the judgment *p* would be presupposed that the account doesn't explain. The account would fail to meet (AC1).

If S doesn't already have this insight, however, then the point that BEL is self-verifying won't provide her with a basis for forming the self-ascription *I believe p*. The transition will remain unintelligible to her and it is not clear why she would start determining whether *p* to find out whether she believes *p* to begin with. The account will again fail to meet (AC1), as the intelligibility puzzle will now remain unresolved.

It might be worth reiterating here a point made earlier, namely that unless the intelligibility puzzle is solved, it can plausibly be denied that TM is a method for finding out about one's own beliefs at all. For it could then be argued that any subject S who is able to self-ascribe beliefs will understand that whether or not *p*, this won't tell her whether she believes *p*, which will prevent her from determining whether *p* when she wants to find out whether she believes *p* to begin with. It will prevent her from ever using TM. It is not sufficient for advocates of TM to rest on the intuitive plausibility of the point that when we are asked whether we believe *p*, we typically proceed to answering whether *p*. For this phenomenon might only be due to the fact that we understand the belief-related question as a world-related one to start with. Thus, if one wishes to defend the view that TM is a method for acquiring self-knowledge of beliefs,

then no matter whether one is internalist or externalist with respect to epistemic justification, a solution to the intelligibility puzzle is required. Since the world-mind inference view doesn't provide one, it remains unsatisfactory.

3.2.2 The bypass view

There is a second, often-cited transparency account that is in part externalist in nature. It has been developed by Fernández (2013). He argues that if one has a particular belief then there is typically some other mental state on the basis of which one has formed the belief. Fernández calls such a state the "ground" for one's belief (2013: 42). For instance, whenever I seem to see a tomato on the table then typically I come to believe that there is a tomato on the table. My visual state is the ground for my belief. Fernández holds that since the visual state that is the ground for my belief that there is a tomato on the table normally gives rise to the belief, I can also form the self-ascription of the belief that there is a tomato on the table on the basis of the state. For my self-ascriptions of the belief will then typically be true, and so qualify as self-knowledge. One can thus know one's own belief *p* by *bypassing* the belief itself.

Underlying Fernández's "bypass view" is a particular account of epistemic justification that combines both internalist and externalist components (2013: 42).[31] More specifically, Fernández holds that S's belief *p* is justified only if she (1) "forms the belief on the basis of a state" (e.g., a perceptual experience, memory, or another belief) that (2) constitutes "adequate support" for the belief (ibid).

Component (1) is internalist, and component (2) externalist in nature. As for (1), in order for S's belief *p* to count as formed on the basis of some mental state E, S needs to be "disposed to believe that she is in E", i.e. upon reflection, she must be "able to arrive at the belief that [she] is in the state" (Fernández 2013: 58). Fernández holds that this doesn't "require that [she] actually believes that [she] is in the state", it is enough that she "experiences being in the state […], and accepts the content of that experience" (ibid). As for (2), Fernández proposes that for the state E to constitute adequate support

[31] Fernández's view could have been counted to the internalist views that I discussed above. However, since it also includes an externalist element, I mention it here.

for S's belief, E needs to correlate, in S, with the state of affairs that makes her belief true. She needn't be aware of the correlation, however.

Returning with this to the above example, suppose I seem to see a tomato on the table and on the basis of my perceptual experience form the self-ascription *I believe there is a tomato on the table*. Since in normal circumstances, when I seem to see a tomato on the table, I will come to believe that there is a tomato on the table, there is a reliable correlation between the visual state and my belief that there is a tomato on the table. My self-ascription of the belief thus meets condition (2) of Fernández's account of epistemic justification.

Furthermore, since upon reflection on the basis for my self-ascription, I'm also "able to arrive at the belief that [I'm] in" the visual state, for I "experience being in the state", condition (1) of the account is met too (Fernández 2013: 58). Thus, on Fernández's view, my self-ascription of the belief that there is a tomato on the table counts as self-knowledge.

Fernández claims that the same line of thought applies when the grounds for *p* that I base my self-ascription of the belief *p* on are not perceptual states but states involved in, e.g., memory, testimony, and reasoning. For, under normal conditions, the following production-of-belief principle (PB) holds.

(PB) For any propositions *p*, *q* and any subjects S, S*:

(i) If S apparently perceives that *p*, then S comes to believe that *p*.
(ii) If S apparently remembers that *p*, then S comes to believe that *p*.
(iii) If S believes that S* is providing her with the information that *p*, then S comes to believe that *p*.
(iv) If S believes that *q* and S believes that *p* follows from *q*, then S comes to believe that *p*. (Fernández 2013: 47)

Hence, on the basis of perception, memory, testimony, and reasoning, I can come to know my own beliefs via bypassing the beliefs themselves.

Fernández holds that the bypass view is modelled on and "strongly supported" by the idea that if, e.g., I wonder whether I believe *p*, my attention will normally be directed at

the reasons for *p*, and not at my own mind (2013: 55). He claims that the bypass view explains this phenomenon, i.e. TM. Because if the support that I have for believing *I believe p* is identical to my grounds for the belief *p*, then it is to be expected that in order to find out whether I believe *p*, I will attend to the world and inquire whether *p*. For this will bring out my grounds for the belief *p*, and therewith provide me with a basis for the self-ascription of the belief *p* (Fernández 2013: 49f).

However, Fernández's view is problematic too. Just like the world-mind inference account, it doesn't manage to solve the intelligibility problem. As mentioned, the problem is that of explaining how subjects can come to determine whether *p* in order to settle whether they believe *p* even though they understand that whether or not *p*, this won't tell them whether they believe *p*.

It might be suggested that on the bypass view, there is no problem here. For according to this account, subjects ask whether *p* to find out whether they believe *p* because they understand that the grounds that they have for believing *p* are identical to their grounds for the belief *I believe p*. Thus, in order to find out whether they believe *p*, they inquire whether *p* because this will reveal their grounds for or against the belief *p*.

The problem with this is that subjects will only be able to grasp that their grounds for the belief *p* are their grounds for the self-ascription *I believe p*, if they believe that their grounds for the belief *p* typically give rise to the belief *p*. If for them the grounds for the belief *p* are only facts of the world, however, then they won't believe this. For, as they will acknowledge, it is false that whenever *p*, then that fact of the world, i.e. *p* will give rise to their belief *p*. They will only believe that their grounds for the belief *p* will typically give rise to their belief *p* if they view their grounds for the belief *p* as mental states/events. For it is only the mental states/events that figure in the antecedents of (PB) that give rise to the belief *p*. Hence, in order to grasp that the grounds for the belief *p* are the grounds for the self-ascription *I believe p* and therewith to attain a basis for determining whether *p* when they wonder whether they believe *p*, subjects will need to believe that the grounds for their belief *p* are mental states/events, i.e. their "apparently perceiving that *p*", their "apparently remembering that *p*" etc., and not facts of the world (Fernández 2013: 47).

But if this is their background for settling whether *p* when they wonder whether they believe *p*, then upon encountering the grounds that would typically give rise to the

belief *p*, subjects will only be able to self-ascribe the belief *p* on the basis of these grounds if they already view them as their own mental states/events, e.g., their own beliefs (in the cases when the grounds of the belief to be self-ascribed are other beliefs). That is, they will already need to have self-ascribed mental states/events.

If that is so, however, then the solution to the intelligibility problem that the advocate of the bypass view might propose presupposes self-ascriptions of mental states/events including beliefs without explaining how these meta-representations are formed. Thus, the bypass view either doesn't offer a solution to the intelligibility[32] problem, or it presupposes what it needs to explain. It hence fails to meet (AC1).

4. The non-epistemic approach

The accounts of TM that I have discussed so far assume that TM is a procedure for acquiring self-knowledge that involves the detection of a pre-existing condition of oneself, i.e. of one's believing or not believing *p*, via evidence pertaining to the obtaining of that condition. Further, according to the proposals mentioned, the self-ascription of the belief *p* is causally related to the self-ascribed belief *p*, and the proposals are intended to explain how one could proceed from the latter to the former so that the former qualifies as knowledge. The accounts discussed are examples of the epistemic approach. I now turn to the *non-epistemic* approach to TM. Two kinds of theories fall within it, what I shall call the *constitution view*, and the *commitment view*.

4.1 The constitution view

Some philosophers argue that in applications of TM only a "madman" could engage in a transition from *p* to the self-ascription *I believe p*, for a "modicum of rational insight will inform" one that even if *p*, this by itself has no tendency to show that one also believes *p* (Boyle 2011: 230; Valaris 2014b). Since we are not madmen, the argument continues, the transition that is involved in TM can't be explained as a matter of, e.g.,

[32] Gertler (2015) claims that "Fernández addresses this worry [i.e. the point that the transition from *p* to *I believe that p* can't explain self-knowledge since it will not appear reasonable to the thinker] in arguing that beliefs formed through the bypass method will seem reasonable to the subject." Gertler overlooks that beliefs formed in this way will only seem reasonable to the subject, if she is already able to view the grounds for her beliefs as her mental states (e.g., beliefs).

inferring a psychological fact from a fact about the world. Rather, in order to make sense of the transition, we need to assume that the first-order belief *p* is constitutively tied to an unconscious higher-order belief that self-ascribes it.[33] This is the constitution view. It holds that TM is not a means for *acquiring* self-knowledge of beliefs but only for bringing the self-knowledge of beliefs that one already has into consciousness. Among the advocates of the view are Boyle (2011), and Shoemaker (2012).

Boyle (2011) holds that in the context of TM, one moves "from *believing p* to *reflectively judging,* i.e. consciously thinking" to oneself *I believe p*, where this step isn't an "inferential transition between *contents*, but a coming to explicit acknowledgment of a *condition* of which one is already tacitly aware" (227). Boyle maintains that "the very same actualization of my cognitive powers that is my believing *p* is, under another aspect, my tacitly knowing that I believe *p*. Hence, to pass from believing *p* to judging *I believe p*, all I need to do is reflect – i.e. attend to and articulate what I already know" (2011: 229). The process of reflection is then what is thought to happen in applications of TM.

Shoemaker (2012) proposes a similar view. To motivate it, he appeals to a familiar phenomenon. He holds that the transition involved in TM

> should remind us of the point that it seems incoherent for someone to affirm a first-order proposition while refusing to assent to the proposition that he believes it. To do so is something like asserting the Moore-paradoxical sentence '*p*, but I do not believe it.' This suggests that if one affirms *p* one must, on pain of irrationality, be prepared to affirm that one believes that *p*. (2012: 241)

For Shoemaker, unlike for advocates of the rationality view (see above), this shows "that beliefs are constitutively self-intimating" i.e. it is "part of being a rational subject that the belief that *p*, together with the possession of the concept of belief and the concept of oneself, brings with it the belief that one believes that *p*" (2012: 241). On this proposal, TM is again just a method for bringing the second-order belief into consciousness.

[33] Setiya (2012) argues that this proposal is in fact about as "mad" as the view it is meant to replace.

The constitution view strikes me as unsatisfactory, however. The problem with it has to do with the fact that a subject S isn't born with the concept of belief. Since that is so, there will be a time in S's development when she has first-order beliefs but does not yet possess the ability to self-ascribe them, and there will be a time when she does have this ability. The question thus arises as to how S can, during her development, become able to form her first self-ascription of a belief *p*. Boyle and Shoemaker claim that first-order beliefs and second-order beliefs that self-ascribe them are constitutively linked but unfortunately they don't explain how this can come about. Since it can't be denied that at some point a transition from first-order to second-order beliefs does take place and since the constitution view doesn't say how the transition works, it leaves a significant explanatory gap. Since it presupposes that subjects are able to self-ascribe beliefs without saying how this can come to be in the first place, I will set the constitution view aside.

4.2 The commitment view

There is a second kind of non-epistemic account of TM that is widely discussed. Unlike the constitution view, it grants that TM is a way of acquiring self-knowledge. But rather than explaining the acquisition of self-knowledge of a belief *p* in terms of one's detecting a pre-existing fact about oneself, the account explains it in terms of one's committing oneself to the truth of *p* and making it the case that one has the belief *p*. This is the commitment view. I will focus on McGeer's (1996, 2008), and Moran's (2001, 2012) versions of the proposal.

McGeer (1996, 2008) takes intentional states, e.g., beliefs to be dispositional states and holds that we are able to bring about a match between our self-ascriptions of mental states and the behaviour our self-ascribed mental states are intended to account for by directing our behaviour appropriately to fit into the dispositional stereotype associated with the mental states (1996: 507).

McGeer holds that this bears directly on the question of how we acquire self-knowledge of beliefs. She proposes that, unlike as it is assumed in the epistemic approach, when we make psychological claims such as 'I believe *p*', then our claim isn't based on some detection of an internal state. Rather, such claims have "forward-looking truth conditions, or satisfaction conditions" in that they involve a commitment on our part to

make it the case that we have the state self-ascribed (McGeer 1996: 508). The "actively constructive" nature of the assertions we make about our own minds, McGeer holds,

> makes clear why our first-person psychological claims underwrite judgments that we know our own minds, even though our first-person claims are not themselves expressions of true, justified beliefs about our own first-order intentional states. We know our own minds because we have been trained to take on the responsibility […] for suiting our words to our deeds and our deeds to our words. (1996: 515)

By "actively involving ourselves in forming, reviewing, revising, suppressing, and selectively acting on them," we make our various first-order states

> into first-order states we authoritatively know about because we are the ones directly involved in generating and sustaining them. Thus, when we are asked to report what's in or on our minds, we do not track or map internal mental states, as we might perceptually track or map objects and events in the world. Rather, we first reflect on how, given the current state of the world, we ought to be minded; and then we actively commit ourselves to displaying the mind – for example, the intentional attitudes – that we think appropriate. (McGeer 2008: 87)

The proposal is intriguing but it too runs into problems. Consider how McGeer's view would account for my using TM to find out what I believe. According to her proposal, in the context of TM, when I wonder whether I believe *p*, I "first reflect on how, given the current state of the world, [I] ought to be minded", i.e. I determine whether *p*, because if *p*, then I ought to believe *p*, and then I "actively commit [myself] to displaying" the belief *p*, where this commitment sanctions my 'forward-looking' self-ascription of the belief *p* (ibid).[34]

If McGeer's view accurately captured what happens when I come to judge that I believe *p* via TM or indeed any other method, then when I self-ascribe the belief *p*, I should at

[34] It is worth noting that McGeer's proposal, just like, e.g., Cassam's (2014) rationality view, relies on the principle: if *p* is true, then I ought to believe *p*. Above, I argued that this principle is false and that a view of self-knowledge that relies on it will hence be problematic. The same point applies to McGeer's view. However, I shall ignore it here and will focus on another problem with her account.

that moment acknowledge that I haven't yet formed the belief. For my self-ascription of the belief would, on McGeer's view, be "analogous to [a] promise" to make it the case that I believe *p* (1996: 508), and if I took it that I already believe *p*, then I evidently couldn't sincerely promise to make it the case that I believe *p.* If I took it that I believe *p*, then there would be no point in committing myself to bringing it about that I have the belief *p*. So on McGeer's view, when I self-ascribe the belief *p*, at that moment, I won't yet take it that I do in fact believe *p*.

However, if that is so, then her proposal contradicts the facts. This is because when I come to believe that I believe *p*, then I clearly don't take it that I don't yet believe *p*. Rather, when I come to believe that I believe *p*, at that moment, for me, I do indeed believe *p*; there is no room for me to hold that I don't yet believe *p*. On McGeer's view, however, from my own perspective, when I self-ascribe the belief *p*, there will at that moment always be room for me to hold that I don't yet believe *p*. Since this is not the case, her proposal is arguably on the wrong track.

There is a second version of the commitment view. It is has been developed by Moran (2001, 2012) and combines non-epistemic and epistemic elements.[35] Moran claims that in applications of TM the transition from *p* to the self-ascription *I believe p* would fail "if the statement I arrive at, 'I believe that *p*,' is delivered in a purely *attributive* mode", i.e. in a way that presupposes a discovery of some fact about myself, because that "would require evidence about me" which reflection on whether *p* alone "does not provide" (Moran 2012: 232). Rather, when I treat the question of my belief about *p* as equivalent to the question of the truth of *p* then I turn, what Moran calls, the "theoretical" question "'What *do* I believe?' into, what he calls, the "deliberative" question 'What *am* I to believe?'" (Moran 2001: 63). Answering the latter question is, in his view, a matter of determining what is true because what is true is what it is rational to believe and hence, what one, as a rational subject, ought to believe. When one then answers the question 'Do I believe *p*?' with 'I believe *p*' by determining that *p* is true, then this is an avowal of one's belief *p* which, Moran claims, retains its first-person reference, and is expressive of a self-ascription of a belief.

[35] Some philosophers interpret Moran as offering a non-epistemic account, see, e.g., Reed (2010). But this is arguably mistaken, as will become clear in a moment, see also Roessler (2013).

Given that the facts appealed to in answering the question as to whether *p* aren't about my mental states but about *p*, how can the avowal be expressive of a self-ascription of my belief *p*?[36] This is where the epistemic element of Moran's view comes in. He phrases the problem at issue thus:

> What right have I to think that my reflection on the reasons in favour of *p* (which is one subject-matter) has anything to do with the question of what my actual *belief* about *p* is (which is quite a different subject-matter)? Without a reply to this challenge, I don't have any right to answer the question that asks what my belief [about, e.g., whether it will rain] is by reflection on the reasons in favour of an answer concerning the state of the weather. (2003: 405)

Moran's response is that

> I *would* have a right to assume that my reflection on the reasons in favour of rain provided me with an answer to the question of what my belief about the rain is, if I could assume that *what* my belief here is was something determined by the conclusion of my reflection on those reasons. (ibid)

And in general, Moran continues, if I were not allowed to make that assumption then I would have to think that even though I'm considering the reasons for *p*, and am coming to some conclusion about the matter, something other than that consideration is settling what my belief about *p* is. But if that were so, then deliberation couldn't play the role in cognition that it does play, Moran claims. For the

> aim of deliberation is to fix one's belief or intention, and it could not do so if in general the conclusion of one's deliberation […] left it as an open question, one needing to be answered in some other way, what one's actual belief about the matter is. And if there is no additional step to be made here, if I am entitled to assume that the conclusion of my reasoning tells me what my belief or intention is, then I think we have the form of the only kind of vindication that [TM] could have. (Moran 2004: 466)

[36] Moran's view too relies on the problematic principle: if *p* is true, then I ought to believe *p*. I will focus on other shortcomings with his proposal here.

Moran's point is thus that I'm entitled to assume that my belief about *p* is determined by the "conclusion of my reflection" on the reasons in favour of *p* (2003: 405), and this assumption vindicates my transition from the conclusion of my reflection about whether *p* to the self-ascription *I believe p*. The "conclusion of my reasoning tells me what my belief" is (Moran 2004: 466).

The problem with Moran's proposal is that in order for the assumption that my conclusion fixes my belief to support the transition at issue, when I conclude that *p*, I will already need to identify my conclusion that *p* as my conclusion that *p*. Otherwise, my assuming that the *conclusion* of my reflection tells me what my belief is won't provide me with a basis for proceeding from my conclusion that *p* to the self-ascription *I believe p*. But, as Moran emphasises himself, when I'm using TM, I will not be aware of my own mental events or states but only of the non-mental facts that pertain to whether *p* (2012: 227-228). Thus, I won't be aware of *p* as my concluding that *p*, because my conclusion is a mental fact: it is my judgment *p*. Since that is so, I won't be able to employ the assumption that my conclusion reveals my belief in order to attain support for moving from *p* to the self-ascription *I believe p*. Indeed, since I understand that *p* might be the case even if no one believes *p*, I will lack a basis for the transition that TM involves.

It might be suggested that the assumption that the conclusion of my reflection reveals my belief leads me to start reflecting on whether *p* to find out what I believe. And once I have then reached the conclusion that *p*, the belief *p* is automatically self-ascribed without my having to identify *p* as my conclusion first.

However, on this proposal too, before starting my reflection on whether *p*, I will already need to understand that what I then begin to do is *reflecting* on whether *p* and so reaching a conclusion on the matter. Otherwise, the assumption at issue won't lead me to start reflecting on whether *p* to find out whether I believe *p* to begin with rather than engage in some other procedure. But this means that I will already need to represent some of my mental processes, namely my reflecting as such. Since Moran doesn't explain how the representation of this aspect of my own mind comes to be, his proposal fails to offer a satisfactory account of how subjects can, via an application of TM, come to self-ascribe and know their own beliefs. Just like all the other proposals mentioned, it doesn't manage to meet (AC1).

5. Conclusion

In this chapter, I first introduced two conditions that any adequate transparency theory of self-knowledge of beliefs should meet. I then discussed a wide range of currently available transparency theories to see whether they are able to do so. I found that they don't even meet the first of the two adequacy conditions or have other shortcomings that provide good reasons to search for an alternative.

CHAPTER 4

An alternative – The implicit dual-reasoning (IDR) theory

In Chapter 3, I argued that a number of extant transparency theories of self-knowledge of beliefs are unsatisfactory. I now want to develop an alternative transparency account that avoids their shortcomings and also explains the privileged nature of self-knowledge of beliefs, which has so far been set aside.

To be able to avoid the problems of the views discussed in the previous chapter, the account will need to meet the two adequacy conditions introduced above. They were the following.

> (AC1) Any adequate account of self-knowledge of beliefs on which TM provides one with self-knowledge needs to solve the intelligibility puzzle, i.e. it needs to explain how a subject S can come to determine whether *p* in order to determine whether she believes *p* even though she understands that whether or not *p*, this won't tell her whether she believes *p*. Furthermore, the account shouldn't presuppose that S is already able to represent her own mental processes or states without explaining how she is able to do so.
>
> (AC2) Any adequate account of self-knowledge of beliefs on which TM provides one with self-knowledge needs to solve the knowledge puzzle, i.e. it needs to explain how the self-ascription *I believe p* that is thought to result from applications of TM can be justified and true.

The view that I will introduce in this chapter will meet both (AC1) and (AC2). It will do so by proposing that TM relies on a particular combination of practical and theoretical reasoning. In a condensed form, the basic idea is this.

In the context of TM, in order to determine whether she believes *p*, S will ask herself whether *p*, for she understands that in general what people will say in response to this question is what they believe about *p*. The question then elicits in S a decision-making process that pertains to what to say in response and requires her to settle whether *p*. When, in the course of settling whether *p*, S finds that *p*, then on the basis that *p* is the

case, she will decide to say that *p*, which gives rise to an epistemic state that leads her to predict that she will say that *p*. This prediction, in conjunction with the background belief that what people will say in response to the question as to whether *p* is what they believe about *p*, then provides her with a basis for inferring that she believes *p*.

What leads S, upon settling whether *p*, from *p* to the decision to say that *p* is practical reasoning. What leads her from that decision and the prediction that she will say *p* to the self-ascription *I believe p* is theoretical reasoning. Hence, an episode of a two-part or dual reasoning sustains and explains the transition from *p* to the self-ascription *I believe p*. I argue that both parts of this reasoning are compatible with S's holding that whether or not *p*, this won't tell her whether she believes *p*, and also don't presuppose any kind of representation of one's own mind. Since that is so, (AC1) is met.

Furthermore, I contend that as soon as the dual reasoning has been followed through once, most components of it are in subsequent applications of TM no longer executed but remain implicit or unarticulated in the transition from the outcome of one's reflection on whether *p* to the self-ascription *I believe p*. The transition at issue then leads to non-inferential self-ascriptions of beliefs that, given the reliable relation between judgments and beliefs with the same content, amount to privileged self-knowledge. (AC2) is hence met also.

I call this account of TM and privileged self-knowledge of beliefs the *implicit dual-reasoning* (IDR) *theory*. I shall develop it in a step-by-step manner.

In section 1, I first introduce the assumptions that underlie the IDR account before employing these assumptions to elaborate the proposal and showing how the latter manages to meet (AC1). In section 2, I turn to the knowledge puzzle and argue that the IDR theory also meets (AC2). Section 3 shows that TM-based self-ascriptions are, according to the IDR theory, non-inferential in nature. In section 4, I elaborate this point and explain how the proposal provides an account of privileged self-knowledge of beliefs. I end, in section 5, by comparing the IDR account with the transparency theories discussed in Chapter 3 and argue that it is preferable to them.

1. From *p* to *I believe p* – Solving the intelligibility puzzle

The view of TM that I want to develop in this chapter relies on three main assumptions. The assumptions pertain to issues that may at first glance seem to have little to do with TM or self-knowledge of beliefs. The connection will, however, become clear in due course.

The three assumptions are the following.

(1) A decision to φ involves an epistemic state that (i) leads one to predict that one will φ, and (ii) sanctions non-conditional judgments on the basis of the assumption that one will φ.
(2) Answering questions involves decision-making.
(3) Subjects (i) have a truth bias, and (ii) assume that what S will say in response to the question of whether *p* is what she believes about *p*.

In the next three sub-sections, I will explain and support (1)-(3). With these assumptions in place, I will then return to TM and self-knowledge of beliefs.

1.1 Decisions and predictions

The first assumption that I will build on below is:

(1) A decision to φ involves an epistemic state that (i) leads one to predict that one will φ, and (ii) sanctions non-conditional judgments on the basis of the assumption that one will φ.

I will consider points (i) and (ii) in turn.

<u>Assumption (1), point (i)</u>

Assumption (1) pertains to decisions. Decisions are the mental events that immediately result from episodes of practical reasoning on whether to φ. They settle whether to φ, which is the defining element of their functional role, and typically give rise to

intentions to φ, which are standing states (e.g., Velleman 1989, 2007; Bratman 1987, 2009; Holton 2009).

When it comes to the nature of intentions, two views can be distinguished: cognitivism and non-cognitivism. Cognitivists hold that the intention to φ involves the belief that one will φ (see, e.g., Hampshire 1965; Harman 1976; Velleman 1989, 2007; Wallace 2001; Setiya 2007; Ross 2009). Non-cognitivists deny that this is so (see, e.g., Bratman 1987, 2009; Holton 2009).

Assumption (1) says that a decision to φ involves an epistemic state that (i) leads one to predict that one will φ and (ii) sanctions non-conditional judgments on the basis of the assumption that one will φ. Assumption (1) hence captures a version of cognitivism about decisions. In the rest of the sub-section, I will offer support for the assumption.

It will be conducive to approach the matter indirectly via first addressing a common objection to cognitivism about intentions. Bratman (1987, 2009) gives the following example to argue against cognitivism about intentions. He imagines himself having the plan to stop at a bookstore, writing

> I might intend now to stop at the bookstore on the way home while knowing of my tendency towards absentmindedness – especially once I get to my bike and go into 'automatic pilot'. If I were to reflect on the matter I would be agnostic about my stopping there […]. (Bratman 1987: 38)

> About this case I am inclined to say: I intend to stop, but I do not believe I will stop (though I do not believe I will not stop). (Bratman 2009: 21)

The example suggests that an intention to φ doesn't entail that one believes that one will φ. It might seem that by extension a decision to φ doesn't imply that one predicts that one will indeed φ either.

However, even if an intention to φ doesn't involve a belief that one will φ, it isn't obvious that the same holds for a decision to φ. To motivate the distinction between intentions and decisions involved here, consider the following. S might have the intention to win in a competition even though she didn't and can't decide to win it. She

might decide to do her best to win the competition, i.e. "to do things that will give her the best chance of achieving her goal of winning the competition", and indeed when she "decided to do her best to win the competition her goal/aim in making that decision is to win. In that sense, her intention is to win. But she hasn't [then] decided to win the competition" (Soteriou 2013: 286).

Similarly, we might hold that while Bratman decided to do his best to stop at the bookstore, and in this sense intends to stop there, he nonetheless didn't *decide* to stop at the store (ibid). On this interpretation of the example, he would still intend to φ without believing that he will φ and so his point against cognitivism about *intentions* would still hold. But there would be nothing in the example to show that one can decide to φ without predicting that one will indeed φ.

I will now provide a positive reason for the view that upon deciding to φ, one predicts that one will φ. Suppose you need to be at the airport at 10:30 am tomorrow to catch a flight to Berlin and now reflect on how best to get to the airport on time. Suppose that it occurs to you that if you take the bus to the airport at 9:30 am, you will be at the airport well in time. On the basis of this conditional belief, you then decide to take the bus. What leads you to the decision to take the bus is your viewing it as highly likely that if you are taking the bus, you will get to the airport on time, and will so achieve your goal of catching the plane.

Suppose now, for the sake of argument, that upon deciding to take the bus, you become agnostic as to whether you will manage to take the bus. Clearly, once agnosticism on the matter arises, you will also become agnostic as to whether you will be at the airport by 10:30 am.[37] Given that, by assumption, you feel the need to indeed be there by that time, you will now re-consider the matter of what to do. The question of what to do won't be settled for you yet. Rather, your desire to be at the airport on time will

[37] It might be suggested that the agnosticism never occurs to you after the decision, and after the decision you never raise the issue of whether you will manage to satisfy your desire via the action you chose. In response to this, note that it will arguably still be *possible* for you to reflect on the matter. Suppose, e.g., right after your decision a friend phones you and questions whether you will be at the airport by 10:30 am. You will then need to take a stance, and at this point the agnosticism mentioned will come into play and lead you to re-open the matter.

motivate you to continue searching for an alternative action that will ensure that you will be at the airport at 10:30 am.

But if that is so then your initial decision to take the bus at 9:30 am didn't manage to settle your reasoning on what to do. A decision that doesn't manage to settle your reasoning on what to do, however, isn't a decision at all. For as noted, it is a crucial part of the functional role of decisions to settle reasoning on what to do. Thus, in general, if there is a decision to φ at all, then upon deciding to φ, there can't be agnosticism, or doubt on whether one will indeed φ. Rather, one needs to predict that one will indeed φ.[38]

How can this epistemic dimension of decisions be explained? The following two factors help offer an explanation and also lend further support to the existence of that dimension of decisions.

Factor 1

Empirical studies suggest that when subjects deliberate on what to do to satisfy a certain desire to ψ, then they don't only consider possible actions that would allow them to do so. They also assess how feasible it is for them to perform those actions (Gollwitzer 1990; Gollwitzer and Bayer 1999; Oettinger and Gollwitzer 2010). The point is perhaps intuitive enough. To avoid selecting an action φ that we are unlikely to perform successfully, and that is hence unlikely to satisfy our desire to ψ, we try to predict the likelihood of our successfully φ-ing if we had to φ under the conditions C that we envisage to hold at the moment when φ would need to be performed to satisfy our desire to ψ. It is plausible to assume that as part of our pre-decision deliberation on what to do to ψ, we consider and assign (typically unconsciously) probability values to conditionals of the form 'if I had to φ at time t under C, then I would successfully φ', where this will involve drawing on our prior experience, knowledge about our abilities, dispositions,

[38] There might be situations when, e.g., I find that all options are equally likely to fail to provide me with a clue as to whether I will satisfy my desire, and I will still decide for one of them. A decision to φ thus doesn't always seem to involve a prediction that one will φ. However, this case is rare. In what follows, I shall ignore such cases, as they won't threaten the proposal below. The general claims I make here about the links between decisions and predictions of success should be read as referring only to typical decisions.

context, etc. When an assessment of the options has occurred and the probabilities have been assigned, then the conditional with the highest probability of success will determine the action that we decide to perform and therewith commit ourselves to performing. This view is supported by a number of studies that show that when "perceived feasibility is high, people strongly commit to attaining the goal [e.g., to *φ*]; when perceived feasibility is low; they form either low goal commitment or none at all" (Achtziger et al. 2012: 125).

With the preceding in mind, the following account of one's prediction of success at *φ*-ing that is involved in one's deciding to *φ* becomes available. Suppose that a subject S has the desire to *ψ*, and upon deliberation comes to think that if she *φ*s at *t* under *C*, then she will *ψ*. Suppose too that she assigns a high probability to the conditional 'if I had to *φ* at time *t* under *C,* then I would successfully *φ*', and on that basis, given that she wants to *ψ*, decides to *φ*. Once she has made her decision, S will therewith have committed herself to *φ*-ing at *t* under *C*. That is, at that moment, for her the matter is settled and from her own perspective, she will have to *φ* at *t* under *C*. Since that is so, the antecedent of the conditional 'if I had to *φ* at time *t* under *C,* then I would successfully *φ*' that fed into her thinking that if she *φ*s at *t* under *C*, then she will *ψ* (on which basis she formed the decision to *φ*), is now satisfied. Given this, her pre-decision assessment that she *would* with high probability successfully *φ* at *t* under *C* becomes replaced with an assessment that she *will* with high probability *φ* at *t* under *C*. That is, her decision triggers a change of the antecedent 'if I *had* to *φ*' to 'I *have* to *φ*' and so of the 'would' in the corresponding succedent to a 'will'. As a result, upon deciding to *φ*, S is going to predict that she will indeed *φ*.

Factor 2

One's pre-decision assessment of the likelihood of success at *φ*-ing if one had to *φ*, which feeds into one's decision to *φ*, is the first factor that helps explain how it can be that when one decides to *φ*, one predicts that one will indeed *φ*. There is a second factor that once one has decided to *φ* sustains one's prediction that one will indeed *φ*. Velleman (1989, 2007) describes it in his work on intentions. He writes that one can accept something as true in two different ways. There is "accepting so as to reflect the truth, and accepting so as to create the truth" (Velleman 1989: 194). The former involves a passive "direction of guidance" and is a feature of belief; the later involves

an active one and is a feature of decision (ibid). Velleman notes that it is common to assume that if a

> mental state is cognitive, representing how things are, then it must be caused by how they are; whereas if it causes what it represents, then the state must be conative […]. This assumption implies that a cognitive direction of fit entails a passive direction of guidance; and, conversely, that only states with conative direction of fit can be active or practical. (1989: 25)

However, Velleman holds that a mental state's direction of fit and its direction of guidance shouldn't be conflated. That is, e.g., a cognitive direction of fit shouldn't be conflated with a passive direction of guidance; and a conative direction of fit shouldn't be conflated with an active direction of guidance. This is because a choice or decision has a *cognitive* direct of fit yet an *active* direction of guidance, Velleman argues. He writes that when, e.g., "I make a choice, a question is resolved in the world by being resolved in my mind. That I am going to do something is made true by my representing it as true. So choice has the same direction of fit as belief but the same direction of guidance as desire" (ibid).

The thought is that when I make a choice or decision to φ, then I take it that I will indeed φ because my decision to φ commits me to φ-ing and becomes therewith partly the cause of my φ-ing: my decision to φ moves me to φ and inhibits me from performing other actions that I don't expect. It thereby makes it true that I will φ rather than do something else. As a result of this, even though with my assumption and certainty that I will indeed φ I seem to "jump to a conclusion" about my action "before the evidence is complete, I jump with the assurance that the conclusion will achieve verity even as I land" (Velleman 1989: 57).

The active nature of my decision to φ, i.e. its feature to make it the case (by imposing constraints on subsequent reasoning and acting) that I will indeed φ, is the second factor that, once I've decided to φ, contributes to my prediction that I will indeed φ. Just like the first one, it helps explain and support the existence of the epistemic dimension of decisions without assuming any extra mechanism in addition to the one already required for decision-making itself.

The view that a decision to φ involves an epistemic state that leads one to predict that one will φ is the first part of assumption (1). I now turn to the second part.

Assumption (1), point (ii)

Assumption (1) holds that upon deciding to φ, one will (i) predict that one will indeed φ, and this then (ii) allows one to form non-conditional judgments on the basis of the assumption that one will φ. So far I have only focussed on (i). What supports point (ii)?

I shall take it that because of factors 1 and 2, upon deciding to φ, one plans one's further action and reasoning on the assumption that one will φ. But what does this amount to?

It is plausible that when one plans one's action and thinking on the assumption that one will φ, then one considers that assumption as one that is to be fulfilled by an action that makes that assumption true. In considering the assumption that one will φ as an assumption that is to be fulfilled by an action that makes that assumption true, one therewith considers that assumption as one that entitles one to make straight-out non-conditional judgments that follow from that assumption, because in discharging the assumption one therewith makes true the judgments that follow from that assumption. Hence, when one makes a decision to φ, one will therewith take oneself to be entitled to make non-conditional judgments that follow from that assumption (see also Soteriou 2013).

However, the decision to φ isn't itself the direct basis for the non-conditional judgments at issue, for it doesn't itself carry the content that one will φ. Rather, the decision involves a separate epistemic state that does so and is supported by factors 1 and 2. I shall refrain from committing myself to the view that this state is a full-blown belief that one will φ. For the evidential basis of the state is arguably quite distinct from that of a stereotypical belief. What I will commit to here and rely on later is just that when a subject S decides to φ, then she predicts that she will indeed φ, and this in turn permits her to form non-conditional judgments that follow from the assumption that she will φ.

This completes my argument for the first main assumption that I will build on below, i.e. the assumption that a decision to φ involves an epistemic state that (i) leads one to predict that one will φ, and (ii) sanctions non-conditional judgments formed on the basis of the assumption that one will φ.

1.2 Communication and decision-making

As mentioned, the account of TM that I want to offer rests on three main assumptions. The second one is the view that:

(2) Answering questions involves decision-making.

To see the motivation for (2), note that in communication, when people are asked questions, they don't normally just blurt out the answers but tend to control their responses. Depending on the context and the person asking, people decide to provide different responses, for depending on the context and the person asking, the kind of response they provide will have different consequences for them. For instance, while telling the truth will often improve cooperation and so contribute to success in projects with a shared goal, there will also be cases when withholding an answer and the truth is more advantageous for achieving the goal that one set oneself. "Communicators, accordingly, choose between expressing and withholding a message, whether truthful or not, that, if believed by the addressee, should have the [communicators'] desired effects" (Sperber 2001: 404).

The claim here is not that every intentional communicative act is preceded by a choice or decision on what to say. This would be implausible. There are intentional acts, including communicative acts, that aren't preceded by a decision (see, e.g., Anscombe 1976; Mele 2009). The claim with assumption (2) is rather that in response to questions in particular, the producer of the response takes into account the source of the question and decides most minimally between expressing or withholding a piece of information.

Note too that the claim is not that the decision-making process or the resulting decision itself is conscious. The proposal is that the monitoring of the source of the question and the decision-making on how to respond occurs typically automatically at the unconscious level. This assumption isn't *ad hoc* but in line with empirical studies that show that subjects monitor, e.g., the competence, reliability, and honesty of an communication partner and respond differently to his/her message depending on the outcome, where the epistemic assessment and decision-making involved happen automatically and unconsciously (see, e.g., Richter et al. 2009; Sperber et al. 2010; Mercier 2013).

In short, then, I shall take the fact that we respond to the same question in different contexts and interactions with different individuals with different answers to yield sufficient motivation for the assumption that when one is asked whether *p*, upon hearing and understanding the question, one will activate (typically unconscious) decision-making processes pertaining to how to respond.

1.3 The truth bias

The third and last main assumption that I will appeal to below is:

(3) Subjects (i) have a truth bias, and (ii) assume that what S will say in response to the question of whether *p* is what she believes about *p*.

I will consider points (i) and (ii) again separately.

Assumption (3), point (i)

The "truth bias" is sometimes viewed as one of the "most cited and documented" cognitive biases in psychological research on deception (Burgoon et al. 2008: 575). It is "the tendency to actively believe or passively presume that another person's communication is honest independent of actual honesty" (Levine 2014: 380).

The existence of the bias is well established (see, e.g., Zuckerman et al. 1981; McCornack and Parks 1986; Levine et al. 1999; Levine and Kim 2010; Vrij 2008; D'Agata and Jacobson 2014). For instance, in a comprehensive meta-analysis of lie detection studies, Bond and DePaulo (2006) found a general truth bias across studies. 56% of messages were judged as honest and 44% as deceptive even though in the studies analysed subjects were presented with an equal number of lies and truths. Similarly, reviewing studies on lie detection from two decades, Levine and Kim (2010) write that the truth-bias score in the studies was consistently between 56% and a high of 72%. A survey of the literature shows further that the bias is particularly pronounced in adults when they interact face-to-face (Buller et al. 1991; Burgoon et al. 1994), when they are unaware that their task is to detect lies (e.g., McCornack and Levine 1990; Levine et al. 2000), and, importantly for my purpose below, when they know the

message source. For instance, when the senders are close friends or relatives, then the addressee exhibits a significantly stronger truth bias than when the message comes from strangers (McCornack and Parks 1986; Millar and Millar 1995), which has led some researchers to also refer to the bias as the "relational truth-bias heuristic" (D'Agata and Jacobson 2014: 156).

The bias isn't absolute, but can be reduced by increasing scepticism (DePaulo et al. 2003). For instance, when test subjects were presented with statements by a salesperson attempting to sell a product, or when they were told that the sender might be dishonest, their truth bias decreased (McCornack and Levine 1990; Millar and Millar 1997). Police officers too, as they frequently assume suspects to be guilty, didn't display a significant truth bias during interrogations (Vrij 2008: 150).

It is telling, however, that in some studies even professional lie detectors tended to believe senders more than, e.g., observers (Hartwig et al. 2004; Bond and DePaulo 2006), and judges who were made highly sceptical of the message still exhibited the bias (McCornack and Levine 1990; Levine and Kim 2010). Overall, the data strongly suggests that in the absence of any reason to be suspicious about the sender's honesty, "most people are truth-biased most of the time" (Levine and Kim 2010: 27).

Assumption (3), point (ii)

Assumption (3) holds that subjects (i) have a truth bias, and (ii) take it that what S will say in response to the question of whether *p* is what she believes about *p*. The preceding pertains to (i). I will now motivate (ii).

During development, neurotypical children learn that when a subject responds truthfully, then this is distinct from her telling the truth. For instance, suppose S is presented with the following scenario, which I adapted from the false-belief task paradigm (Wimmer and Perner 1983). Two individuals, A and B, are in a room with two boxes and an object M. B, who owns M, places M into box 1 and then leaves the room. Once B has left the room, A takes M out of box 1 and puts it into box 2. Afterwards B returns and is asked about the location of M. Suppose S is then questioned about what B will say. If she took it that people tend to tell the truth, she would be led to predict that B will say that M is in box 2. This would be the wrong prediction, for B has

a false belief about M's location, and will thus on the basis of that belief say M is in box 1. Four-year olds are able to correctly predict that B will say M is in box 1. Importantly, when they do so, they don't take it that B is untruthful in saying that B is in box 1. Rather, they assume that B is honest and that in general a subject's truthful response isn't necessarily tied to what is the case but to what she believes to be the case. That is, they seem to make the implicit default assumption:

(IDA) What subjects will say in response to the question of whether *p* is what they believe about *p*.

(IDA) underwrites the familiar practice that we adopt to determine other people's beliefs: to find out whether they believe *p*, we tend to simply ask them whether *p*. It seems clear that we wouldn't engage in this practice unless we assumed (IDA). (IDA) arguably belongs to the stock of generalisations that the mindreading system has available to work out other people's attitudes to make sense of and predict their actions.

This completes my argument for the third and final main assumption that I will depend on below, namely the view that subjects (i) have a truth bias, and (ii) assume that what another person S will say in response to the question of whether *p* is what she believes about *p*.

1.4 TM revisited

I have introduced and motivated the following three assumptions:

(1) A decision to φ involves an epistemic state that (i) leads one to predict that one will φ, and (ii) sanctions non-conditional judgments on the basis of the assumption that one will φ.
(2) Answering questions involves decision-making.
(3) Subjects (i) have a truth bias, and (ii) assume that what S will say in response to the question of whether *p* is what she believes about *p*.

With (1)-(3) in place, I will now return to TM and the issue of how a subject can come to form the self-ascription *I believe p* on the basis of *p*.

I want to start approaching the issue without assuming that one has some special, non-third-personal method of working out one's own beliefs. Indeed, for the sake of argument, I shall initially take it that in finding out one's own beliefs, one has only the resources available that one employs for forming third-personal belief ascription.

As will become clear, these resources, assumptions (1) to (3), and a combination of (a) practical reasoning and (b) theoretical reasoning that is based on these assumptions are sufficient to make the transition from *p* to the self-ascription *I believe p* intelligible. I will now go through the two kinds of reasoning that I take to underlie TM.

(a) Practical reasoning

Suppose a subject, e.g., I want to find out whether I believe *p*. I might then wonder what to do to settle the matter. Suppose I understand (IDA), i.e., that typically, when people are asked whether *p*, what they will say in response is what they believe about *p*. With this in mind and with no other means available for working out my own belief, to find out whether I believe *p*, I will then decide to ask myself whether *p* so as to settle the matter.

Once I decide to ask myself whether *p*, I will predict that I will ask myself whether *p*. This is because of the points made in the preceding sub-section.

Further, I will understand that typically when a subject S asks a question, she does so because she requires an answer in order to achieve some goal or other, and the provision of a correct answer will facilitate her achieving that goal. In the case at hand, I am the source of the question myself, and I'm also aware of this, for I predict that I will ask myself whether *p*. Given that I understand that typically when a subject S asks a question, she does so because she requires an answer to the question in order to achieve some goal or other, I will then take it that the same holds for me.

Since the provision of a correct answer will facilitate my achieving some goal or other, I will hence want to answer the question as to whether *p* correctly, for doing so will by my own lights increase the probability of success in my projects. Since that is so, I will need to determine whether *p*, for I understand that what the correct answer to the question as to whether *p* is will evidently depend on whether or not *p* is the case.

Suppose I then reflect on whether *p* and find that *p*. Since *p*, and since I want to correctly answer the question of whether *p*, I will decide to say *p* in response to the question. Further, since in general one's decision to *φ* comes with a prediction that one will *φ*, this means that I will predict that I will say *p*.

Indeed, I will not only predict that I will say *p*. I will also understand why I will say *p*. I will understand that I will say *p* to correctly answer the question as to whether *p*, for to correctly answer the question is what I want and what leads me to the decision to say *p* to begin with. In my decision-making, I'm aware of what I want and therewith understand why I will say *p*.

To be clear, being aware of *what* I want doesn't require being aware *that* I want it. The point here is that in my decision-making I'm aware of the *content* of my desire, not of the desire, the attitude, itself. Consider neonates. Clearly, they are aware of what they want, and on the basis of what they want form decisions to act. Yet, it is implausible to assume that they are also aware *that* they want what they want, for they arguably still lack the capacity to represent mental states such as desires. Similarly, in my decision-making on how to respond to the question as to whether *p*, I can be aware of *what* I want, namely correctly answer the question, without necessarily being aware *that* I want this. It is in virtue of the awareness of what I want, which doesn't presuppose any representation of my own mental states, that, upon forming the decision to say *p*, I understand why I will say *p*.

The steps of the practical reasoning leading up to my decision to say *p* might then be summarised thus:

(1) Do I believe *p*? [Expression of my occurrent desire to determine whether I believe *p*.]
(2) What subjects will say in response to the question of whether *p* is what they believe about *p*. [This is (IDA); a background belief.]
(3) Ask (self) whether *p*. [This is a decision based on a combination of the desire expressed in (1) and the belief (2) applied to myself.]
(4) I will ask (self) whether *p*. [This is an epistemic state/prediction that is based on decision (3).]

(5) I wouldn't ask whether *p* unless I needed a correct answer to achieve some goal or other. [This is a judgment derived from the background belief that normally when a subject S asks a question, she does so because she requires an answer to achieve some goal or other, where her finding the correct answer will facilitate her achieving that goal.]

(6) Correctly answer the question as to whether *p*. [This is a desire based on my interest in achieving my goals.]

(7) If *p*, then saying *p* in response to the question as to whether *p* is correctly answering that question. [This is a background belief.]

(8) *p.* [This is a judgment that results from my settling whether *p*.]

(9) Say *p* (to correctly answer the question as to whether *p*). [This is a decision based on a combination of desire (6), belief (7), and judgment (8).]

To emphasise, (1)-(9) is an episode of *first-order* reasoning. That is, when I engage in the reasoning, I don't need to understand that I *desire* to, e.g., correctly answer the question as to whether *p* or that I *judge* that *p* in order to form the decisions (3) or (9). For as noted, if the desire and the judgment had to be represented as such in order for subjects to settle what to do then this would imply that neonates, who arguably still lack the ability to represent something as a desire or judgments (e.g., Block 2007), can't settle what to do – which seems absurd.

(b) Theoretical reasoning

Suppose, then, that via the just-mentioned episode of practical reasoning, I come to decide to say that *p*. On the basis of this decision, I can now, by a piece of theoretical reasoning, work out my own belief about *p*. For, given the points made in the preceding section, upon forming the decision, I will predict that I will say *p* (to correctly answer whether *p*). And if I predict that I will say *p* (to correctly answer whether *p*), then this prediction, together with (IDA), i.e., the initial belief that what a subject will say in response to the question as to whether *p* is what she believes about *p*, will provide me with a basis for inferring that, since I will say *p* (to correctly answer whether *p*), I believe *p*. The theoretical reasoning that leads me to this conclusion can be summarised thus:

(1)* What subjects will say in response to the question of whether *p* is what they believe about *p*. [This is (IDA); a background belief.]
(2)* I will say *p* (in response to/to correctly answer the question of whether *p*). [This is an epistemic state/prediction based on the decision to say *p*.]
(3)* I believe *p*. [This is a judgment.]

As noted above, premise (1)* is already needed to make sense of other people's utterances and to find out about their beliefs via probes. Premise (2)* captures the epistemic state that is involved in my decision to say *p*. Since this state sanctions non-conditional judgments, it sanctions (3)*. With judgment (3)* in place, I have worked out my belief about whether *p* by a piece of *dual reasoning*, i.e. a combination of practical reasoning, namely (1)-(9), and theoretical reasoning, namely (1)*-(3)*.

However, the inference from (2)* to (3)* might strike one as problematic, for might it not be that I will say *p* even if I don't believe *p*? Might I not lie?

Here the point about people's truth bias comes in. I argued above that empirical data and theoretical considerations support the view that people are truth-biased in that they take it, by default, that if a subject S is asked whether *p*, then her response will be truthful. I noted that this bias is most pronounced in interactions with close friends or relatives leading some psychologists to call it the "relational truth-bias heuristic" (D'Agata and Jacobson 2014: 156). Perhaps this aspect of the bias is to be expected. For the more interdependently one is related to the subject whose statement one is presented with, the higher the likelihood that insincerity will backfire for the subject exhibiting it. Hence, in highly interdependent relationships, both the incentive for insincerity (for the sender) and the need to engage vigilance processing (for the addressee) decrease. It is to be expected then, and borne out by the data, that in such relationships, the addressee will display little suspicion regarding the sender's message. In close relationships, in which the flourishing and the detriment of one is immediately also that of the other, both parties will refrain from deception and both parties will increase trust.

Given that oneself is arguably as close as one can possibly be to anyone, presumably one will be highly confident when it comes to the truthfulness of what one will say to oneself in response to the question as to whether *p*. Indeed, since in the case of self-directed queries and answers, both sender and addressee are one and the same subject, a

default of complete trust is to be expected. This is because the sender's insincerity will immediately negatively affect him/herself, and there will thus be no incentive for him/her to be untruthful. Conversely, for the addressee, there will be no need for epistemic monitoring to detect deception. These points help alleviate the initial worry that it might be that one will, in response to the question of whether *p*, say *p* even when one doesn't believe *p*: under the circumstances at issue, from one's own point of view, the possibility of one's own insincerity simply won't arise and so won't undermine one's inference from what one will say to what one believes.

In sum, then, the preceding discussion offers an account of how answering the self-directed question 'Is *p* the case?' can allow one to find out about one's belief about *p*. By appealing to an episode of dual reasoning, it makes intelligible how one can from *p* proceed to the self-ascription *I believe p*. Since such a transition is precisely what is claimed to happen in applications of TM (see, e.g., Evans 1982; Moran 2001), the preceding discussion provides an account of how one can via TM come to work out one's own belief *p* on the basis of *p*.

1.5 Objections

To further develop and defend the dual reasoning account just introduced, I will now discuss five objections to it.

Objection (1)

According to the account, since I decide to ask myself whether *p* and decide to say that *p*, I will predict that I will execute these actions and this prediction supports a piece of theoretical reasoning to the conclusion that I believe *p*. However, the objection continues, for the prediction that I will say *p* to support an inference to the conclusion that I believe *p*, it will itself need to be supported by evidence indicating that I will indeed say *p*. Yet, no such evidence is available prior to my saying *p*, and in fact various factors might then still lead me to not say *p*. Since that is so, my prediction that I will say *p* isn't evidentially supported in the right way to itself provide sufficient support for inferring that I believe *p*.

Response

Note that upon deciding to say *p*, from my own point of view, there is no room for doubt or agnosticism on whether I will say *p*. If there were, then, for the reasons mentioned above, the matter wouldn't be settled for me yet and so no decision proper would have occurred in the first place. If I have decided to say *p* at all then for me I will indeed say *p*. And this constitutes, from my own point of view, as good an epistemic basis for non-conditional judgments as if I had already said that *p*. The explanation for this is that my view that I will say *p* is supported by the evidence that led me (prior to forming the decision to say *p*) to predict that I *would* say *p* if I *had* to do so under the circumstances at issue. The prediction will have been strong because there is little that could prevent me from saying things, e.g., in inner speech, if I had to. Since upon deciding to say *p*, I will by my own lights have to say *p* because the decision involves a commitment to saying *p*, the high pre-decision probability that I *would* say *p* if I had to is transformed into a high probability that I *will* say *p*. This is what leads me, upon deciding to say *p*, to strongly predict that I will indeed say *p*. Because of the certainty involved, this prediction constitutes from my own point of view a sufficient epistemic basis for the same kind of non-conditional judgments that I would derive from my already having said that *p*.

Objection (2)

The dual reasoning account relies crucially on the assumption that since I decide to ask myself whether *p* and decide to say *p*, I will predict that I will execute these actions prior to executing them. But, the objection continues, for this prediction to be able to support an inference to the conclusion that I believe *p*, it will need to have conceptual content. That is, I will need to conceptualise what I'm going to do *as* my *asking* myself a question and *as* my *saying* that *p*. Yet, my being able to conceptualise something as my asking myself a question, and as my saying *p* will arguably already presuppose some self-ascription of mental events or states and therewith violate (AC1).

Response

According to the proposal, when I'm using TM, I do need to have a grasp of what it is for me to ask myself a question and say *p*. However, this doesn't require me to represent

any of my own mental events or states, for I can learn what it is for me to say that *p* to S by being told by others that a particular verbal act I'm executing is saying that *p* to S. In order to learn this, I only need to keep track and conceptualise my own overt action of uttering that *p*. Further, in order for me to understand what it is to be asking myself a question, all that is required is that I understand that saying something in a certain way, i.e. with a verb-subject inversion, to a certain person, is asking that person a question. None of this requires me to represent any of my own mental events or states.

Objection (3)

It might be argued that on the dual-reasoning account, it is assumed that one moves at some point in one's reasoning from *p* to the self-ascription *I believe p*. But, the objection continues, one can't do that because one will also understand that *p* might be the case even if no one believes *p* and this will prevent one from engaging in the transition envisaged.

Response

Recall that on the view proposed, the self-ascription *I believe p* is formed on the basis of a combination of practical and theoretical reasoning. In the episode of practical reasoning, one wants to correctly answer the self-generated/directed question as to whether *p*, one finds that *p*, and thus decides to say *p* to answer the question. In the episode of theoretical reasoning, one takes it that what subjects will say in response to the question of whether *p* is what they believe about *p*, predicts that one will say *p* (to correctly answer the question of whether *p*), and so concludes that one believes *p*.

Crucially, the theoretical reasoning doesn't involve any move from *p* to the self-ascription *I believe p*. Hence, it is compatible with one's holding that *p* might be the case even if no one believes *p*. Similarly, one's move from *p* to the decision to say *p* in the practical reasoning too is compatible with one's holding that *p* might be the case even if no one believes *p*. For one's holding that *p* might be the case even if no one believes *p* doesn't undermine one's forming a decision on *p*'s basis. In one's first-order decision-making, one doesn't need to first represent *p* as one's own (or anyone's) belief to form the decision. Assuming otherwise is, as I emphasised above, developmentally implausible, for, as noted, neonates engage in decision-making but arguably still lack

the ability to represent mental events or states. This suggests that first-order decision-making doesn't require any representation of the attitudes involved. If this is right then both episodes of reasoning that together lead me to form a self-ascription of the belief *p* are compatible with my holding that *p* might be the case even if no one believes *p*.

But why do I then still ask myself whether *p* in order to determine whether I believe *p*?

I ask myself whether *p* because I assume that what I will say in response will be what I believe about *p*. This assumption results from applying to myself (IDA): what subjects will say in response to the question of whether *p* is what they believe about *p*. I re-cycle this principle to work out my own belief. The important difference between my use of it to determine my own beliefs and my use of it to determine other people's beliefs is that I need to wait for and interpret other people's responses to my question first in order to work out their belief about *p*. When it comes to my own response to the question, however, this isn't required. For I have an insight into what I will say prior to uttering it in virtue of deciding to perform that action in the first place.

Further, unlike in the case of other people, in my own case, I'm both sender and addressee at the same time. Since that is so, there is no need for an interpretation of the response at issue to come into play either, because if I'm sender and addressee, I evidently can't, e.g., deceive myself with my response or tell jokes to myself, misread my intention etc. anymore than I can tickle myself. Given that in my own case both the sender of the message and the addressee is I, for the reasons mentioned above, this will lead me to trust in the truthfulness and authenticity of my response to the probe, making any interpretive decoding to work out the underlying motive redundant.

This is very unlike in the other-related case in which there is even in cooperative social environments always the potential and incentive for deception, and one's trust in the truthfulness and authenticity of someone else's response hence won't be similarly absolute (Sperber et al. 2010). The basic point that the self is distinct from the other and one's own mind is not the same as someone else's will prevent this from happening. The fundamental self-other difference means that the production of other-ascriptions of beliefs will always involve at least some interpretive, inferential processing. In contrast, the identity of oneself with oneself (the identity of sender and addressee in self-directed probes) has as a result that this processing is suspended.

Hence, even though for forming self-ascriptions of beliefs I redeploy resources that I already rely on for forming other-ascriptions of beliefs, the use of these resources and the nature of the result of it will remain fundamentally different. It will be interpretive in the other-related case and non-interpretive in the self-related case.

The dual-reasoning account, then, explains how I can come to ask myself and settle whether *p* to determine whether I believe *p* even when I acknowledge that whether or not *p*, this won't tell me whether I believe *p*. The proposal thus solves the intelligibility puzzle and the problem captured in objection (3).

Objection (4)

It might be argued that if the dual-reasoning account were on the right track, then there shouldn't be a problem about TM and the view that subjects who understand that whether or not *p*, this won't tell them whether they believe *p*, nonetheless begin determining whether *p* to find out whether they believe *p*. There shouldn't be a problem because people would already be familiar with the episodes of reasoning introduced above. Yet, the objection continues, this doesn't appear to be the case. Most subjects using TM seem to move from *p* directly to the self-ascription *I believe p* without any inference and clue about the reasoning mentioned.

Response

Note that in the preceding, there was no claim that the sequence of practical and theoretical reasoning is or needs to be conscious. Typically, it isn't. Typically, when subjects use TM, upon wondering whether they believe *p*, they will ask themselves whether *p*, reflect on the matter, and, upon finding that *p*, self-ascribe the belief *p*. No other component of the dual reasoning enters consciousness.

Indeed, I want to suggest that normally the other components of the reasoning don't even occur unconsciously; they aren't executed at all. The proposal is this.

Once S understands that what subjects will say in response to whether *p* is what they believe, when she wonders whether she herself believes *p*, S will ask herself whether *p* and follow through the dual reasoning introduced. This will take place *unconsciously*.

That is, both the decision-making and the judgment-formation to get from *I will say p* to the self-ascription *I believe p* will occur at the unconscious level. As soon as this has happened once, S is able to treat the link between *p* and the self-ascription *I believe p* as sanctioned and will in subsequent applications of TM refrain from executing the entire dual reasoning. This is because when S has the desire to determine whether she believes *p*, believes that what subjects will say in response to whether *p* is what they believe about *p*, desires to correctly answer self-sourced questions and so on, then, given the way her cognitive system is set up, the structure of the decision-making and the judgment-formation (to get from the prediction that one will say *p* to the self-ascription *I believe p*) will remain invariant in different applications of TM and won't affect the outcome of the production of the self-ascription. From the cognitive system's point of view, executing the decision- and judgment-forming processes hence becomes redundant when the task is to produce a self-ascription of a belief about *p*. As a result, these parts of the dual reasoning remain unarticulated or as I shall say *implicit* in the transitions involved in TM.

The motivation for this view comes from empirical studies. There is much evidence that judgment- and decision-making systems tend to avoid computations and adopt shortcuts (*heuristics*) to simplify information processing if they can (see, e.g., Kahneman and Frederick 2002; Evans 2008; Kahneman 2011; De Ney et al. 2013). This is due to the fact that the resources (e.g., attention) that are available to the cognitive system as a whole are necessarily limited forcing it to be a "cognitive miser" and reduce processing costs whenever it can (Fiske and Taylor 2013).[39] Given that the system that forms self-ascriptions of beliefs is just another judgment-forming system, there is reason to assume that it will operate under similar constraints as other judgment-forming systems and also adopt mental shortcuts. With this is mind, it becomes plausible to assume that the system will leave the components of the dual reasoning that don't make a difference to the outcome of the formation of a self-ascription of a belief in subsequent applications of TM (i.e., once the inferential link between *p* and *I believe p* is sanctioned) unarticulated.

To be clear, none of this amounts to a knockdown argument against the view that the dual reasoning occurs unconsciously. The point here is simply that given what we know

[39] This point also relates to and is further supported by Kent Bach's (1984) work on default reasoning.

about how judgment-forming systems operate, it is less likely that this view is correct, and more tenable that the mentioned aspects of the dual reasoning remain suspended.

The components of the initial processing that will still occur are then only (a) the transition from the question of whether one believes *p* to the question as to whether *p*, (b) one's settling whether *p*, (c) the judgment about whether *p*, and (d) the self-ascription of the belief about *p*. Components (a)-(d) of the dual reasoning are not only executed but also present in consciousness, for the initial query as to whether one believes *p* is raised at the conscious level and settling whether *p* will require considering different kinds of information depending on the different values of *p*, making answering whether *p* a matter of central cognition, i.e., global-workspace, conscious processing. All other parts of the dual reasoning, however, remain implicit in applications of TM and one moves directly from *p* to the self-ascription *I believe p*.

I shall refer to the formation of self-ascriptions of beliefs that involves such a direct transition as an *unreflective self-attribution*. And I shall refer to the formation of self-ascriptions of beliefs in which the dual reasoning is fully executed and (provided this happens consciously) TM becomes intelligible to its user as a *reflective self-attribution*.

The distinction between unreflective and reflective self-attributions helps tackle objection (4), i.e. the point that most subjects using TM seem to move from *p* directly to the self-ascription *I believe p* without any inference and idea about the dual reasoning mentioned. This is because, according to the account introduced, when the dual reasoning occurs and puts into place the structure for a direct transition from *p* to *I believe p* then this happens unconsciously and only once. Since subsequent applications of TM lead, due to computational constraints, furthermore only to unreflective self-attribution it is to be expected that subjects using TM are ignorant of the dual-reasoning basis of the method.

Objection (5)

It might seem that there remains a crucial problem with the dual-reasoning account of TM. For it might be argued that to be able to explain both unreflective and reflective self-attributions, the account still needs to assume some mechanism that detects the judgment *p* and produces the self-ascription *I believe p* on the basis of that judgment.

Since the proposition *p* might be judged, supposed, doubted, etc., the mechanism will need to be able to track the judgment *p* and differentiate it from these other attitudes. Otherwise, it can't produce the self-ascription *I believe p* on its basis. But then, the objection continues, the transition involved in applications of TM will arguably already presuppose a representation of attitudes, e.g., judgments, and the account introduced fails to meet (AC1).

Response

This objection is a common argument against TM-based theories of self-knowledge (see, e.g., Goldman 2006: 240, 2012: 419; Carruthers 2011: 82f; Paul 2014: 299). However, the DA view (see Chapter 2), which I shall henceforth treat as a component of the dual-reasoning account, offers a plausible response to it. For according to the DA view, one has direct access to judgments in conscious first-order thinking and this involves the operation of a sub-personal attitude-type tracking mechanism which detects and differentiates the attitude types of the representations tokened without representing any attitude. The mechanism at issue is able to do so via the tracking of neural activation patterns that occur when the representations are tokened.

I propose that the same mechanism is in the context of TM responsible for detecting the judgment *p*. When upon reflection on whether *p*, one concludes that *p*, then this first-order attitude-type tracking mechanism enables the system responsible for the transition from the judgment *p* to the self-ascription *I believe p* to identify the judgment *p* via the neural properties of the representation about *p*. On the basis of these properties, the system can then automatically produce the self-ascription of the belief *p*. Since the mechanism that the DA view describes is able to detect and differentiate attitudes without already representing them and since the dual-reasoning account can appeal to this mechanism to explain how in applications of TM the judgment *p* is tracked, the account is able to avoid objection (5).

1.6 Summary

We now have an explanation of how TM works according to which the transition from *p* to the self-ascription *I believe p* that is part of TM is sustained by a piece of practical and theoretical reasoning. Both the practical reasoning and the theoretical reasoning are

consistent with one's holding that whether or not *p*, this won't tell one whether one believes *p*. The intelligibility puzzle is thus solved. Furthermore, the proposal doesn't presuppose that one is able to represent one's own mental processes or states without explaining this ability. Hence, the account meets (AC1). Finally, according to the account, most components of the dual reasoning that sustain TM are normally not executed (unconsciously or consciously). Rather, when the reasoning has been performed once (at the unconscious level), they are in subsequent applications of TM, due to processing constraints under which the cognitive system operates, taken for granted and remain implicit in the transition involved in TM. Upon asking oneself whether *p*, one will then proceed from the outcome of one's reflection on whether *p* immediately to the self-ascription *I believe p*. This does justice to the intuition that the transition involved in TM is direct. It is also why I shall call the account of TM introduced the *implicit dual-reasoning* (IDR) theory of TM.

2. Solving the knowledge puzzle

If the IDR theory is to be an account of self-knowledge of beliefs then it also needs to meet (AC2). That is, it needs to explain how self-ascriptions formed in the way just described can amount to knowledge. I will now offer such an explanation to supplement the IDR theory.

As noted above, I assume that a judgment or belief *p* counts as knowledge if it is justified and true. The answer to the question as to whether and how TM-based self-ascriptions of beliefs can be knowledge then depends on the account of epistemic justification that one adopts. I will assume the following externalist view.

- S is justified in forming a belief *p* if she (a) forms the belief because she is undergoing a mental event *M*, and (b) *M* reliably correlates in S with the type of state of affairs that makes her belief *p* true.

For instance, if S forms the belief (i.e. judges) that there is a cat on the sofa because she sees a cat on the sofa, and if her visual experience about the cat on the sofa reliably correlates with there actually being a cat on the sofa, then her subsequent belief that there is a cat on the sofa meets both (a) and (b), and is hence justified.

It seems to me that the combination of (a) and (b) yields a plausible notion of justification. Internalists will of course find the proposal problematic. But, as mentioned above, the question of whether internalism or externalism is true is still a matter of debate (see, e.g., Kornblith 2001; Conee and Feldman 2001; Bonjour and Sosa 2003; Goldman 2009). I won't delve into the debate here. Instead I want to apply the view of justification just introduced to TM-based self-ascriptions of beliefs to see whether it helps solve the knowledge puzzle.

Given (a) and (b), in the context of TM, for the self-ascription *I believe p* to be justified, the self-ascription will need to be formed on the basis of a mental event that reliably correlates with the state of affairs that makes the self-ascription true. The candidate event will be the conclusion of one's reflection on whether *p*, which is a judgment *p*. Since in the context of TM, by assumption, the self-ascription *I believe p* is formed because one undergoes the event of judging *p*, condition (a) is automatically met. In order for (b) to be met also, one's judgment *p* needs to reliably correlate with the state of affairs that makes the self-ascription *I believe p* true, i.e. with one's having the belief *p*. While not all judgments issue into beliefs with the same content, the claim that judgments normally give rise to and hence reliably correlate with beliefs is widely accepted (e.g., Peacocke 2003; Cassam 2010; Carruthers 2011). Furthermore, frequently, when upon wondering whether *p*, one swiftly recalls that *p*, then that will be a judgment too. For it will be an episode of recalling or bringing to consciousness a pre-existing belief. In these cases, the correlation between a judgment and the corresponding belief clearly holds. Given this, it should be uncontroversial that in the context of TM the judgment *p* does indeed reliably correlate with the truth-maker of the self-ascription *I believe p*. In the context of TM, the formation of the self-ascription *I believe p* thus meets both (a) and (b), and we now have an account of how TM-based self-ascriptions can be justified.

Moreover, since these self-ascriptions will typically be true, we have an account of how TM-based self-ascriptions can qualify as self-knowledge of beliefs. The proposal supplements the IDR theory of TM, which deals only with the intelligibility of the method. Hence, the IDR theory isn't just an account of TM but an account of self-knowledge of beliefs.

3. Are TM-based self-ascriptions inferential?

If the IDR theory is an account of self-knowledge of beliefs, what is the nature of the self-knowledge that it is an account of? Is self-knowledge acquired via TM in the way the account proposes *inferential* or *non-inferential*?

Above I distinguished between reflective and unreflective self-attribution. Reflective self-attributions lead to inferential self-knowledge. But I shall now argue that unreflective self-attributions, i.e. typical applications of TM, result in non-inferential self-knowledge of beliefs.

Before considering whether these self-attributions are inferential or not, it is useful to draw a distinction between two different senses of 'non-inferential'. Self-knowledge might be *psychologically* or *epistemically* non-inferential. Self-knowledge of a belief *p* is *psychologically* non-inferential if the acquisition of it doesn't involve any kind of conscious or unconscious inference from evidence pertaining to mental states of affairs. And self-knowledge of a belief *p* is *epistemically* non-inferential if one's justification for believing that one believes *p* doesn't come, even only partly, from one's having justification to believe other, supporting, propositions (Cassam 2011).

With this distinction in mind, does TM-based self-knowledge, acquired in the way the IDR theory proposes, amount to inferential or non-inferential knowledge? According to the IDR theory, unreflective self-attributions rely on conscious inferences about whether *p*. However, these inferences don't in fact threaten the view that TM-based self-ascriptions are psychologically non-inferential. For, by assumption, they are all first-order in nature, i.e. they don't pertain to anything mental but only to whether *p*, which is a non-mental state of affairs. Since only non-mental facts figure in the conscious reasoning on whether *p* which don't support any claims about anyone's having any belief, whatever justification the beliefs that play a role in the conscious reasoning on whether *p* might have, it won't be a justification for the self-ascription *I believe p*. As a result, the conscious inferences involved in settling whether *p* also don't undermine the view that unreflective self-attributions are psychologically immediate.

Are they *epistemically* inferential though? Unreflective self-attributions are epistemically non-inferential just in case one's justification for believing that one

believes *p* doesn't come from one's having justification to believe other, supporting, propositions. Now, according to the IDR theory, in unreflective self-attributions, when I proceed from *p* to the self-ascription *I believe p*, then what justifies the self-ascription is that (i) it is formed on the basis of the judgment *p*, and that (ii) the judgment *p* reliably correlates with the belief *p*.[40] That is, even if all my other judgments and beliefs lacked any justification, as long as the self-ascription of the belief *p* is formed on the basis of the judgment *p*, and the judgment reliably correlates with the belief *p* (which it does, even if it is completely unsupported), the self-ascription would still be justified and knowledge.[41] Since, according to the IDR account, my unreflective self-attribution doesn't involve any kind of conscious or unconscious inference from evidence pertaining to mental states of affairs, and since my justification for believing that I believe *p* doesn't come, even in part, from my having justification to believe other, supporting, propositions, it follows that unreflective self-attributions are both psychologically and epistemically non-inferential.

In response, it might be argued that, on the IDR theory, the direct move from the judgment *p* to the self-ascription *I believe p* surely still remains an inference. And if that is so, then, according to the IDR theory, unreflective self-attributions will be inferential after all.

However, note first that in general, for a transition from *p* to *q* to count as an inference, it isn't sufficient for *p* to merely cause *q*. Boghossian (2014) mentions a nice example to illustrate this:

> Suppose I see Aline. This causes me to believe that I see Aline, which causes me to drop the coffee I had been holding, which causes a stain on my shirt, which leads me to believe that my shirt is stained. My belief that I see Aline is part of the

[40] Cassam (2015) holds that "for my self-knowledge to qualify as non-inferential in the epistemological sense it would have to be the case that the justification for my second-order belief doesn't come from my justification for believing any other proposition" (11). He goes on to argue that in applications of TM, this isn't the case, and concludes: "TM itself only delivers inferential knowledge" (2015: 13). It is worth noting that in virtue of tying the justification of TM-based self-ascriptions to (i) and (ii), the IDR theory qualifies as an account of non-inferential self-knowledge even if one adopts Cassam's notion of 'non-inferential'. Cassam is thus wrong in his claim that TM only delivers inferential knowledge.

[41] It might be argued that a judgment based on an entirely unjustified premise can't amount to knowledge (Harman 1973: 47), but see Byrne (2011: 206-207) for a response to deal with this worry.

> causal explanation for why I believe that my shirt is stained. But we wouldn't want to say that I inferred that my shirt is stained from the fact that I see her. (3)

To specify what is missing for a causal transition between thoughts to qualify as an inference, Boghossian goes on to propose that "S's inferring from *p* to *q* is for S to judge *q because* S *takes* the (presumed) truth of *p* to provide support for *q*" (2014: 4).[42] On this account,[43] for my transition from the judgment *p* to the self-ascription *I believe p* to count as an inference, I must arrive at the self-ascription in part because I take the presumed truth of *p* to provide support for the self-ascription.

However, in unreflective self-attributions, this is not the case. For if I were asked whether *p* supports the self-ascription of a belief *p*, I will acknowledge that it doesn't do so because it doesn't follow that if *p* then anyone believes *p*. Since, by my own lights, *p* doesn't provide support for the self-ascription *I believe p*, the transition involved in unreflective applications of TM is, given the Boghossian account of inference, not inferential in nature.

Indeed, according to the IDR theory, the transition is purely causal. In the context of TM, the judgment *p* causes the self-ascription *I believe p* via the operation of a sub-personal attitude-type detection mechanism without any "element of taking"[44] that might be viewed as "essential to inference" (Boghossian 2014: 15). The direct transition from *p* to the self-ascription *I believe p* hence doesn't make unreflective self-attributions inferential. I conclude that the unreflectively acquired self-knowledge that the IDR theory describes is non-inferential in nature.

To be fair, the inferences that the IDR theory invokes to explain the transition involved in TM might suggest otherwise. But as noted, these inferences pertain specifically to *reflective* self-attributions. And reflective self-attributions are only one of two ways in which TM might, on the IDR view, be used. Reflective self-attributions and the dual

[42] Note that Boghossian's proposal pertains to person-level, conscious thinking. In the context of TM, the transition from *p* to the self-ascription *I believe p* is an instance of such thinking. Hence, his proposal on what an inference is can be applied to the issue at hand.

[43] Wright (2014) argues that Boghossian's 'taking' condition isn't required for an inference. But see Hlobil (2014: 428) for problems with Wright's response.

[44] See Boghossian (2014) for an account of the 'taking' at issue here.

reasoning introduced are meant to show how TM can be intelligible and be set up in the first place. For instance, the assumption of (IDA) is on the IDR view crucial for explaining how subjects could come to ask whether *p* to find out whether they believe *p* and move from *p* to the self-ascription *I believe p*. But even though the assumption plays the role of a mediating premise in the dual reasoning that puts in place and sustains TM, neither the assumption nor the reasoning contribute to the *justification* of one's self-ascriptions formed via TM. Note too that from the fact that inferences are involved when a particular mechanism for the transition between thoughts is put in place for forming self-ascriptions of beliefs, it doesn't follow that the outputs of that mechanism are then themselves inferential in nature. According to the IDR theory, they aren't. Rather, when the dual reasoning is executed once, information-processing constraints will lead the cognitive system to establish a simple detection mechanism that operates non-inferentially.

In sum, then, the intelligibility and origin of TM and the justification of the judgments and beliefs that the method leads to should be kept separate. Once they are kept separate, it isn't difficult to see that self-knowledge acquired via TM in the way the IDR theory proposes remains (in unreflective self-attributions) non-inferential in nature both in the psychological and the epistemic sense.

4. Privileged self-knowledge revisited

The discussion in the preceding section has immediate consequences for the question of whether the IDR theory can explain privileged self-knowledge of beliefs. I will now argue that given the points just made, the account is indeed able to do so.

Privileged self-knowledge of beliefs has two features: *authoritativeness* and *immediacy*. The authoritativeness of self-knowledge of beliefs is illustrated by the fact that one's judgments about one's own beliefs seem to be more strongly justified and more likely to amount to knowledge than other people's judgments about them. The immediacy of self-knowledge of beliefs has to do with the apparent fact that typically one seems to be able to know one's own beliefs without any observation, interpretation, inference or evidence that one has the beliefs.

The IDR theory explains both the authoritativeness and immediacy of self-knowledge of beliefs by proposing a particular account of TM. The account of TM that the theory

offers shows that and how TM-based self-knowledge of beliefs can be non-inferential. Further, even in cases of reflective self-attributions, which are inferential, no self-observation or interpretation is required. Hence, the IDR theory explains the immediacy of self-knowledge of beliefs. As a by-product, it also accounts for the authoritativeness of self-knowledge of beliefs. For other people can only work out, e.g., my beliefs by observing me and drawing inferences from my behaviour, the context, circumstances, etc. Since that is so, their judgments about my beliefs will be susceptible to errors that my unreflective self-attributions aren't susceptible to. This explains why my unreflectively formed TM-based self-ascriptions are more strongly justified than other people's ascriptions of beliefs to me. That is, it explains the authoritativeness of self-knowledge of beliefs. Since the IDR theory explains both the authoritative and the immediate nature of self-knowledge of beliefs, it explains privileged self-knowledge of beliefs.

It is worth emphasising that it is able to do so in a way that captures the attractiveness of TM. As noted above, TM is attractive because it suggests that self-knowledge of beliefs can be explained by appealing only to resources already required in other domains. The IDR theory captures this point because it holds that the formation of self-ascriptions of beliefs only relies on one's having the belief-concept, the self-concept, and the ability to engage in first-order (practical/theoretical) reasoning and the production of other-ascriptions of beliefs. As a result, the IDR theory offers an explanatorily and ontologically economical account of privileged self-knowledge of beliefs.

5. The IDR theory vs. other transparency theories

The preceding section has brought out one reason for endorsing the IDR theory, namely its ability to explain privileged self-knowledge of beliefs in an attractively parsimonious way. There are more. To make them explicit, I will now compare the theory with the transparency accounts discussed in the previous chapter.

There is a general point worth highlighting upfront. What seems to me to be the most significant difference between the proposals discussed in Chapter 3 and the IDR theory is that the latter offers a more detailed analysis of TM than any of the currently available accounts. It might be argued that the theory offers in fact an implausibly complicated explanation of TM. But note that the theoretical footwork done in the preceding sections allows the IDR account, unlike any other transparency view, to meet both (AC1) and (AC2). In doing so (as part of meeting (AC1)), it also provides a

defence of TM as a means for acquiring self-knowledge of beliefs. The complexity that the IDR theory involves is thus arguably what is needed to solve the explanatory problems surrounding TM. Apart from the issues of complexity and explanatory depth, there are other important differences between the IDR theory and the alternatives discussed. To show this, I will briefly revisit each of the transparency theories mentioned in Chapter 3.

The rationality view

The IDR theory, unlike the rationality view proposed by, e.g., Gordon (2007), offers an explanation of why Moore paradoxical statements such as '*p*, but I don't believe *p*' seem incoherent to us. These statements strike us as incoherent because once we have performed the dual reasoning, we have an understanding of the point that when *p* is the case then we believe *p*. This understanding is itself typically unconscious (leading to the intelligibility puzzle and the puzzle of why Moore statements appear absurd to us), but it can be made explicit via reflection along the lines mentioned in the preceding sections. Furthermore, according to the IDR account, unlike on the rationality view proposed by, e.g., Finkelstein (2012) and Cassam (2014), in order to be able to form TM-based self-ascriptions, one needn't make problematic assumptions such as: if *p* is true, then one ought to believe *p*.

The consciousness-based view

The IDR theory also differs from the consciousness-based view (e.g., Peacocke 2003; Silins 2012; Smithies 2012), according to which S's reason for forming the self-ascription *I believe p* on the basis of *p* is her phenomenally conscious judgment *p*, in that it doesn't assume the existence of phenomenally conscious judgments. This is a good thing because the assumption turned out to be controversial. Also, on the consciousness-based view as defended by, e.g., Peacocke (2003), in applications of TM, the judgment *p* is marked as a judgment at the conscious level. It is, on this view, only after the judgment is experienced that the self-ascription *I believe p* is formed. The phenomenology of the judgment is meant to serve the subject as a reason for the self-ascription. In contrast, according to the IDR theory, a sub-personal mechanism tracks the judgment via the vehicle properties of the representation *p* that is tokened when one concludes that *p*. What one is at the conscious level aware of as the basis for the

transition from the judgment *p* to the self-ascription *I believe p* is only the worldly fact *p*. This is similar to Valaris' (2014a) consciousness-based view. However, unlike his proposal, the IDR theory doesn't presuppose the ability to view *p* as one's own take on the fact rather than as the fact *p* itself.

The mental-action view

The IDR theory is also different from the mental-action view in that it doesn't assume the existence of some kind of mental-action awareness of judging that *p* (Peacocke 2008). And even though it holds that decisions have an epistemic dimension, which makes the proposal similar to Soteriou's (2013) and Roessler's (2013, 2015) views, unlike their accounts, the IDR theory manages to explain how the transition from *p* to the self-ascription *I believe p* works without presupposing a self-ascription of mental processes or states. It is able to do so because, in contrast to Soteriou's, and Roessler's proposals, the IDR theory holds that self-ascriptions of beliefs aren't only based on one's insight into what one will do (which is attained via one's decision and intention to do it). Rather, they are also based on a combination of this insight with a truth-bias, and the general principle that what people will say in response to the question of whether *p* is what they believe about *p*.

The world-mind inference view, and the bypass view

Turning now to the externalist proposals discussed, while advocates of the world-mind inference view (Byrne 2005, 2011) hold that the transition from *p* to the self-ascription *I believe p* is self-verifying, the IDR theory allows for false self-ascriptions of beliefs. For, according to the account, the self-ascription *I believe p* is based on a judgment *p* and judgments don't always lead to beliefs. Another difference is that whereas the world-mind inference view claims that the formation of self-ascriptions of beliefs via TM involves an "inference" (Byrne 2005, 2011), the IDR theory allows for non-inferential self-attributions because it holds that there are not only reflective but also unreflective self-attributions. Further, more importantly, the IDR account offers an explanation of how the transition from *p* to the self-ascription *I believe p* could be set up in the first place and be intelligible to the subject engaging in it even when she doesn't already represent any aspect of her own mind as such. The world-mind inference view

doesn't manage to do so. This point makes the IDR theory also significantly different from the bypass view (Fernández 2013).

The constitution view, and the commitment view

As for the non-epistemic proposals that I mentioned, the IDR theory explains, whereas the constitution view (e.g., Boyle 2011; Shoemaker 2012) leaves open, how the second-order, self-ascribing beliefs, that are on the constitution view constitutive of first-order beliefs, can come to be formed in the first place. Moreover, unlike both the constitution and the commitment views (e.g., McGeer 1997, 2008; Moran 2001, 2012), the IDR theory is an epistemic account. It construes and defends TM as a means for acquiring knowledge of a pre-existing mental condition of oneself. Relatedly, e.g., on McGeer's view, when one self-ascribes a belief *p*, one promises to make it the case that one believe *p*, which is compatible with one's holding that one doesn't yet believe *p*. I argued that this is at odds with the datum that when one self-ascribes a belief *p*, one takes it that one does indeed already believe *p*. In contrast to McGeer's view, the IDR theory is able to capture this point. Finally, the IDR theory doesn't, unlike, e.g., Moran's proposal, presuppose any kind of self-ascription of mental processes or states without explaining it. The IDR theory is hence clearly distinct from and preferable to the commitment view.

6. Conclusion

In this chapter, I developed a transparency theory of self-knowledge of beliefs. The account, the IDR theory, is able to meet both conditions of adequacy that I introduced in the previous chapter. As argued, none of a wide range of extant transparency proposals is able to do so. Indeed, as far as I can tell all the currently available accounts of this kind have this problem. The IDR theory hence fills a significant gap in the literature. I noted that the account also manages to avoid many other shortcomings that bedevil extant competitors of its kind and explains, in an attractively economical way, an aspect of self-knowledge of beliefs that is philosophically especially puzzling, namely that it is at least sometimes privileged in nature.

CHAPTER 5

Transparency-independent theories

So far I have only discussed accounts of self-knowledge of beliefs that invoke the transparency method. I argued that the IDR theory is preferable to extant proposals of this kind. There are, however, other theories of self-knowledge of beliefs that don't appeal to the transparency method. According to these accounts, one can come to know one's own belief *p* without first having to engage in conscious first-order thinking to determine whether *p*. I shall refer to them as *transparency-independent theories.* Two different views can be distinguished when it comes to these proposals. There are what I shall call the *asymmetry view* and the *symmetry view*.

According to the asymmetry view, self-knowledge of beliefs is typically independent from other-knowledge of them in that it is developmentally prior to and involves a different kind of mechanism than other-knowledge of beliefs. Unlike other-knowledge of beliefs, self-knowledge of beliefs is acquired without any kind of observation and interpretation of behaviour. One common account that belongs to the asymmetry view is what I shall call the *inner-scanner theory.* According to this proposal, when one wonders whether one believes *p*, one doesn't start thinking about whether *p*. Rather, a sub-personal mechanism is activated that scans or searches one's mind for the presence of a belief *p*. If it detects the belief, then it produces a self-ascription of the belief. Since no inference is thought to be involved in this process (e.g., Armstrong 1968, 1999; Goldman 1993; Lycan 1996; Nichols and Stich 2003; Goldman 2006), the ability to self-ascribe beliefs is on the inner-scanner theory fundamentally different from the ability to other-ascribe them.

In contrast, on the symmetry view, the ability to self-ascribe beliefs and the ability to other-ascribe them are dependent in that the former is developmentally posterior to or simultaneous with the latter. Both also involve the operation of the same kind of mechanism. That is, self-knowledge of beliefs is, just as other-knowledge of them, thought to be dependent on observation, interpretation, and inferences (e.g., Ryle 1949; Dennett 1992; Gopnik 1993; Gazzaniga 1995; 2000; Stephen and Graham 2000; Wilson 2002; Dretske 2003; Cooper 2007; Lawlor 2009; Williams and Happé 2010; Carruthers 2009a, 2011; Mandelbaum 2014). Arguably the best-developed version of the symmetry

view is the *interpretive sensory-access theory* (Carruthers 2009a, 2011). It holds that we come to know our own attitudes by interpreting ourselves (e.g., our own behaviour, circumstance, or sensory-imagistic states) and employing the same faculty that we use to work out other people's mental states, namely the mindreading system.

In this chapter, I have two goals. The first one is to further defend and develop the IDR theory by relating it to the just mentioned two transparency-independent views. I argue that the two most recent versions of the inner-scanner theory, i.e. Nichols and Stich's (2003), and Goldman's (2006) accounts, are problematic, for they leave unexplained how the inner scanner or detection mechanism that they postulate can come to classify a first-order representation that it detects as, e.g., a belief in the first place. Furthermore, both of the accounts produce predictions that are contradicted by the empirical data. I contend that the IDR theory is able to avoid both problems. After that, I turn to the interpretive sensory-access theory and show that this account too is problematic. For it doesn't satisfactorily explain our access to attitudes in first-order reasoning, and its proposal on how we acquire self-knowledge of beliefs that we can draw on in conscious thinking is at odds with what we know about the processing constraint under which the cognitive system operates. The IDR theory is preferable when it comes to dealing with both of these points. The argument against the interpretive sensory-access theory and for the IDR theory will bring out that the findings that are often taken to undermine the existence of privileged self-knowledge of attitudes (e.g., evidence on self-interpretation and confabulation) don't in fact threaten the existence of privileged self-knowledge acquired in the way the IDR theory proposes. In fact, I contend that once the way in which the IDR theory explains the acquisition of self-knowledge of beliefs is integrated with empirical data on how judgment-forming systems work, an argument that supports the existence of privileged self-knowledge of beliefs emerges.

My second goal in this chapter is to embed the IDR theory, which covers only self-knowledge of a sub-class of beliefs, into a general account of self-knowledge of attitudes. I will do so by arguing for a hybrid view that combines the IDR theory with a revised version of the interpretive sensory-access theory. The resulting proposal conjoins the asymmetry view and the symmetry view. It holds that self-knowledge of attitudes is developmentally posterior to and relies on the same mechanism as other-knowledge of attitudes but when it comes to beliefs and TM is used, self-knowledge nonetheless remains privileged in nature.

In section 1, I introduce and critique the two just-mentioned inner-scanner accounts. I scrutinise the psychological findings that have been claimed to support them and compare the accounts with the IDR theory. In section 2, I introduce and critique the interpretive sensory-access theory. Section 3 then develops the mentioned hybrid view.

1. The inner-scanner theory – Two proposals

Different versions of the inner-scanner theory have been proposed. I will focus on the two most recent, empirically motivated ones only. They have been developed by Nichols and Stich (2003), and Goldman (2006).

Nichols and Stich (2003) distinguish between *detecting* mental states and *reasoning* about them. Detecting mental states is the capacity to attribute current mental states to someone. Reasoning about them is the "capacity to use information about a person's mental states (typically along with other information)" to explain and "make predictions about the person's past and future mental states, and behaviour" (Nichols and Stich 2003: 152).

Detecting and reasoning about mental states may pertain to one's own or other people's mental states, and they involve different mechanisms. Nichols and Stich hold that third-person detection of mental states and first- and third-person reasoning about one's own and others' mental states require the operation of the theory-of-mind (ToM) or mindreading system. The detection of one's own mental states, e.g., of one's own beliefs, however, doesn't. It is performed by

> a Monitoring Mechanism (MM) (or perhaps a set of mechanisms) that, when activated, takes the representation *p* in the Belief Box as input and produces the representation *I believe that p* as output. […] To produce representations of one's own beliefs, the Monitoring Mechanism merely has to copy representations from the Belief Box, embed the copies in a representation schema of the form: *I believe that* ___, and then place the new representations back in the Belief Box. (Nichols and Stich 2003: 160-161)[45]

[45] The term 'Box' refers to a functional role type in a block diagram.

Nichols and Stich argue that the MM operates at the sub-personal level and produces self-ascriptions of mental states without any kind of inference. In the latter respect, it is distinct from the mechanism producing other-ascriptions of attitudes.

Nichols and Stich emphasise, however, that the MM isn't the only mechanism that is able to produce self-ascriptions of attitudes. The ToM system can do so as well. It becomes operative whenever reasoning about one's own mental states is required, e.g., when one is trying to determine the causes of one's own behaviour (Nichols and Stich 2003: 162).

While both the ToM system and the MM might produce self-ascriptions of attitudes, Nichols and Stich hold that the MM is not only different from the ToM system in that the former is responsible for the detection of mental states whereas the latter is responsible for reasoning about them, the MM is also an "innate cognitive mechanism" that comes online "significantly before ToM is fully in place" (2003: 194, 163). Indeed, the two are thought to be entirely independent allowing for two-way dissociations in which the MM and so the ability to detect one's own attitudes is intact, while the ToM system and so the ability to detect or reason about other people's attitudes is impaired, and vice versa.

Nichols and Stich claim that studies from (i) developmental psychology and data on (ii) autism and (iii) schizophrenia provide evidence for the dissociations that their account predicts. The findings suggest, Nichols and Stich hold, that both normally developing young children and autistic subjects have difficulty tracking others' mental states but not their own. Reversely, in passivity-symptomatic schizophrenics, other-ascriptions of mental states appear intact but self-ascriptions of them are impaired.

The data pertaining to (i)-(iii) are the main support for Nichols and Stich's monitoring-mechanism theory. I will discuss them below. Now I just want to register a more general worry about their account. As it stands, it doesn't explain how the MM is able to differentiate and conceptualise representations according to their different attitude types. As noted, the MM's task is to "copy representations from the Belief Box, embed the copies in a representation schema of the form: *I believe that* ___, and then place the new representations back in the Belief Box" (Nichols and Stich 2003: 161). The same holds for other attitudes, e.g., desires, hopes etc. The only aspect that changes is the

'Boxes', and the first part of the representation schema. If that is so, however, then how is the MM able to determine and classify the attitude type of a particular representation? How does it 'know' that the content of a representation *p* is *believed* as opposed to, say, supposed or doubted? Nichols and Stich don't say. Their view thus remains crucially incomplete.

One response that comes to mind here harks back to Chapter 2. It is that representations might have attitude-type specific neural properties that the MM could then track and use to determine whether a particular representation carries content that is believed rather than, e.g., supposed or doubted.

Goldman (2006) has developed this suggestion into his own kind of inner-scanner account. He agrees with Nichols and Stich that subjects have a "special method of accessing or detecting their current mental states" which is operative prior to and dissociable from the mechanism involved in working out other people's mental states (Goldman 2006: 224). And he too supports this claim by using the three kinds of empirical studies that Nichols and Stich cite.

However, unlike Nichols and Stich, Goldman holds that the method for forming self-ascriptions of mental states involves what he calls "introspection", an "inner recognition" of mental states based on their intrinsic properties, which is similar to the perceptual processes of seeing or hearing objects or sounds (2006: 246). He thinks that just as perception, introspection involves a transduction mechanism, which is causally sensitive to a particular kind of input and produces, on that basis, a particular kind of output. The outputs of the transduction mechanism are "representations of token mental states that classify them" according to "(1) the general category of the token state (e.g., belief, desire, pain, anger, heat sensation, visual representation), (2) the content of the state, and (3) the strength or intensity of the state" (Goldman 2006: 246). As for the inputs, when it comes to determining, e.g., a mental-state type, the transduction mechanism uses neural-activation patterns that are detected by a sub-personal attention network. For instance, a high level of activation in one class of neural cells "generates the introspective classification 'pain' (or 'sharp pain'), a high level of activation in a different class of cells generates the introspective classification 'tickle,'" and so on for other mental states including attitudes such as beliefs and desires (Goldman 2006: 252).

There are commonalities here between Goldman's proposal and the IDR theory. I will come back to them below. Now I shall just note the following shortcoming of his proposal. On Goldman's view, for a self-ascription of the belief *p* to be formed, the belief *p* must be tokened in first-order reasoning because otherwise the neural-activation pattern associated with the belief won't be present for the attitude-type detection mechanism to latch onto. Suppose then my belief *p* now figures in my first-order reasoning and the neural activation at issue is occurring. How can the attitude-type detection mechanism take the neural activation pattern to be indicative of my believing rather than, say, my supposing that *p*, and produce on its basis the self-ascription *I believe p*? The link between the neural-activation pattern and the second-order belief must somehow, e.g., via learning, be established. The problem is that the connection will arguably need to be established while I'm only aware of *p* and the belief *p* figures in first-order reasoning. How can the connection at that moment be set up given that I understand that when *p*, it doesn't follow that anyone also believes *p*? The problem here is a version of the intelligibility puzzle. Goldman offers no explanation that could provide a solution. It thus remains unclear on his view how a meta-representation[46] such as *I believe p* can be formed on the basis of neural properties instantiated when *p* figures in first-order thinking.

Both Goldman's, and Nichols and Stich's accounts rest on the assumption that one can form a meta-representation, a self-ascription *I believe p*, on the basis of a first-order representation *p*, but they don't say how the classification of *p* as a belief can come about. Both proposals are hence in an important respect incomplete. There is another problem with them. It pertains to the empirical data that bear on the predictions that the proposals produce.

1.1 The data

On both Nichols and Stich's, and Goldman's views, it is predicted that the ability to self-ascribe attitudes can become dissociated from the ability to other-ascribe them in that the former might be functional while the latter isn't and vice versa (see, e.g., Nichols and Stich 2003: 163; Goldman 2006: 224). I shall refer to this as the

[46] How attitude-type detection and differentiation via neural properties works in general is left open by the IDR theory too. This issue is not the problem here.

dissociation thesis. I shall refer to the negation of the dissociation thesis as the *parallelism thesis*.

As noted, Nichols and Stich claim that three kinds of findings support the dissociation thesis: data from (i) developmental psychology, (ii) studies on autism, and (iii) research on schizophrenia. I also mentioned that Goldman appeals to the same kind of results to support his own account. I will now take a critical look at the findings and argue that they don't corroborate the dissociation thesis but in some cases rather support the parallelism thesis.

To forewarn the reader, I will spend a bit of time discussing the empirical data pertaining to the dissociation thesis. This is because, as will become clear below, the viability of the IDR theory hinges in part on the falsity of that thesis.

1.1.1 Developmental studies

Nichols and Stich, and Goldman make much of two developmental studies in particular. The first was conducted by Wimmer et al. (1988) and pertains to children's ability to track what a subject knows on the basis of what she has seen (Nichols and Stich 2003: 175f; Goldman 2006: 236f). Wimmer et al.'s experiment involved two conditions. In the first one, children were shown a box and told to either look inside the box or refrain from looking inside it. Afterwards, they were asked 'Do you know what is in the box or do you not know that?' 3-year-olds did well on the task. In the second condition, they were then presented with a scenario in which another child either looked or didn't look into a box before they were asked: "'Does [name of child] know what is in the box or does she [he] not know that?'" (1988: 383) The children did worse on the second version of the task. Nichols and Stich, and Goldman hold that this shows that the ability to self-ascribe mental states, e.g., knowledge states, is dissociable from the ability to other-ascribe them.

However, there is an alternative interpretation of Wimmer et al.'s findings. In the study, children were asked double-barred questions: 'Does [name of the other child] know what is in the box, or does she [he] not know that?/Do you know what is in the box or do you not know that?' It might be that they found the other-related question in this set-up more complicated to process than the self-related one. If so, that would help explain

their poorer performance on the other-related task in a way that doesn't lend support to the dissociation thesis.

This interpretation of the findings is confirmed by a study conducted by Pratt and Bryant (1990). Pratt and Bryant ran the same kind of experiment as Wimmer et al., but unlike Wimmer et al., they used a simpler question format. They asked the children "'Does [name of the other child] know what is in the box?/Do you know what is in the box?'" (1990: 977) Pratt and Bryant found that 3-year-olds now performed just as well on the other-related questions as on the self-related ones. They then conducted another experiment to make a direct comparison between the simpler question and the more complex, double-barrelled question that Wimmer et al. used, and found that the children had much greater difficulty processing the more complex question. This undermines Nichols and Stich's, and Goldman's interpretation of Wimmer et al.'s data. And the first kind of findings that they cite for the dissociation thesis doesn't in fact support the thesis.

The second kind of developmental studies that both Nichols and Stich, and Goldman take to corroborate the dissociation thesis pertains to children's understanding of pretence. They cite, for instance, an experiment by Gopnik and Slaughter (1991) in which 3-year-olds first had to pretend that an empty glass contained orange juice. The experimenter then briefly turned the glass over and asked the child to pretend that it was now filled with hot chocolate. After that, the child was presented with the question "'When I first asked you […] what did you pretend was in the glass then?'" (Gopnik and Slaughter 1991: 106) Gopnik and Slaughter found that most of the children correctly identified what they had pretended. Nichols and Stich contrast this study with an experiment by Rosen et al. (1997), in which 3-year-olds were shown animation characters that were sitting on a bench but pretending to be on an airplane. The experimenters then asked the children "'Are they [the characters] thinking about being on an airplane or about sitting on a bench outside their school?'" (Rosen et al. 1997: 1135) The majority of the 3-year-olds gave the wrong answer saying that the character was thinking about sitting on a bench. Nichols and Stich, and Goldman argue that, taken together, these two studies suggest that 3-year-olds have no difficulty in self-ascribing pretence but do have problems ascribing pretence to others. This lends support to the dissociation thesis.

However, one problem with Nichols and Stich's, and Goldman's use of the data is that the two studies involved different groups of subjects, different experimental set-ups, and different task demands (Carruthers 2011). For instance, in Gopnik and Slaughter's study, the experimenters asked the child about what someone "pretended" (1991: 106), while in Rosen et al.'s study, children were probed about what a character was "thinking" (1997: 1135). Quite possibly, the 3-year-olds in Rosen et al.'s study found the thought-related question more difficult to understand than Gopnik and Slaughter's test subjects the pretence-related question. And if asked the pretence-related questions, they would have performed well (Carruthers 2011).

Indeed, in more recent studies similar to Rosen et al.'s experiment it turns out that 3-year-olds don't do poorly in detecting pretence in others. For instance, Ganea et al. (2004) presented 3-year-olds with a scenario in which subjects wanted to be like a kangaroo and either acted like a kangaroo or didn't act like it. The children were then asked what the subjects wanted to be like, how they were in fact moving like, and what they were pretending to be. Ganea et al. found that 3-year-olds performed significantly better than chance in identifying the pretence of the subjects.

Finally, in these kinds of experiments much may again hinge on how the experimenter phrases the test questions. For instance, Sobel (2007) presented 3-year-olds with characters who either intended to act (e.g., like a kangaroo) or incidentally acted like a kangaroo (e.g., out of joy the character jumped up and down). Sobel found that when the children were asked the standard question 'Is [the character] pretending?' they performed poorly, but when the procedure involved a forced choice, e.g., 'Which of the two characters is pretending?' they correctly chose the character that was pretending. It is fair to say then that the second kind of developmental data that Nichols and Stich, and Goldman mentioned don't support the dissociation thesis either.

In fact, there is counterevidence to the thesis. Consider children's understanding of beliefs. If the dissociation thesis were right, then when it comes to, e.g., children's grasp of their own and other people's false beliefs, meta-analyses of false-belief task studies should reveal that the ability to self-ascribe false beliefs can become dissociated from the ability to other-ascribe them. However, Wellman et al. (2001) conducted an

extensive meta-analysis of 178 distinct theory-of-mind studies involving self- and other-conditions,[47] and found that the

> essential age trajectory for tasks requiring judgments of someone else's false belief is paralleled by an identical age trajectory for children's judgments of their own false beliefs. Young children, for example, are just as incorrect at attributing a false belief to themselves as they are at attributing it to others. (665)

This challenges the dissociation thesis and corroborates the parallelism alternative.[48] I will now turn to the second kind of empirical findings that Nichols and Stich, and Goldman take to speak in favor of the dissociation thesis.

1.1.2 Autism

Nichols and Stich (2003: 184f), and Goldman (2006: 237f) hold that results from studies on autism too support the dissociation thesis. They note first that it is well known that subjects with autism tend to have an impaired understanding of other people's mental states. Nichols and Stich, and Goldman then proceed to mention studies that indicate that some of these subjects nonetheless retain an intact awareness of their own minds.

They focus on two kinds of studies in particular. One is based on introspective reports, and the other pertains to meta-memory. For instance, in introspective-sampling studies by Hurlburt et al. (1994), three adults with autism were asked to carry a beeper and describe their immediate experience whenever they heard the beep. The subsequent assessment of their reports indicated that some of the subjects were aware of what was happening in their minds. Similarly, autobiographical texts by high-functioning individuals with Asperger syndrome, which is a form of autism, suggest that they can recall their own experiences, wants, and beliefs from early childhood (see, e.g., Grandin 1984: 145ff; Dewey 1991: 204). Nichols and Stich take the introspective-sampling and autobiographical data to show that in "autism […] third-person mindreading is seriously

[47] This meta-analysis is frequently cited; e.g., Carruthers (2011) and Musholt (2012) mention the study also.

[48] For more evidence supporting a parallelism see, e.g., Rakoczy (2010), Carruthers (2011), and Musholt (2012).

defective, though first-person mental state detection is not significantly impaired" (2003: 189). Goldman (2006: 238) concurs.

However, this claim isn't supported by the data at issue. Note first that in the introspective-sampling study, the three autistic adults involved were in fact also asked to perform ToM tasks, e.g., false-belief tests. It turned out that the individuals who had difficulties in the ToM tasks exhibited a corresponding difficulty in the introspective-sampling task (Hurlburt et al. 1994: 394). To be fair, Frith and Happé (1999), who were Hurlburt's collaborators, write that in the study at issue, the subject with the least developed ToM skills was still able to report "current ongoing inner experience during interviews" (14). Nichols and Stich take this point in particular to support their MM account (2003: 187). However, Frith and Happé (1999) note too that the experiences reported were only "visually based", i.e. they were episodes of conscious visual imagery (12). The experiences weren't attitudes. Indeed, Hurlburt (2009) writes that none of the subjects in the introspection-sampling studies he has conducted ever confidently reported propositional attitudes. Since the dissociation thesis pertains to attitudes, the data at issue offers little support for the thesis.

The aforementioned autobiographical descriptions also don't support the dissociation thesis. Even if we ignore the point that the reports of the autistic subjects might have been *post-hoc* reconstructions or confabulations (Carruthers 2011), all of the subjects who provided the reports were high-functioning autistic individuals. This makes it likely that they had some understanding not only of their own but also of other people's mental states. Unfortunately, their theory of mind ability wasn't controlled for. Since that is so, for all we know, they also had some grasp of other people's mental states at the time during their childhood that their reports referred to. The reports thus don't yield support for the dissociation thesis.

But Nichols and Stich (2003: 188f) cite a second kind of autism study. It pertains to meta-memory performance in autism and was conducted by Farrant et al. (1999). Farrant et al. asked children with autism and neurotypical controls to memorise a set of items and afterwards to report on the mental methods that they employed to recall the items. Nichols and Stich write that, just like the "other children in the study, most of the autistic children answered this question with some explanation that adverted to thinking, listening, or exploiting a strategy", and, more generally, there was no significant

difference between the performance of the autistic and the non-autistic children on the meta-memory tasks (2003: 190). Nichols and Stich hold that these results

> fit perfectly with the Monitoring Mechanism theory. For the Monitoring Mechanism can be intact even when the mental mechanisms subserving third-person belief attribution are damaged. […] [T]hese findings […] indicate that people afflicted with autism do indeed manifest one of the patterns of dissociation. (ibid)

Proust (2013) too maintains, on the basis of Farrant et al.'s study, that "[c]hildren with autism have normal metacognition, but an impaired capacity for mindreading" (66).

However, such claims are in fact unsupported by the study at issue. For Farrant et al. note themselves that the autistic children in their study exhibited "essentially ceiling effects on the false belief tasks" (1999: 127). Evidently, the autistic children in the study didn't have a significantly impaired ability to ascribe mental states to others. Their successful performance on the meta-memory tasks hence doesn't contrast with an impaired ability to other-ascribe mental states. Furthermore, their good performance on the meta-memory tasks in fact correlated with a good ability to also keep track of other people's memory capacities. As Farrant et al. note:

> In Experiment 1, there was some slight evidence to suggest that the children with autism might be less knowledgeable than controls concerning other people's memory abilities. However, in Experiment 5 a more extended test of knowledge of others' memory abilities produced no evidence of impairment. (1999: 127)

That is, the apparently intact ability of the autistic children to track their own memory processes corresponded with an apparently intact ability to track other people's memory. Given this, there is no support in the study for the dissociation thesis whatsoever.

In fact, a number of findings on autism strongly contradict the thesis. For instance, many studies that specifically tested self versus other mentalising skills in autism have found that autistic subjects have significant problems in reflecting on both their own and other people's false beliefs and intentions (see, e.g., Baron-Cohen 1989; Perner et al. 1989; Kazak et al. 1997; Phillips et al. 1998; Williams and Happé 2009; Williams

and Happé 2010; Williams 2010). This has led some psychologists to suggest that in general "one cognitive mechanism (or process) is involved in both recognising one's own and others' mental states" (Williams and Happé 2010: 318). More recent data on meta-memory monitoring ('feeling-of-knowing') (Grainger et al. 2014), agency tracking (Zalla et al. 2015), and the self-other distinction (Lombardo and Baron-Cohen 2010; Lyon and Fitzgerald 2013) in autistic subjects with mindreading deficits lend further support to the parallelism thesis.

But Nichols and Stich, and Goldman discuss a third kind of data to argue for the dissociation thesis.

1.1.3 Schizophrenia

Nichols and Stich (2003: 190), and Goldman (2006: 238) claim that research on schizophrenics, especially on those with passivity symptoms, supports the dissociation thesis. Passivity symptoms include experiences of thought insertion and delusions of control. It has been argued that subjects with these symptoms have an impaired ability to monitor their own intentions to act (e.g., Frith and Done 1989; Corcoran et al. 1995; Frith and Corcoran 1996). Nichols and Stich agree and emphasise the further finding that these subjects are nonetheless able to ascribe intentions to others. They appeal to a set of experiments by Corcoran et al. (1995), and Frith and Corcoran (1996) in which "patients with passivity features (delusions of control, thought insertion, etc.) could answer" intention-specific "theory of mind questions quite well" (527). On the basis of these studies, Nichols and Stich write that "in schizophrenic subjects with [passivity symptoms], first-person mental state detection is severely impaired but third-person mindreading is not" (2003: 192). Goldman agrees (2006: 238).

However, there is again reason to doubt that the data shows what Nichols and Stich, and Goldman take it to show. For instance, while passivity symptoms are typically assumed to be indicative of an impaired ability to track one's own intentions to act, there is reason to believe that schizophrenics with these symptoms don't always suffer from them and hence don't always lack the ability to monitor their own intentions to act. For instance, Mlakar et al. (1994) write,

> it is only a small number of actions that [passivity-symptomatic] schizophrenic patients experience as being outside of their control. For example, none of the patients in our study reported that the drawings they made were produced by alien forces. (563)

This suggests that schizophrenics with passivity symptoms are sometimes able to monitor their own intentions to act. Until it has been shown that the same passivity-symptomatic schizophrenics who are in a particular study able to grasp the intentions of others are also at the same time unable to monitor their own intentions, which has not been established by any study so far, the data on passivity-symptomatic schizophrenics lends little support to the dissociation thesis.

Furthermore, even if we grant that passivity-symptomatic schizophrenics have a general deficit in monitoring their own intentions, it is questionable whether these subjects indeed also have a generally intact ability to keep track of other people's mental states. To assess whether or not this is so, it helps if we take a look at all the studies pertaining to the issue. The number of studies that specifically concern the ability of passivity-symptomatic schizophrenics to infer other people's mental states is relatively low. At this point, there are as far as I can tell only seven studies: (1) Corcoran et al. (1995), (2) Corcoran and Frith (1996), (3) Frith and Corcoran (1996), (4) Corcoran et al. (1997), (5) Murphy (1998), (6) Pickup and Frith (2001), and (7) Brüne et al. (2008).

The data on the deficit of passivity-symptomatic schizophrenics in monitoring their own mental states pertains only to intentions to act. There is no study probing their ability to monitor, e.g., their own beliefs. Since that is so, from the seven studies that concern their ability to infer other people's mental states, only the ones that address their ability to infer other people's intentions will be of use for anyone wishing to support the dissociation thesis. Since Murphy (1998) and Pickup and Frith (2001) only probe other-related false-belief understanding, this reduces the relevant studies to five: (1) Corcoran et al. (1995), (2) Frith and Corcoran (1996), (3) Corcoran and Frith (1996), (4) Corcoran et al. (1997), and (5) Brüne et al. (2008).

Brüne et al.'s study can be set aside because all the subjects they tested "responded well to antipsychotic treatment and were only mildly symptomatic at the time of testing", which suggests that their ability to monitor their own intentions to act was not severely

impaired (2008: 1992). The relatively normal performance on other-related ToM tasks that Brüne et al.'s study revealed lends thus little support to the dissociation thesis. With respect to the remaining four studies, the results are mixed:

(1) Corcoran et al. (1995) tested passivity-symptomatic schizophrenics on their capacity to infer intentions behind indirect speech and found that these subjects were "perfectly capable of inferring the intentions of others from indirect speech" (10).

(2) Frith and Corcoran (1996) presented passivity-symptomatic schizophrenics with cartoon stories which involved deception. In order to understand the behaviour of the characters in the stories, subjects had to infer the intentions of the characters. "Patients with symptoms of passivity (e.g. delusions of control) […] did not differ from normal controls" in their understanding of the stories (Frith and Corcoran 1996: 521).

(3) Corcoran and Frith (1996) tested passivity-symptomatic schizophrenics' understanding of the mental states of others by probing their appreciation of Grice's (1989) conversational maxims. Test subjects were given statements and answers to choose from. To correctly answer, they had to grasp the intention of the person asking the question. With respect to the overall maxims performance, passivity-symptomatic schizophrenics "obtained a lower composite score than did the normal control group" (Corcoran and Frith 1996: 313).

(4) Finally, Corcoran et al. (1997) presented passivity-symptomatic schizophrenics with cartoon jokes in which the main character was deceived by another character. To 'get' the joke, subjects had to grasp the intention of the deceptive character. Corcoran et al. found that in comparison to normal control subjects, the "schizophrenic patients" found the "jokes significantly more difficult to understand. This effect was most marked in patients […] reporting passivity experiences" (1997: 319).

Taken together, the findings are inconclusive when it comes to supporting the claim that passivity-symptomatic schizophrenics are able to work out other people's intentions but not their own. Studies (1) and (2) suggest that they are able to infer other people's intentions (these are the experiments that Nichols and Stich, and Goldman cite).[49] But

[49] It is worth noting that in study (2), which Nichols and Stich, and Goldman take to indicate that passivity-symptomatic schizophrenics are able to infer other people's intentions, the "patients with

studies (3) and (4) indicate that they have difficulties doing so. Given this, the currently available data bearing on the issue of whether passivity-symptomatic schizophrenics are able to work out other people's intentions but not their own intentions don't support general claims such as "[p]assivity-symptomatic schizophrenics, who are impaired at first-person mindreading, exhibit normal performance on third-person mindreading tasks" (Goldman 2006: 238).

The data suggests rather that sometimes these subjects are able to other-ascribe intentions but not at other times. Given the point made earlier that they seem sometimes also able to track their own intentions correctly but not at other times, there is little support for the dissociation thesis in the data. Indeed, the findings are in fact better in line with the parallelism thesis.

1.2 A comparison with the IDR theory

In the preceding discussion, I made two points against Nichols and Stich's, and Goldman's versions of the inner-scanner theory. I argued that their accounts are (a) incomplete in that they don't explain how the attitude-type detection mechanism that they postulate can come to conceptualise a first-order representation *p* as, e.g., a belief *p*. Furthermore, both accounts are (b) committed to the dissociation thesis, which is not supported but in some cases contradicted by the experimental findings. I shall now argue that the IDR theory manages to avoid both of these problems.

Consider (a). Recall that the IDR theory consists in part of the DA view, and so assumes, just as the mentioned inner-scanner accounts do, the existence of a sub-personal attitude-type detection mechanism. The DA view, and therewith the IDR

---

symptoms of passivity […] were also medicated", i.e. they were at the time taking neuroleptics (Frith and Corcoran 1996: 525). There is evidence that neuroleptics improve patients' ability to monitor their own actions and thoughts. For instance, Keefe et al. (2003) investigated the effect of antipsychotic medication on "source monitoring performance and its relation to the reduction of certain psychotic symptoms associated with the inability to identify self-generated mental event" and found that "[s]everal aspects of source monitoring assessed in [their] study improved with antipsychotic treatment. Significant improvements were found for the source discrimination of self-generated and heard items, and the recognition of heard items" (Keefe et al. 2003: 383, 387). If neuroleptics increase the ability of passivity-symptomatic subjects to monitor their own mental states and actions, then it becomes questionable whether in the passivity-symptomatic subjects in study (2) the ability to correctly other-ascribe intentions still contrasted with a comprehensive inability to correctly monitor their own intentions.

theory, postulates such a mechanism in order to account for first-order theoretical and practical reasoning, which involves the combination of representations with different attitude roles. The mechanism at issue is taken to operate much like the one that Goldman describes. For it too involves the detection and differentiation of the attitude types of representations by tracking the neural properties of the representations. Unlike Goldman's, and Nichols and Stich's accounts, however, the IDR theory explains how the mechanism at issue can come to classify a first-order representation *p* as a belief *p* and form a self-ascription *I believe p* on the basis of the representation *p*. The account explains this by proposing that the transition from the first-order to the second-order representation is grounded in and sustained by a particular combination of practical and theoretical reasoning. In offering an account of the link between a first-order representation about *p* and a self-ascription *I believe p*, the IDR theory provides an answer to a central question that both versions of the inner-scanner view leave open.

Turning now to (b), according to the IDR theory, TM-based self-ascriptions of beliefs can only be formed when one is also able to ascribe beliefs to others. This is because the formation of TM-based self-ascriptions is thought to recycle assumptions that underlie the formation of other-ascriptions of beliefs, and are part of the mindreading system. The formation of TM-based self-ascriptions relies, for instance, on one's understanding of the principle that what subjects will say in response to the question of whether *p* is what they believe about *p*. Since, according to the IDR theory, the mechanism implementing self-ascriptions of beliefs redeploys resources used for and grounded in other-ascriptions of beliefs, one prediction of the account is that the ability to form self-ascriptions of beliefs shouldn't be developmentally prior to but developmentally simultaneous with or posterior to the ability to form other-ascriptions of them. In contrast, the inner-scanner mechanisms that Nichols and Stich, and Goldman postulate are claimed to be operative already before and independently of the ability to other-ascribe attitudes (Nichols and Stich 2003: 163; Goldman 2006: 224-225). Their views thus imply the dissociation thesis, whereas he IDR theory implies the falsity of the thesis. As shown above, the data don't support but in some case speak against Nichols and Stich's, and Goldman's proposals. They are in line with and in some cases strongly support the parallelism thesis and therewith the IDR theory (see, e.g., the meta-analysis by Wellman et al. 2001, and, on autism, Williams 2010).

Thus, even though the IDR theory assumes, just as Nichols and Stich's, and Goldman's accounts do, a sub-personal mechanism for the detection of the attitude type of a representation, this mechanism is significantly different from the ones that their proposals describe. The features in virtue of which it is different make the IDR theory preferable to Nichols and Stich's, and Goldman's inner-scanner proposals.

2. The interpretive sensory-access theory

There is another major empirically oriented transparency-independent theory of self-knowledge of attitudes that is a competitor to the IDR account. It is Carruthers' (2011) interpretive sensory-access theory. Since I will at the end of this chapter endorse a revised version of it, I will offer a more detailed exposition of his proposal than of the other accounts discussed in the preceding sections.

2.1 The case against introspection and for the ISA theory

It is a common intuition that we can have introspective self-knowledge of attitudes, i.e. we can know our own attitudes without first having to observe ourselves or draw inferences from our own behaviour or circumstances (e.g., Searle 1983; Moran 2001, 2012; Nichols and Stich 2003; Byrne 2005, 2011; Goldman 2006; Boghossian 2008; Shoemaker 2012). Not everyone accepts this, however. Carruthers (2011), for instance, argues that experimental findings undermine the intuition.

One kind of data that he uses to make the point comes from Gazzaniga's (1995) experiments with so-called 'split-brain' patients. These are subjects whose left brain-hemisphere is partly or entirely disconnected from the right hemisphere.[50] In one of Gazzaniga's experiments, the brain's contra-lateral processing of visual information[51] was exploited. A split-brain patient called 'Joe' was asked to focus on a spot straight ahead, and then two pictures were shown to him simultaneously, one in his left visual field, the content of which was only available to his right brain hemisphere (which lacks the ability to produce language), and one in his right visual field, the content of which

[50] See Gazzaniga (1995) and Bayne (2008) for details on this intriguing condition and the functions of the two hemispheres.

[51] 'Contralateral processing' means that the right hemisphere processes information from the left visual field, and the left hemisphere processes information from the right visual field.

was only available to the left hemisphere (which houses a language production system and a mindreading capacity). When the instruction 'Walk!' was presented in his left visual field, he suddenly stood up and walked out of the van in which the experiment took place. When asked why he left the van, which was in fact triggered by the instruction to walk that was flashed to his right brain hemisphere, he replied without apparent insincerity that he wanted to go to the house to get a coke. He seemed to unknowingly interpret himself and confabulate an intention for his action while being under the impression that he had direct access to the cause of his behaviour.

Carruthers argues that this and the results of similar studies (2011: 40f, 326f) undermine our warrant for holding that we know our own attitudes introspectively. For we may, just like Joe, have the *impression* that we introspect our own attitudes, even though we in fact self-ascribe them interpretively. In the absence of any convincing reason for believing that we can come to know our own attitudes non-interpretively, the intuition that we are able to do so should no longer be accepted, Carruthers holds.

He proposes an alternative account, the "interpretive sensory-access (ISA) theory" (2011: 1). According to the ISA theory, the faculty that produces self-ascriptions of attitudes is the same interpretive mechanism that produces other-ascriptions of them, namely the mindreading system. The only difference is that in one's own case, the mindreading system has more sensory information available upon which to base its interpretation. This is because it is one of the judgment-making systems that are connected to the global workspace and can thus directly access all the representations broadcast. Hence, in addition to using overt behaviour, in one's own case, the mindreading system can also utilise, e.g., one's own affective and sensory-imagistic representations (e.g., visual imagery or inner-speech tokens), and the conceptual contents embedded in them (Carruthers 2011: 68f).

It doesn't have direct access to most of one's own attitudes, however, because, according to the ISA theory, which has the IA view[52] as one of its components, most of one's own attitudes aren't broadcast. Apart from the attitudes within the mindreading system's own database, the only attitudes that the system can directly access are sensorily-embedded judgments and felt desires. Any other attitude can't be broadcast and is only accessible to the system via a typically unconscious process of interpretation

[52] See Chapter 2.

of the attitude's expressions, e.g., in overt behaviour or imagery (Carruthers 2011: 51, 75f).

According to the ISA theory, the mindreading system doesn't have direct access to other attitudes because it was selected for working out other people's mental states and for operating specifically only on interpretation-dependent sensory input. The thought is that the system that produces self-ascriptions of attitudes initially evolved for the formation of other-ascriptions and was then subsequently turned inwards. As a result, since the system producing other-ascriptions of attitudes operates interpretively, so does the system producing self-ascriptions of them (Carruthers 2011: 66f).

Why believe any of this? There are theoretical considerations and empirical evidence that Carruthers mentions to support the ISA theory. I'll briefly consider them in turn.

The theoretical considerations speaking in favour of the ISA theory pertain to its explanatory simplicity and generality as well as its ability to integrate well-established theories from the cognitive sciences. For instance, the split-brain data suggests that the system producing self-ascriptions of attitudes sometimes operates interpretively while the subject is under the impression of having non-interpretive access to her own attitudes. Competitor accounts to the ISA theory, e.g., dual-method views, which hold that self-knowledge of attitudes is at least sometimes non-interpretive, need to postulate two different mechanisms to account for knowledge of attitudes, one for non-interpretive self-ascriptions of attitudes and one for interpretive ascriptions of them. In contrast, the ISA theory assumes only one, an interpretive capacity, i.e. the mindreading system. Since the intuition of direct, non-interpretive self-knowledge of attitudes is undermined *inter alia* by the split-brain data, and since that data suggests that the mindreading system also produces self-ascriptions of attitudes, in the absence of a positive argument for non-interpretive self-knowledge of attitudes, preference should be given to the ISA view. For it is the simpler and explanatorily more general proposal, Carruthers thinks (2011: 44f).

Another point that he offers in support of the ISA theory is that the account manages to integrate widely accepted theories from the cognitive sciences. For instance, since the ISA theory contains the IA view, it coheres well, Carruthers thinks, with research on the global workspace and working memory, which suggests that only sensory-imagistic

representations are broadcast (2011: 48f). Further, he argues that the ISA theory is in line with theories about the evolution of the ability to ascribe mental states. For example, we know that there is such a thing as an outward-focused, other-directed, interpretation-dependent capacity to work out mental states. And we know that, according to the currently arguably most plausible proposal of its origin, namely the social-intelligence hypothesis, this faculty evolved for social purposes, e.g., for Machiavellian purposes (e.g., manipulation and deception of others) (Byrne and Whiten 1988, 1997), cooperative purposes (e.g., cooperative breeding/communal rearing) (Hrdy 2009; Richerson and Boyd 2005), or for both kinds of purposes. The pressure that could have led to the selection of an outward-direct mindreading system is thus well understood. In contrast, we don't have a good grasp of the selection pressure that could have led to the evolution of an extra, non-interpretively operating mechanism for the production of self-ascription of attitudes.[53] The proposal that the ability to self-ascribe attitudes emerged from turning inward an outward-focused mindreading faculty offers a straightforward explanation of the origin of the ability to self-ascribe attitudes, and is well in line with work on other-ascriptions of them (Carruthers 2011: 67f). This again speaks in favor of the ISA theory.

Arguably the strongest kind of support for the theory, however, is the empirical findings that pertain to the account's predictions. One of the predictions that the ISA theory produces is that the dissociation thesis is false and that there is a co-development of the ability to form self- and other-ascriptions of attitudes. I have already argued above that the evidence confirms this particular prediction. I will hence focus on other ones now.

Carruthers holds that the "central, key, prediction made by the ISA theory" is that there should be "[f]requent confabulation" of attitudes (2011: 7). For, according to the account, a subject's self-ascriptions of attitudes are formed via an unconscious interpretation of overt behaviour or other sensory data (e.g., sensory-imagistic representations). And whenever they are formed on the basis of misleading sensory information, then a subject's subsequent self-ascriptions of attitudes should be mistaken. The ISA account predicts, for instance, that whenever subjects are unknowingly made to act as if they had a certain attitude, then they should, if asked

[53] Shallice's (1988) proposal might come to mind here according to which self-ascriptions of attitudes are required for executive functions. I will discuss and reject the proposal in the next chapter. See also Carruthers (2011: 67) for an argument against it.

about their motivation, self-ascribe the attitude even though they don't have it. Carruthers mentions a number of studies to show that this prediction is borne out. I will briefly consider a selection.

He cites, for instance, an experiment by Brasil-Neto et al. (1992), which involved TMS of areas of the motor cortex. Subjects were told that when they heard a click (which was the sound of the TMS magnet being activated), they should lift one or the other index finger. They could choose which finger to move. Unbeknownst to them, TMS was then, via a headset, applied to their motor cortex either on the right or the left, which led them to involuntarily move the index finger of their opposite hand. Even though their movement was caused involuntarily, in each case, the "subjects claimed to have been aware of *deciding* to lift that finger" (Carruthers 2011: 334). Carruthers holds that this is because, for their mindreading system, the "best explanation of the available data is that they chose to lift the index finger that subsequently moved", and so "that is what they report. But they are unaware that they make these reports as a result of self-interpretation. Rather, they think that they are *aware* of their decisions" (ibid).[54]

Carruthers also discusses studies by, e.g., Wegner and Wheatley (1999) in which subjects could, by being prompted to entertain action-relevant thoughts (pertaining to hand movements) shortly before they witnessed themselves perform the action, be made to believe that they intentionally caused the action even though that they didn't in fact do so. Carruthers holds that this is at odds with the view that subjects have immediate, non-interpretive access to their own decisions but in line with the ISA theory. For it suggests that the mindreading system produced a self-ascription of a decision on the basis of an unconscious self-interpretation of overt behaviour (2011: 340).

While the studies discussed so far pertain only to decisions, Carruthers also cites research that concerns judgments. He mentions, for instance, two experiments conducted by Briñol and Petty (2003). In one of them, subjects were asked to nod or shake their heads while listening to a message. When the message was convincing, nodding strengthened confidence in the message and head shaking reduced it. However, when the message was unconvincing the reverse happened. Carruthers argues that this is because subjects interpreted their "nodding behaviour as confirming their own initial

[54] There are problems with Carruthers' interpretation of this study. I discuss them elsewhere; see Peters (2014).

negative reactions to the message [e.g., in inner speech utterances such as 'How stupid!'] while head-shaking is interpreted as disagreement with those reactions" (2011: 344).

In a second study, Briñol and Petty (2003) had subjects write down three good or bad qualities that they thought they had as potential professionals (e.g., hard-working, polite, impatient, lazy etc.). They were asked to do so either with their dominant (right) or their 'weak', non-dominant (left) hands. Those features written down with the left hand tended to look 'shaky'. It turned out that subjects subsequently indicated more confidence in the judgments that they had written with their dominant hands than they did in judgments written with their left hands. Carruthers argues that this is because the "the mindreading system has learned of the relationship between low confidence and hesitancy, and the sentences written with the left hand looked hesitant" (2011: 344).

He claims that if people had non-interpretive access to their own "judgments" then the subjects in the two studies just mentioned "should just have accessed and reported on their current judgment. But they didn't" do so; they instead drew inferences about their judgments from their head nodding/shaking and handwriting (ibid).

There are a number of ways in which opponents of the ISA theory might respond to the findings just introduced. They might hold, for instance, that the data show at best that *sometimes* subjects unknowingly confabulate their own attitudes and work them out interpretively but not always. They might also argue that the data establishes at best that subjects tend to fail to correctly *remember* their attitudes, which is compatible with their having non-interpretive access to them when the attitudes are occurrent.

Carruthers' response to these objections is that they rely on the assumption that we have at least occasionally non-interpretive access to our own attitudes, but given, e.g., the confabulation data, which suggests that we might be under the impression of having non-interpretive self-knowledge of attitudes even though we don't in fact have such knowledge, this assumption is in need of a positive motivation. He holds that in the absence of such a motivation, e.g., the theoretical considerations mentioned above (explanatory parsimony etc.) speak against a dual-method account. Carruthers' overall argument for the ISA theory hence takes the form of an inference to the best explanation.

2.2 ISA vs. IDR

I now want to take a critical look at the ISA theory and relate it to the IDR account. According to the ISA theory, we have non-interpretive access only to sensorily-embedded judgments, felt desires, and the attitudes that are within the mindreading system's own database (Carruthers 2011: 53f). The access that we have to any other attitude is "always interpretive (and often confabulatory), utilizing the same kinds of inferences (and many of the same sorts of data) that are employed when attributing attitudes to other people" (Carruthers 2011: 1). If the ISA theory is correct, then the IDR proposal is wrong. For the latter holds that in principle any belief that can be retrieved in conscious thinking can via TM be known in a privileged way, i.e. without interpretation and inferences. The question is thus whether the ISA theory is correct. In this section, I will argue that it isn't. We do have privileged self-knowledge of beliefs whenever we use TM in the way the IDR theory proposes.

What speaks against this view? Carruthers (2011) rejects TM-based theories in general with an argument that I have already briefly touched on above. He thinks that when in response to the world-related question one judges *p*, then in order for the mechanism responsible for forming the self-ascription *I believe p* to form the self-ascription, the information that a "judgment is occurring" with the content *p* "would have to be accessible to that mental faculty", for it can't produce the self-ascription if *p* is, e.g., only supposed (Carruthers 2011: 82). But if that is right, Carruthers holds, then this commits the advocate of TM to the assumption of "some form of inner sense" for the detection of attitudes such as judgments, and the existence of such a thing is undermined by the case for the ISA theory (ibid).

However, I have already dealt with this kind of objection to TM-based theories in Chapter 4. I there argued that, on the IDR theory, the mechanism responsible for the judgment-detection doesn't produce representations of attitudes. Its outputs are first-order in nature and it tracks attitudes-types, e.g., judgments via the neural properties of the representations tokened. Note the existence of this mechanism is not undermined by the case for the ISA theory, for, as I argued in Chapter 2, its existence in fact needs to be granted in order to be able to explain how judgment- and decision-making systems can correctly combine representations of different attitude-types in *first-order*

theoretical and practical reasoning. Since that is so, Carruthers' argument against TM-based theories of self-knowledge doesn't threaten the IDR account.

But do the empirical findings that the ISA theory relies on undermine the account? As noted, the data on confabulation and self-interpretation show only that people *sometimes* lack non-interpretive access to their own attitudes. They don't suffice to establish that this is always the case. Hence, nothing in the evidence mentioned undermines the proposal that subjects are able to self-ascribe beliefs via TM in the way the IDR theory suggests.

It might be argued that if people could come to self-ascribe beliefs thus, i.e. non-inferentially, then one would expect that subjects shouldn't rely on an interpretation of their own behavior when they are asked to report, e.g., their confidence about a message or about their own judgments. Yet, as Carruthers argues, Briñol and Petty's (2003) studies suggest that subjects do rely on an interpretation of their behaviour (e.g., head nodding/shaking) or other evidence about themselves (e.g., their handwriting) to determine how confident they are about a claim or about their own judgments pertaining to, e.g., their positive and negative character traits that might affect their professional career.

The evidence seems to suggest that people can't without self-interpretation determine their own attitudes in the way the IDR theory suggests. For if they had non-interpretive access to their judgments or beliefs then in, e.g., Briñol and Petty's (2003) handwriting study, they should have non-interpretively known their current judgments about their positive and negative character traits. But they didn't seem to be able to do so.

However, again the data doesn't in fact undermine the IDR view, for I might have one way of coming to know *that* I am confident that *p* (and that I judge *p*), and another way of coming to know *how* confident I am that *p*. In Briñol and Petty's (2003) handwriting study, subjects were only asked about how confident they were about their judgments. The data thus show at best that one's knowledge of the degree of one's confidence that *p* is sometimes affected by the outcome of a self-interpretation (e.g., of one's shaky handwriting). They don't show that one relies on self-interpretation of behaviour to come to know *that* one is confident or judges that *p* at all. The findings at issue hence

don't undermine the view that we sometimes acquire self-knowledge of beliefs in the way the IDR theory proposes.

Do, e.g., the parsimony considerations speak against the IDR account then? This is not the case either. For the IDR theory holds that all the resources that are required for the production of TM-based self-ascriptions are already in place when one has a self-concept, is able to engage in first-order practical/theoretical reasoning, and can ascribe beliefs to other people on the basis of what they are saying. That is, the IDR theory of TM-based self-ascriptions doesn't postulate any additional mechanisms that could otherwise make the ISA theory more explanatorily or ontologically parsimonious.

In fact, parsimony considerations provide good grounds for preferring the IDR view to the ISA theory as an account of self-knowledge of beliefs that can be retrieved in conscious thinking. For instance, it is widely accepted, though hardly ever adequately defended (see Chapter 3), that TM is one way in which we can come to know our own beliefs. Given this, how would the ISA theory explain what is going on in applications of the method?

According to the ISA theory, upon wondering whether one believes *p*, one asks oneself whether *p*. The answer to the question then becomes expressed in sensory-imagistic representations, e.g., in an inner-speech token '*p*'. The mindreading system takes the sensory-imagistic representation as input, interprets it as an expression of one's belief *p*, and issues the self-ascription *I believe p* as output (Carruthers 2011: 82ff). So on the ISA view, three kinds of processing become operative. There is first reasoning on whether *p*. Then there is the expression of the outcome of the reasoning in, e.g., inner speech, and then this expression is decoded for its underlying attitude via an interpretation by the mindreading system.

Compare this with the account that the IDR theory offers. According to the IDR theory, in cases of unreflective self-attributions, one proceeds directly from the outcome of one's reflection on whether *p*, without any kind of interpretation of behaviour, context, or sensory-imagistic states, to the formation of the self-ascription *I believe p*. The occurrence of the judgment *p* is in the context of TM enough to activate the disposition to produce the self-ascription.

To be clear, it is crucial for the IDR account that one has direct access to one's own judgments in conscious first-order thinking. The ISA theory denies this. It relies on the

IA view. However, as I argued in Chapter 2, the IA view is false. The ISA theory hence rests on a mistaken picture of the kind of access that we have to our own attitudes in conscious first-order thinking. As I contended in Chapter 2, we have *direct* access to our own judgments in conscious first-order thinking. Since that is so, there is little in the case for the ISA theory that could threaten the IDR account of TM-based self-ascriptions.

Note that according to the IDR account of TM-based self-ascriptions, apart from asking oneself whether *p*, only one kind of processing becomes operative, namely reasoning on whether *p*, where the outcome of the reasoning is then the basis for the formation of self-ascription of a belief about *p*. Since the IDR theory holds that no expression of a judgment and no mindreading need to occur in order for one to form TM-based self-ascriptions, it offers a more straightforward explanation of what happens in applications of TM than the ISA theory does.

Further, since the IDR theory holds that other-related resources for forming ascriptions of beliefs only need to be applied to oneself for one to be able to non-inferentially self-ascribe beliefs via TM, the ISA assumption that the mindreading system has to become activated to interpret an inner-speech token seems entirely unmotivated. An argument that already figured in the discussion above on why the dual reasoning that the IDR theory postulates isn't typically executed but remains suspended can be used here again to illustrate that the assumption now at issue is psychologically implausible and arguably false. For initiating interpretive processing would unnecessarily add computational complexity to the processing of the system responsible for producing TM-based self-ascriptions without adding any benefit. Indeed, this would only increase the possibility of erroneous self-ascriptions. Of course, it might still be that such processing is nonetheless employed. But there are good empirical grounds to deny that this will happen, for, as noted above, there is much evidence that, due to resource limitations, the human cognitive system tends to avoid computations if it can, and opts for simple rather than difficult methods to solve cognitive tasks (Fiske and Taylor 2013). In psychology, this is often referred to as "attribute substitution": when presented with a difficult question, subjects tend to unconsciously substitute it for and instead answer an easier one (e.g., Kahneman and Frederick 2002; Evans 2008; Kahneman 2011; De Neys et al. 2013).

The fact that the human cognitive system has a general tendency to short-circuit

information processing and adopt heuristics in judgment- and decision-making supports the assumption that the judgment-forming system responsible for producing self-ascriptions of beliefs will similarly adopt simplifying strategies to perform its function, and employ a non-inferential rather than an interpretive method to produce self-ascriptions of beliefs if available. This assumption would, given the processing constraints under which the human cognitive system operates, be more realistic than the ISA proposal that interpretive processing is activated. The argument against the ISA theory can then be summarised in the following seven points.

(1) The case for the ISA theory doesn't undermine the IDR theory of self-knowledge of beliefs.
(2) We don't know whether TM-based self-ascriptions of beliefs are formed either in the way the IDR theory proposes or in the way the ISA theory suggests.
(3) Still, we have reason to assume that the human cognitive system uses computationally simple rather than complex methods to perform a given cognitive task.
(4) TM-based self-ascriptions formed in the way the IDR theory proposes are more computationally tractable than TM-based self-ascriptions formed in the way the ISA theory proposes.
(5) Given that the IDR theory doesn't require any more assumptions than the ISA theory and the IDR theory is better in line with the empirical evidence on how the human cognitive system processes information, there is reason to believe that TM-based self-ascriptions are formed in the way the IDR theory proposes.
(6) Since self-knowledge of beliefs acquired in the way the IDR theory proposes is privileged in nature, at least sometimes, namely when using TM, subjects have privileged self-knowledge of beliefs.
(7) Since the ISA theory disallows this, points (1)-(6) speak against the theory.

3. The IDR-ISA hybrid view

The discussion in the previous section provides some motivation for adopting a hybrid account of self-knowledge of attitudes that combines the IDR theory and a revised version of the ISA theory. The combination that I have in mind would require the following two revisions of the ISA theory.

As it stands, the ISA theory involves the claim that apart from sensorily-embedded

judgments and felt desires, no attitudes figure in the workspace. In Chapter 2, I argued that this is false. There are attitudes in the workspace and directly accessible in first-order reasoning that don't fall into these two categories. The just-mentioned claim of the ISA theory thus needs to be abandoned.

The point doesn't yet threaten the more central claim of the ISA theory that we can only come to know the attitudes at issue interpretively, however. For even though the attitudes might be in the workspace and directly accessible to first-order judgment- and decision-making systems, it could still be that the mindreading system, as one of the judgment-forming systems connected to the workspace, does not, unlike the other systems, have direct access to attitudes when they are broadcast. For it could have evolved specifically for the purpose of interpretively working out attitudes and hence lack the capacity of the other systems to non-inferentially access the attitude types of the representations broadcast.

However, I argued that at least sometimes one's direct access to attitudes in first-order reasoning can be the basis for non-interpretive self-knowledge of beliefs. This is the second revision of the ISA theory. It is also the point where the IDR theory comes in as the second component of the hybrid account. As noted above, the procedure for forming self-ascriptions of beliefs that the IDR theory describes relies on the resources of the mindreading system. Since that is so, one way of integrating the IDR account with the revised version of the ISA theory is to hold that the procedure is one of the methods via which the mindreading system produces self-ascriptions of beliefs. On this proposal, even though the mindreading system evolved, just like the ISA theory holds, for the purpose of interpretively working out mental states, and typically does operate interpretively, once one has executed the dual reasoning introduced above, the processing of the mindreading system is functionally re-structured. It is restructured so that, in the context of TM, the system becomes able to suspend interpretive and inferential processing and make direct use of the judgments broadcast in the workspace. The mindreading system attains the capacity to produce privileged self-knowledge of beliefs, because the sub-personal attitude-type detection mechanism that is already operative in first-order reasoning is then integrated into that system. The mindreading system can subsequently redeploy the mechanism for unreflective self-attributions.

On the resulting view, self-knowledge of attitudes is acquired via the operation of the same interpretive mechanism that produces other-knowledge of attitudes, i.e. the

mindreading system. As a consequence, typically, even in one's own case, ascriptions of attitudes will be formed interpretively and so not be privileged in nature. Nonetheless, the mindreading system can also produce non-inferential self-ascriptions of beliefs via TM in the way the IDR theory proposes. I shall call this kind of dual-method account of self-knowledge of attitudes the *IDR-ISA hybrid* view.

Why should we endorse it? I have already argued for the IDR theory in Chapter 4 and in the preceding sections here. The main reason for adopting the revised ISA theory is that, unlike other presently advocated accounts of self-knowledge of attitudes, which mostly assume, without adequately defending, that we have non-interpretive self-knowledge of attitudes, the ISA theory is well supported by both empirical data, e.g., developmental evidence for the parallelism thesis, findings on confabulation and self-interpretation etc., and theoretical considerations pertaining, e.g., to theoretical parsimony/generality, and the evolution of mental-state ascriptions. Furthermore, the ISA theory captures nicely how we come to know those of our own beliefs and other attitudes that we can't easily retrieve in conscious thinking, e.g., repressed beliefs and desires (see, e.g., Lawlor 2009; Cassam 2014, 2015 for illustrative examples). For, in order to find out about these attitudes, it seems we do need to interpret ourselves, or consult other people (e.g., psychotherapists) who may tell us about them. These points speak for adopting the ISA theory as a second component of a hybrid proposal formed together with the IDR theory. Finally, both accounts would benefit from the integration. For if it is combined with the IDR theory, the revised ISA theory inherits a better picture of subjects' first-order access to attitudes in conscious thinking (the DA view) and a more plausible explanation of TM-based self-ascriptions of beliefs. In exchange, the TM- and belief-specific IDR theory becomes part of a well-motivated general theory of self-knowledge of attitudes.

4. Conclusion

In this chapter, my aim was to further defend and develop the IDR theory. To do so, I related the account to two major transparency-independent views of self-knowledge of attitudes, namely the inner-scanner theory, and Carruthers' ISA theory. As for the inner-scanner theory, I focused on Nichols and Stich's (2003), and Goldman's (2006) versions of the view. I argued that their proposals leave unexplained how the second-order representations involved in self-knowledge of beliefs are formed in the first place and also produce predictions that are not supported but in some cases in fact

contradicted by the empirical data. The IDR theory avoids both problems and is hence preferable. As for the ISA theory, I noted that it doesn't satisfactorily explain our access to attitudes in first-order reasoning and the formation of TM-based self-ascriptions of beliefs. I maintained that the IDR theory is more plausible than the ISA view in both respects. In the final step of the overall argument of this chapter, I then proposed to integrate the IDR theory with a revised version of the ISA theory in order to embed it into an empirically and theoretically well-supported general account of self-knowledge of attitudes. The result of this integration, the IDR-ISA hybrid, combines the asymmetry view and the symmetry view. It holds that self-knowledge of attitudes is developmentally posterior to or simultaneous with other-knowledge of them because it requires turning onto oneself the faculty that is already in place and evolved for other-knowledge of attitudes, i.e. the mindreading system. Nonetheless, even though this system typically operates interpretively, when it produces self-ascriptions of beliefs via TM in an unreflective manner, then the resulting self-knowledge will remain privileged in nature.

CHAPTER 6

The function of self-knowledge of beliefs

In the preceding chapter, I adopted a view of self-knowledge on which the capacity for self-knowledge of attitudes emerged from turning inward onto oneself a faculty that evolved for the purpose of working out other people's mental states, i.e. the mindreading system. What the view leaves unexplained is why this faculty was turned inward to begin with. Since my focus is on privileged self-knowledge of beliefs, why is there such a thing as privileged self-knowledge of beliefs at all? A satisfactory theory of privileged self-knowledge of beliefs shouldn't leave this a mystery.

Typically, when naturalistically minded theorists try to answer why a particular trait exists, they search for the evolutionary function of it. The question of the evolutionary function of the trait is then in turn answered by mentioning some adaptive or reproductive advantage that the trait confers onto an organism or subject with it. If we apply this to self-knowledge of beliefs, a satisfactory naturalistic theory of privileged self-knowledge of beliefs should say what adaptive or reproductive advantage such knowledge confers onto a subject who has it.

In the present chapter, I want to offer an answer. To make the issue more tractable, I will initially concentrate mostly on the more general question of the function of self-ascriptions of beliefs. I will return to the specific question of the function of privileged self-knowledge of beliefs at the end of the chapter. With respect to the question of the function of self-ascriptions of beliefs, there isn't much work available. Nonetheless, the following three proposals can be distinguished.

(1) Self-ascriptions of beliefs are required for conscious beliefs.
(2) Self-ascriptions of beliefs are required for controlling one's cognition.
(3) Self-ascriptions of beliefs are required for moral responsibility.

In what follows, I first argue that these proposals are unsatisfactory and then develop my own view on the matter. It holds that:

(4) Self-ascriptions of beliefs are required for belief reports.

The ability to verbally communicate one's own beliefs to others not merely by expressing but by reporting them has in cooperative social environments adaptive advantages because it leads to better joint decision-making and group coordination with mutual benefits for both self and others. I will argue that the function of TM-based self-ascriptions and so TM-based privileged self-knowledge of beliefs, in particular, is derived from (4). It is to enable the most efficient and most reliable verbal communication of our own beliefs to others. To make the case for this view, I will draw on recent cognitive-scientific research on the function of conscious metacognition and integrate it with the IDR-ISA hybrid view introduced in Chapter 5.

In sections 1-3, I will discuss proposals (1)-(3). In section 4, I develop (4), and relate it to TM. In section 5, I then introduce an account of the function of TM-based privileged self-knowledge that combines (4) with the IDR-ISA hybrid view.

1. Self-ascriptions of beliefs and conscious beliefs

The first proposal concerning the function of self-ascriptions of beliefs will only be considered briefly. It takes its starting point from theorising about consciousness.

Some philosophers hold that a self-ascription of a mental representation is required for the latter to become conscious. They typically take it that this is required for the representation to be both phenomenally and access conscious (see, e.g., Carruthers 2000; Rosenthal 2002, 2005; Gennaro 2012). For instance, Rosenthal (2002, 2005) claims that "a state is access conscious" and phenomenally conscious only if one is "conscious of being in that state" via a higher-order thought about it, where "the content of one's HOT [higher-order thought] must be, roughly, that one is in that very state" (Rosenthal 2002: 409). If higher-order thoughts that self-ascribe mental states are required for the states to become conscious, then it might be suggested that self-ascriptions of beliefs, which are higher-order thoughts also, are required for and have the function of making beliefs conscious.

There is a sense in which the proposal can't be right, however. For consciousness is an occurrence or event. It can come and go. In contrast, a belief *p* is a state, not an event; it persists over time. It doesn't, e.g., disappear when one attends to something else, falls asleep, or becomes unconscious. Given that consciousness is an event but beliefs are

states, beliefs can't themselves be conscious, for they can't be both states and events (Crane 2013). Since there are arguably no such things as conscious beliefs, self-ascriptions of beliefs can't be required for the beliefs self-ascribed to be conscious.

However, one might take a conscious belief to be a belief that is brought to consciousness. When a belief *p* is brought to consciousness then what occurs in consciousness is an event of affirming and committing oneself to the truth of a proposition *p*. In Chapter 1, I called such an event a conscious judgment. One might hold that self-ascriptions of beliefs are required for the judgments that correspond to the beliefs to be conscious.

But there is good reason to reject this proposal too. For mental events in general can arguably be conscious in the absence of any kind of self-ascription of either mental states or events. For instance, it seems hard to deny that neonates, who lack concepts of mental states and events and hence can't self-ascribe them, can nonetheless still undergo conscious events, e.g., pain experiences (Block 2007). If that is right, then presumably the same holds for events such as judgments. That is, judgments can be conscious without any self-ascription of some mental state or event.

More generally, it is worth noting that there is no uncontroversial theory of consciousness. The HOT account of consciousness is one proposal among others. It isn't forced on us. An equally common alternative theory of consciousness that has already been discussed above would be the global workspace theory (e.g., Baars 1988; Dehaene and Naccache 2001). The theory would hold that a conscious judgment is one that is the target of attention and broadcast in the global workspace. Note that this doesn't require any representation *of* the attended to and broadcast representations (Prinz 2011). One might also adopt a self-representational theory, according to which a representation *x* about *y* is conscious when "(i) *x* represents *y* and (ii) *x* represents *x*'s representation of *y*" (Kriegel 2013: 36). Note that on this view, the self-representing representation needn't necessarily be represented *as such*.[55] Yet another option would be to hold that higher-order quasi-perceptual (e.g., Armstrong 1968, 1999; Lycan 1996), or indexical-conceptual (Lurz 2006) representations about representations are sufficient

[55] Kriegel (2013), for instance, who is a recent defender of the self-representational theory of consciousness, emphasises that he does "not mean to suggest that the representation must represent itself *as* itself" (36).

for the latter to become conscious.

None of the three proposals just mentioned requires any self-ascription of a mental state or event (i.e. a conceptualisation of the mental state or event *as* a mental state or event) for judgments to be conscious. This is only required on the HOT account. Since there are good reasons for rejecting the HOT account, it is perhaps better to set aside the view that self-ascriptions of beliefs are needed for conscious beliefs or judgments and look for alternative suggestions on the function of them.

2. Self-ascriptions of beliefs and cognitive control

A second proposal on the function of self-ascriptions of beliefs is that they are required for the control of one's own cognition. I shall distinguish three arguments that might be mentioned to support this view: (i) the argument from executive functions, (ii) the argument from belief revision, and (iii) the argument from self-blindness.

2.1 The argument from executive functions

The argument from executive functions applies to the evolutionary function of meta-representations in general. It is worth briefly discussing the argument here because it might be thought that it equally allows specifying a function for self-ascriptions of beliefs.

A number of theorists hold that meta-representations are needed for executive functions, i.e. for the monitoring and controlling of one's own cognitive processes. The thought is that it is "part of the evolutionary *adaptive significance*" of meta-representations to "enable correction of errors made in first-order […] processing" (Rolls 2014: 489). They are to allow a subject to supervise first-order cognitive processes, adjust or fine-tune behaviour, initiate new strategies when difficulty is encountered, and so on (see, e.g., Shallice 1988; Wimmer 1989; Perner 1991; Gennaro 2012).

However, it isn't clear whether meta-representations are indeed required for these purposes. For instance, Proust (2013) argues convincingly that humans, and, e.g., rhesus macaques and bottled-nosed dolphins, can monitor and assess their own cognitive states and activity, and employ their self-evaluations to structure their action via sensitivity to first-order representations, which she calls "noetic feelings", e.g., feelings of confidence, familiarity, or processing fluency (2013: 318). These feelings aren't meta-

representational, Proust holds, for they don't represent the contents of the first-order states and processes they track but only the properties of the vehicles of the contents (e.g., trace strength)[56] (2013: 71). Further, the monitoring and control of cognitive processes in general arguably only requires that the cognitive processes be structured into different layers. As Carruthers (2009a) argues, the controlling and the monitoring of

> the progress of a task may just require the supervisory system to possess a (first-order) representation of the goal-state, together with some way of comparing the current output of the system with the represented goal-state and making adjustments accordingly. The status of the goal *as a goal* needn't be represented. (67)

Finally, the correction of errors made in first-order reasoning that, e.g., Rolls (2014) focuses on, doesn't seem to require meta-representation either. For, as Rosenthal (2008) notes, if, e.g., a step

> in a multistep chain of reasoning is erroneous, that step will by itself likely result in some first-order dissonance with other antecedent beliefs. That dissonance will serve to locate the error, and so make possible the adjusting of the multistep chain at that point. Interactions among first-order states that reflect the intentional content of those states can iron out errors independently of any higher order monitoring. (836)

I conclude that the argument that meta-representations are required for executive functions remains unconvincing.

2.2 The argument from belief revision

It might be suggested that self-ascriptions of beliefs in particular could still have an important function when it comes to the control of one's own cognition. In rational subjects, for instance, self-ascriptions of beliefs might be required for belief revision in episodes of conscious thinking. Shoemaker (2012) advocates this proposal. He maintains that it is "a condition of being a rational subject that one's belief system will regularly be revised with the aim of achieving and preserving consistency and internal

[56] The term 'trace' here refers to the structural changes in neural cells following learning. The strength of the neural connections involved is the trace strength.

coherence" (245). While this might sometimes happen "automatically", in some cases, he claims, in cases of conscious thinking, when "the revision of the belief system requires an investigation on the part of the subject, one that involves conducting experiments, collecting data relevant to certain issues, or initiating reasoning aimed at answering certain questions", this is "an intentional activity on the part of the subject" that "requires awareness of, and so beliefs about" whether "certain contents are believed" and others "are not believed" (Shoemaker 2012: 245). That is, it requires self-ascriptions of beliefs.

It isn't difficult to see, however, that self-ascriptions of beliefs aren't needed for the purpose Shoemaker has in mind either. Peacocke (1996) offers an example that helps illustrate this.[57]

> Suppose you come home, and see that no car is parked in your driveway. You infer that your spouse is not home yet; you store that information, or misinformation, and move on to think about other matters. Later, you may suddenly remember that your spouse mentioned in the morning that the brakes of her car were faulty, and wonder whether she may have taken the car for repair. At this point, you suspend your original belief that she is not home yet. For you come to realise that the absence of her car is not necessarily good evidence that she is not home. If the car is being repaired, she would have returned by public transport. Then finally you may reach the belief that she is home after all, given your next thought that she would not have taken any risks with faulty brakes. (129)

This is an episode of conscious thinking. Furthermore, there are

> initially thoughts about the world; there is then the bringing to bear of additional information; there are thoughts about what would justify what; there is suspension of certain attitudes; and finally some resolution. This is a (modest) piece of reasoning which results in revision of beliefs in the light of thoughts about relations of evidence and support […]. (Peacocke 1996: 129)

[57] Peacocke mentions the example in his critique of Burge's (1996) account of critical reasoning, but it equally applies to Shoemaker's view.

Since that is so, the piece of reasoning seems to fit the description of the episodes of thinking that Shoemaker has in mind. Importantly, however, it doesn't involve any kind of self-ascription of beliefs; "the thoughts it involves all seem to be thoughts about the world, not about the thinker's thoughts" (ibid). There is no need for a conceptualisation of the representations involved as beliefs. First-order reasoning, an awareness of what is true and what is false, and the ability to accept true and reject false first-order propositions are sufficient. Against Shoemaker's claim, the revision of one's own beliefs in conscious thinking doesn't seem to require self-ascriptions of beliefs either.

It might be objected that this response to Shoemaker's argument misunderstands his point. It might be argued that his point is that it is a "condition of being a rational subject" that in one's revision of one's belief system one is guided by an assessment of one's attitudes and reasoning *as such* (Shoemaker 2012: 245). And in the above example, when one thinks about one's wife, one does then not think of one's reasoning and attitudes as such.

However, why should thinking about one's attitudes as such be a *condition* of being rational? When you engage in the above reasoning about whether your wife is at home, it seems you are perfectly rational in your thinking and belief revision. What would be missing for you to count as rational? If the reply is that what is missing is that you think reflectively, i.e. about your own reasoning and the beliefs involved *as such*, then it seems to become part of the definition of being rational that one represents one's own beliefs. The proposal on why subjects, qua being rational, self-ascribe beliefs will become explanatorily vacuous.

2.3 The argument from self-blindness

Shoemaker (1996, 2012) has offered a second argument to show that self-ascription and self-knowledge of beliefs are required for the control of one's own cognition in that they are necessary for rationality. The argument is more specifically directed against causal theories of self-knowledge of beliefs, which are theories that assume a causal connection between a belief *p* and a self-ascription of it. The strategy behind the argument is to show that causal theories entail the possibility of something that in fact isn't possible. In a nutshell, Shoemaker's point is the following.

(1) Causal theories imply that "self-blindness" – a condition "in which someone who is perfectly rational, and suffering from no cognitive deficiency, is introspectively blind to his own beliefs, and so incapable of self-ascribing them except on the basis of third-person evidence" (Shoemaker 2012: 242) – is a conceptual possibility.
(2) Self-blindness is conceptually impossible.
(3) Thus, causal theories are false.

Causal theories imply the conceptual possibility of self-blindness, Shoemaker holds, because the connection between first-order beliefs and self-ascriptions is on these accounts contingent in nature. Given that any contingent process can go wrong, self-blindness must on these views be possible, he claims.[58]

Shoemaker then proceeds to provide an argument for the view that self-blindness is conceptually impossible. It relates to the "incoherence of affirming something while denying that one believes it" and involves an appeal to Moore's paradox (Shoemaker 2012: 240).

The paradox has already been mentioned above. It concerns the proposition that a speaker would assert if she uttered '*p*, but I don't believe that *p*'. Shoemaker holds that if a subject were to believe the proposition expressed by the utterance, her belief in the proposition would be self-falsifying because she would believe *p* but at the same time believe that she doesn't believe *p*, which seems incoherent. Since rational creatures are creatures that avoid incoherence, they will avoid forming Moore-paradoxical beliefs, Shoemaker reasons.

[58] Shoemaker assumes that proponents of the causal theory are committed to the view that there could be a subject that is rational, and cognitively and conceptually just as a normal subject, yet lacks introspective access to her attitudes. However, this assumption is false. Advocates of casual theories need only commit themselves to the view that there could be creatures that do not have introspective access to their own attitudes (Gertler 2011: 157f). This commitment does derive from their view that the relation between a belief *p* and the second-order belief is contingent. But note that allowing for an introspectively blind creature isn't the same as allowing that such creature would also be rational and cognitively and conceptually like a normal subject. Advocates of the causal theory are free to reject the possibility of a self-blind subject (ibid).

Would a subject S, who is self-blind to her own beliefs, be able to avoid producing Moore-paradoxical statements? Shoemaker argues that there might be cases where the world-related evidence available to S supports the proposition, e.g., that it is raining, and the third-person evidence that she has available about her belief supports the opposite proposition that she does not believe that it is raining. In such circumstances, S might be led to assert 'It is raining, but I don't believe that it is raining'. However, Shoemaker continues, S is, as a self-blind subject, by assumption "perfectly rational, and suffering from no cognitive deficiency" (2012: 242). Since that is so, she ought to be able to recognise the puzzling character of her assertion, and be able to reason as follows:

> *p*. Since *p* is true, it will, *ceteris paribus*, be in the interest of anyone to act on the assumption that *p*, if one is in circumstances (call these relevant circumstances) in which whether one so acts is likely to affect the satisfaction of one's interests.
>
> To act on the assumption that *p* is to act as if one believes that *p*. And part of acting as if one believes that *p* is acting in ways that indicate to others that one believes that *p*; for given that *p* is true, it will be in anyone's interest to act this way in relevant circumstances. So acting will help one enlist the aid of others who believe that *p* in the pursuit of one's goals. Others who believe that *p*, and share one's goals, will cooperate with one in ventures undertaken on the assumption that *p*, and since *p* is true such ventures will tend to be successful. Acting in ways that indicate to others that one believes that *p* will include saying, in appropriate circumstances, that one believes that *p*. Since this applies to everyone, it applies to me. And since I am in appropriate circumstances, I should say that I believe that *p*. (Shoemaker 2012: 241-242)

The same line of reasoning will lead S to avoid saying 'I don't believe *p*' when for her *p* is true, and when she has thus reason to say '*p*'. As a result, S will tend to avoid Moore-paradoxical statements. Indeed, Shoemaker holds that once S infers from *p* the conclusion that it is in her interest to act on the assumption that *p* is true, this will, in combination with her desires and her other beliefs, lead her "to precisely the sorts of behaviour that the second-order belief" that she "believes that *p* will lead to" (2012: 244). And since "if everything is *as if* a creature has knowledge of its beliefs and

desires, then it *does* have knowledge of them" (Shoemaker 1996: 34), Shoemaker concludes that S will have self-knowledge of beliefs.

The supposition that there could be a self-blind subject, i.e. a subject that is both rational and cognitively and conceptually just as one of us but without self-knowledge of beliefs, is thus undermined. And, since causal theories are, on Shoemaker's view, committed to the possibility of self-blind subjects, these theories will be undermined also, he maintains. According to Shoemaker's reasoning, one can't be both rational, and cognitively and conceptually just as one of us but without self-knowledge of beliefs. The latter is required for rationality.

However, there is good ground to reject his argument. Note first that the outcome of S's reasoning in Shoemaker's example is that she says that she believes *p* and acts in various other ways *as if* she believes *p*. Suppose that at some point she reflects on why she says 'I believe *p*' and acts in the way she does. Via a reversal of the initial reasoning that led her to produce her utterance, she will soon find that she merely acts as if she believes *p* but doesn't have any idea as to whether she does in fact believe *p*. If that is so, however, then S is arguably only feigning self-knowledge of beliefs. For she lacks what each of us is familiar with from the first-person perspective when we come to know our own beliefs, namely the insight that we do indeed believe *p*. Since Shoemaker's argument only succeeds if it is granted that S has self-knowledge of her belief *p*, and since there is reason to deny that this is so, his argument can be rejected. The argument doesn't show that rationality requires self-ascriptions and self-knowledge of beliefs. It provides at best an account of how someone who lacks the ability to self-ascribe and self-know beliefs can nonetheless come to avoid the apparent incoherence of asserting something while denying that s/he believes it.

3. Self-ascriptions of beliefs and moral responsibility

The third and last extant proposal on the function of self-ascriptions of beliefs that I want to consider is closely related to the just-mentioned view that self-ascriptions and self-knowledge of beliefs are required for rationality. It is the view that they are needed for moral responsibility. Bilgrami (1998, 2006, 2012) has offered a widely discussed argument in support of this view. In what follows, I will focus on his argument only.

Bilgrami claims that self-knowledge of attitudes in general is a condition for the possibility of responsible agency. To argue for this claim, he begins by noting that we sometimes adopt "reactive attitudes", e.g., resentment and indignation, toward each other's action and engage in "evaluative practices", e.g., blame and punishment, that these attitudes justify (Bilgrami 2006: 74). He continues that to "blame or resent a particular action is to presuppose that it has been freely enacted", and to presuppose that an action has been freely enacted is in turn to presuppose that both the "action" and the "intentional states […] that potentially go into the production" of it are "*self-known*" (Bilgrami 2012: 268). His thought is the following. To know what I'm doing, and so to be held responsible for my action, I must "not merely know that [I have] acted […] but [I] must also know the intentional states which cause and explain (rationalise) the action", because these states, in triggering and being the reasons for my actions, make my actions into, and allow classifying them as, the intentional actions that they are (Bilgrami 1998: 222). Bilgrami's point is that "there is no correctly describing action except in terms of the intentional states that explain (rationalise) it" (1998: 223). For instance, suppose I lift a glass of water and then drain it. If I don't know the attitudes that rationalise my action then I can't correctly describe my behaviour as, e.g., my quenching my thirst by drinking a glass of water as opposed to some other action or indeed simply some random interactions of objects in space-time. The thought is that a correct description requires me to determine the intention that underlies the action. And this in turn requires me to know that I *desire* that my thirst be quenched and that I *believe* that drinking the water will have the effect of quenching my thirst (Bilgrami 2006: 100). Hence, if I don't know the intentional states that rationalise my action, then I don't know how my action is correctly described and so I don't know what I'm doing (Bilgrami 2006: 93f, 100). Since self-knowledge of attitudes including beliefs is required for knowing what one is doing, and one's knowing what one is doing is required for one's being responsible, self-knowledge of beliefs is required for being responsible, Bilgrami concludes:

> [s]elf-knowledge [of attitudes] is necessary for responsibility *for no other reason* […] than that our *evaluative* justifications of the practices of assigning punishment and blame seem to be apt only when self-knowledge is present. (1998: 216)

Bilgrami's argument is unpersuasive, however. Its central assumption that self-knowledge of attitudes is required for knowing what one is doing is arguably mistaken. Consider my waving my hand while standing by the road when a car is approaching. My action might be many different things. For instance, it might be my saying goodbye to someone or my greeting someone. Suppose that what leads me to the action at issue is the desire to stop a taxi and the belief that waving my hand by the road will stop a taxi. What leads me to act in the way I do might be the following simple piece of practical reasoning.

(i) I need to stop a taxi.
(ii) If I wave my hand by the road, I will stop a taxi.
(iii) Decision: wave the hand by the road.

Suppose I then go to the road and wave my hand. How do I know that what I'm doing is waving my hand to stop a taxi rather than saying goodbye to someone, or greeting someone? Do I need to represent the attitudes that feed into my decision-formation process?

Note that holding that someone *needs* to do something doesn't presuppose a mental-state attribution to him/her. For one might, e.g., need to take a medicine without wanting to do so, or one might need to go to work without wanting to (Perner and Roessler 2010). Similarly, my holding that I need to stop a taxi doesn't presuppose that I represent a desire. I might just treat it as a fact that I have to stop a taxi, e.g., in order to quickly get somewhere. My reasoning from (i) to the decision (iii) thus doesn't require a self-ascription of a desire.

Do I need to self-ascribe the belief (ii) that leads me to my decision (iii) in order to know that I'm waving my hand to stop a taxi rather than to, e.g., say goodbye to someone or greet someone? This is not the case either. For to find out what I'm doing, I might simply go back to the piece of practical reasoning that led me to exhibit the action in the first place. The premises of the reasoning will tell me why I'm doing what I'm doing because they inspired me to do what I'm doing in the first place and already contain a specific description of the action that they lead me to perform. I know that my hand waving is to stop a taxi, because I know that what caused me to perform this

behaviour[59] is that (i) I need to stop a taxi, and that (ii) if I wave my hand by the street, I will stop a taxi. Given that I know that (i) and (ii) lead me to wave my hand, I know that I wave my hand to stop a taxi.

Since neither (i) nor (ii) involves a representation of an attitude but only a representation of what for me are worldly facts, I can know what I'm doing without self-ascribing the attitudes that lead me to the action. And since that is so, I can arguably also be morally responsible without knowing my own attitudes. Bilgrami's argument for the view that self-knowledge of beliefs is required for moral responsibility is thus undermined.

4. Self-ascriptions of beliefs and belief reports

With the discussion of the preceding sections in mind, one might wonder whether self-ascriptions and self-knowledge of beliefs are really necessary for anything. Cassam (2014) writes, for instance, that in "every case in which it looks as though self-knowledge might be necessary for the achievement of some high ideal it turns out not to be" (222). And he goes one to propose that the "right reaction to this is not disappointment but reflection on why it ever seemed a good idea to defend a claim of this form. Self-knowledge is still valuable if it leads to other goods, even if those other goods could be achieved without it" (ibid).

The problem is that if self-knowledge or more generally self-ascriptions of beliefs merely lead to goods that could be achieved without them, then it is unclear what selection pressure could have brought about their evolution in the first place. The ability to form self-ascriptions of beliefs might of course just be a "spandrel", a trait that isn't adaptive itself but merely arose as a by-product of another trait that did evolve as a direct result of adaptive selection (Gould and Lewontin 1979). But it seems to me that before giving up on the search for a function for which self-ascriptions of beliefs were and are required, it is worth exploring whether they could have been and could still be needed for communicative purposes.

[59] Empirical studies suggest that subjects often confabulate a reason and intention for their action *post-hoc*. However, so far, none of them undermines the view that subjects know the reasons for their action before and during the action they decide to perform. It might be that they swiftly forget about their motives after the choice or action (see, e.g., Johansson et al. 2005; Hall et al. 2010).

More specifically, I want to propose that self-ascriptions of beliefs were selected for and have the function of allowing a subject to verbally transmit her own beliefs to others by reporting them. This straightforward view, which I shall defend in the rest of the chapter, will bring the discussion of the function of self-ascriptions of beliefs back to the IDR theory and TM.

It draws on and further develops recent cognitive-scientific research on conscious metacognition. Since that is so, a brief introduction to work on metacognition in general will be useful before returning to the question of the function of self-ascriptions of beliefs.

4.1 Metacognition

Different theorists use the term 'metacognition' in different ways. A broad and a narrow construal can be distinguished. Broadly construed, 'metacognition' refers to cognition about one's own cognition. It consists of a monitoring process, which supervises and assesses cognitive performances, and a control process, which regulates cognitive activities and overt behaviour (see, e.g., Koriat 2007; Frith 2012).

Defined more narrowly, 'metacognition' refers to monitoring and control processes that involve the use of metacognitive representations, i.e. representations of properties of one's cognitive processes, e.g., the reliability of a perceptual representation, decision uncertainty, or processing fluency (Proust 2013; Shea et al. 2014). In what follows, I adopt the narrow notion of the term.

Metacognition can take place consciously or unconsciously. Conscious metacognition occurs when one, e.g., feels confident or doubtful about one's judgment or decision, when one experiences a 'feeling of knowing' pertaining to a particular piece of information or when one notices a decrease in processing fluency in a task and in all these cases adjusts one's thinking and acting accordingly (e.g., Hart 1965; Brown and McNeill 1966; Koriat 2007; Frith 2012; Proust 2013).

There is evidence that much metacognition can also occur unconsciously. For example, within the cognitive system, tactile and visual information is integrated in a Bayesian, statistically optimal manner, where more weight is assigned to the channel with less

noise, which means that the noise in each channel is unconsciously represented and taken into account (e.g., Ernst and Banks 2002; Alais and Burr 2004).

Further, subjects decelerate their performance in experimental tasks after committing errors even though they aren't aware of committing them (e.g., Chun and Jiang 1998; Spehn and Reder 2000; Logan and Crump 2010; Yeung et al. 2012). They also tend to adjust their behaviour to the current difficulty of a task despite not having any experience of mental effort (Naccache et al. 2005), and make use of unconscious judgments about the reliability of their memory in regulating their performance on recognition tasks (Paulus et al. 2013).

On the basis of these and other kinds of empirical data, including from animal research, it has been argued that unconscious metacognition is prevalent in human and non-human animals (e.g., Frith 2012; Shea et al. 2014). This raises an intriguing question. Given that much metacognition occurs unconsciously, what is the function of *conscious* metacognition?

4.2 The inter-subjective cognitive control view

Frith (2010, 2012) argues that since conscious metacognitive representations, unlike unconscious ones, have the crucial feature of being reportable, they allow us to communicate their contents to others. In doing so, they enable us to discuss epistemic aspects of our perceptual and decision-making processes with others, which enhances social interactions and improves joint decision-making. This, Frith (2012) suggests, is the "main, if not the only, function" of conscious metacognition (2216).

Shea et al. (2014) elaborate the proposal. They hold that when

> sensorimotor systems have to be coordinated between two or more interacting agents it is no longer possible for learning automatically to make use of all and any metacognitive information located anywhere within the different agents: my system 1 [i.e. unconscious] learning processes have direct access to the metacognitive information in my head, but not to the metacognitive information in your head, and vice versa. […] [I]nter-agent control will typically be more effective when it can use metacognitive representations, if relevant metacognitive

> representations within system 1 processes in each agent are selected for broadcast to the other agent, so that decisions about which sensorimotor processes to deploy can be taken in a space of shared metacognitive information. (188-189)

For Shea et al., the function of conscious metacognition is to make metacognitive representations available for inter-subjective cognitive control, i.e. for the controlling of the sensorimotor systems of two or more agents involved in a shared task. For instance, when a metacognitive representation of the reliability of a sensory signal is conscious, then a subject can report it in terms of confidence by saying 'I'm sure' or 'I'm doubtful'. When "agents are cooperating, these reports can be used to optimise control by, for example, giving more weight to the more confident observer" (Shea et al. 2014: 189).

Frith and Shea et al. argue that this is what conscious metacognition was selected for. It was selected for enhancing inter-subjective cognitive control and optimising inter-personally coordinated action. The use of "metacognitive representations about one's own cognitive processes for intra-subjective cognitive control came second, and arose as a side effect" of the selection of conscious metacognition for "inter-personally coordinated action" (Shea et al. 2014: 189). I shall thus call Frith's, and Shea et al.'s proposal on the function and evolution of conscious metacognition the *inter-subjective cognitive control* (ICC) view.

There is increasing empirical support for the proposal. Studies show that pairs of subjects who discuss their confidence ratings about their individual classifications of visual stimuli do significantly better in detecting a subtle visual signal than the best one working on her own (Bahrami et al. 2010; Fusaroli et al. 2012). There is also evidence that the sharing of metacognitive experiences (e.g., a feeling of confidence) allows subjects to develop more accurate models of the world even in the absence of objective feedback (Bahrami et al. 2011; Firth 2012).

Given the intuitive plausibility and empirical support that the ICC view enjoys, it seems a natural question to ask whether it can also provide an account of the function of self-ascriptions of beliefs.

4.3 A dilemma

Frith and Shea et al. propose the ICC view as an explanation of the function and origin of conscious metacognition in general. They don't discuss self-ascriptions of beliefs, however. Nonetheless, self-ascriptions of beliefs are also conscious metacognitive representations. They are conscious second-order judgments about properties of first-order representations, namely about their contents and attitudinal role. Since that is so, the ICC view should also hold for self-ascriptions of beliefs. If the view is applied to second-order judgments such as *I believe p*, then it amounts to the proposal that these conscious metacognitive representations have the function of and evolved for enabling one to report to others that one believes that *p*.

While this proposal is *prima facie* plausible, it leads to a problem. For metacognitive representations seem unnecessary for the purpose just mentioned. If I want to know whether S believes *p*, I can arguably just ask her whether *p*. S's sincere answer to the world-directed question will reliably tell me what she believes about *p*. And since the question as to whether *p* is a first-order, world-directed question, S herself will be able to find the answer that provides me with an insight into her beliefs by engaging just in first-order reasoning to settle whether *p*. No self-ascriptions of beliefs and conscious metacognition on her part are required.

Since that is so, it becomes less obvious why it should be the function of self-ascriptions of a belief *p* to enable one to broadcast to others that one believes *p*. If these kinds of conscious metacognitive representations evolved for this purpose, then some selection pressure would need to have been in place. But since others seem already able to find out whether one believes *p* via one's first-order answers on whether *p*, where could that pressure have come from?

The following rejoinder comes to mind. There might be situations when S's sincere answer to a world-related question doesn't reveal what she believes. For instance, suppose S is charged with murder and at the end of the trial the evidence on whether she committed the crime remains inconclusive either way. Suppose that when the jury retreats for deliberation on the verdict, S asks her lawyer 'Am I guilty?' Since the evidence is inconclusive with respect to the facts and her lawyer (by assumption) acknowledges that this is so, he might sincerely respond that whether she is guilty

hasn't been determined yet.[60] However, if S asked him instead 'Do you *believe* I'm guilty?' then most likely he will react quite differently. He will then express his own judgment on the evidence, his intuition on the issue and on her culpability. The answer to the question in the latter case will provide S with an insight into her lawyer's mind that will allow her to predict his behaviour when he is, e.g., asked to place a bet on a guilty or not guilty verdict. In contrast, his answer to the former question won't do so.

The point that a subject's sincere answer to a world-related question might not always reveal what she believes is due to the fact that belief formation isn't only sensitive to epistemic, truth-related reasons and evidence, but also to non-doxastic factors, e.g., affective states, intuitions, biases etc. As a result, in general, when upon reflection on whether *p*, one sincerely answers the factual question as to whether *p*, there might still be room for a difference between what one asserts and what one believes on the matter. If that is right then, in line with the ICC view, it could be correct after all that the function of self-ascriptions such *I believe p* is to enable one to broadcast to others that one believes *p*. For one's first-order judgment and answer on whether *p* won't always suffice to provide other people with an insight into one's belief about *p*.

But now another problem arises. Above I argued at length that we often form self-ascriptions of beliefs via TM, i.e. we find out about whether we believe *p* by asking ourselves whether *p*. If our sincere answer to the first-order question as to whether *p* won't necessarily reveal what we believe when others ask us, then why do we nonetheless ask ourselves the very same question to find out what we believe?

It seems the ICC view of the function of self-ascriptions of beliefs leads to the following dilemma. If the advocate of the view grants that one's sincere answer to the question as to whether *p* reveals one's belief on whether *p* to others, then since answering this question doesn't require that one forms a self-ascription of a belief about *p*, s/he would grant that one can reliably convey to others one's belief about *p* without conscious metacognition. The ICC proposal that self-ascriptions of beliefs have the function of enabling one to communicate to others that one believes *p* (or not-*p*) would remain unmotivated; they wouldn't be needed for this purpose. If alternatively the advocate of the ICC view does not grant that one's sincere answer to the question as to whether *p* reveals one's belief about *p*, then the ICC view would contradict the plausible

[60] He might add that this is an odd question to hear from *her*, the alleged perpetrator.

claim that we typically use TM as a means for working out our own belief on *p*. It seems that either way, the ICC view of the function of self-ascriptions of beliefs remains unsatisfactory.

4.4 The dilemma resolved

Since the ICC account of the function of self-ascriptions of beliefs is *prima facie* very plausible, it is desirable to find a solution to the dilemma. I will use the rest of this section to introduce one.

The first step is to draw a distinction between two different kinds of information on whether *p*, between what I shall call *inter-subjective* and *intra-subjective* evidence on whether *p*.

Inter-subjective evidence on whether *p* is information pertaining to *p* that counts as evidence for (or against) *p* and is also accepted *ceteris paribus* in one's social environment as evidence for (or against) *p*. For instance, if S claims that whales are mammals and justifies her claim by holding that whales lactate and only mammals lactate, then her claim is based on inter-subjective evidence because it is based on information that is evidence for whales being mammals that is also accepted in her social environment as evidence for whales being mammals.

In contrast, *intra-subjective* evidence on whether *p* is information pertaining to *p* that might count as evidence for (or against) *p* by one's own lights but doesn't do so by the standards of others. For instance, if S claims that whales are mammals on the basis of an affective state or an intuition that she herself finds compelling but can't further justify to others, then her claim is based on intra-subjective evidence. For it is based on information that others won't accept as evidence for whales being mammals, even though for her, it is. Intra-subjective evidence includes contents carried by non-doxastic states, e.g., affective states, intuitions, implicit biases, etc.

During development, people learn to keep inter-subjective and intra-subjective evidence apart. Crucially, keeping the two apart doesn't require awareness of one's own mind. It only requires awareness of worldly, non-mental states of affairs and of how the obtaining of them is supported.

For instance, when a subject S wonders whether *p*, she will turn to the world to settle the matter. She will appeal to states of affairs whose obtaining speaks either for or against *p*. The obtaining of some of them will be supported in a way that other people acknowledge. For example, others acknowledge the obtaining of the state of affairs that only mammals lactate because it is a scientific fact about mammals. States of affairs of this kind will for S subsequently count as inter-subjective evidence for or against *p*. Any remaining states of affairs that from her point of view obtain and speak to the question of whether *p* will then count as intra-subjective evidence for or against *p*.

To keep inter-subjective and intra-subjective evidence apart thus only requires keeping track of the nature of the support for the obtaining of states of affairs. This doesn't presuppose self-ascriptions but only other-ascriptions of mental states. It only requires keeping track of what others accept or would, given their commitments, accept as being the case. When this ability is in place and exercised, the distinction between the two kinds of information is established. For intra-subjective evidence can then be negatively demarcated from inter-subjective evidence.

Subjects who distinguish between inter-subjective and intra-subjective evidence might bring both kinds of information to bear when answering the question as to whether *p*. The activation and use of either, or both, will depend on the source of the question. There are two different sources of the question as to whether *p*, for the question might originate from other subjects or from oneself. I will now argue that if the question originates from others, then only inter-subjective evidence will be consulted for answering it. If the source of the question is oneself, however, then both inter-subjective and intra-subjective evidence will be employed to settle the matter. The motivation for assuming that different kinds of information are brought to bear in answering one and the same question is the following.

Suppose that when other people ask S whether *p*, then she responds, on the basis of *intra*-subjective evidence on *p*, that *p*. Suppose further that they subsequently rely on S's claim in the pursuit of their own interest. Suppose finally that, as it turns out, *p* is not the case, and as a result they fail in their project.

It seems clear that given S's involvement in the events that led to this outcome, she will

be held responsible. S will be asked about her support for her claim that *p*, and her support for the claim will be subject to scrutiny by others, where the result of the assessment will determine the kind of criticism that she will eventually receive.

The problem for S is that there will be no shared basis between her and other people when it comes to the support that she might cite for her claim in order to mitigate their subsequent criticism of her. For, by assumption, the information S employed to arrive at her claim was intra-subjective evidence only, and intra-subjective evidence serves at best for her as support for *p*, but not necessarily also for others.

Compare this scenario with the following situation. S is again asked whether *p* but now to answer the question, she doesn't use intra-subjective but *inter*-subjective evidence on whether *p*. Two things will be different. First, she will be less likely to give a false answer to begin with, for inter-subjective evidence is information accepted within one's social environment as evidence for *p*. Since that is so, the probability of it being correct is higher (Pollock 1982: 142). Second, and more importantly, in the case of a false answer, there will be a shared basis between S and others with respect to the support pertaining to whether *p*, because inter-subjective evidence for (or against) *p* is information that counts as evidence for (or against) *p* by both her own and others' lights. As a result, others might in the event of a mistaken claim find no fault on her part and suspend criticism. Hence, if S draws on inter- rather than intra-subjective evidence in her search for an answer to the question as to whether *p* then she will reduce the risk of negative consequences. Given the different outcomes that the use of one particular kind of information might imply, one would expect that when others ask S whether *p*, she will in her search for an answer tend to employ inter-subjective evidence only.

Intra-subjective evidence might then still unbeknownst to her feed into her search for an answer, and when her cognitive resources are drained (e.g., when she is under time pressure, when her attention is engaged elsewhere, when she is tired etc.), her tendency to consult inter-subjective evidence might be disengaged and intra-subjective evidence (e.g., biases) determines her response. Nonetheless, in situations when resources are available, then, given the possibly detrimental consequences of her relying on intra-subjective evidence, S will be disposed to employ only inter-subjective evidence to settle the matter as to whether *p*.

One consequence of this is that when others ask S whether *p*, her sincere answer won't necessarily provide them with an insight into what she takes to be the case when, e.g., the inter-subjective evidence is inconclusive. The above example of the lawyer whom S asks about her guilt illustrates this. In fact, whenever S holds beliefs that are based on non-doxastic factors, e.g., affective states, intuitions, biases etc., which will arguably often be the case, as people are generally not "model epistemic citizens", whose beliefs are what they rationally should be (Cassam 2014: 83), others will be less likely to find out about S's beliefs[61] via asking her whether *p*.

But consider now the case when S asks *herself* the same question, e.g., in the context of TM. Since in self-directed queries as to whether *p*, there is no one else to whom S is responsible and by whom she could be reprimanded if her answer turns out to be false. The just-mentioned basis for keeping apart inter-subjective and intra-subjective evidence is hence absent.

As before, S will be more likely to answer correctly if she relies in her search for an answer on inter-subjective rather than intra-subjective evidence. However, importantly, unlike when others are the source of the query, in the case of a self-generated and self-directed question, it will now in fact be disadvantageous for her to disallow intra-subjective evidence (e.g., affective states, intuitions, biases etc.) to feed into her answer to the factual question. For if it turns out that the inter-subjective evidence is inconclusive, then her obtaining *some* answer will still be better than getting none, because it will resolve uncertainty on the matter. This is critical, because the point of beliefs (in conjunction with desires) is to enable one to act in ways that lead to the satisfaction of one's desires, and any proposition that one hasn't marked as either true or false but left truth-value undetermined is of little use for action-planning purposes, as one can't rely on it either way, i.e. as either a true or false proposition. In the case of a self-directed query, drawing upon intra-subjective evidence when determining whether *p* has thus the significant benefit of settling matters and enabling action. Hence, there is

[61] It might be suggested that when S consults inter-subjective evidence to answer other people's question as to whether *p* and finds that *p*, then she will *also* have the belief that *p*, and so others will gain at least an insight into this particular belief of hers. However, inter-subjective evidence might indicate to S that *p* and lead her to sincerely respond that *p* even though she herself remains unconvinced by it, e.g., because for her the intra-subjective evidence for not-*p* is stronger. In this case, she won't then believe *p* at all despite sincerely claiming that *p*. In such a case, her response won't correspond to any of her beliefs.

reason to expect that when the source of the question as to whether *p* is oneself, then both kinds of information will be employed to resolve the matter.

In the typical case, when cognitive resources are available, inter-subjective evidence will be brought to bear until the point when it is exhausted, and then intra-subjective evidence is consulted to settle the issue one way or the other. When resources are drained, as before, intra-subjective evidence might become effective earlier and trump inter-subjective considerations.

The points just made help solve the dilemma mentioned above. The problem resulted from the combination of the following two views.

(1) If the advocate of the ICC account of the function of self-ascriptions of beliefs grants that one's sincere answer to the question as to whether *p* is sufficient to provide others with an insight into one's belief about *p*, then since answering this question doesn't require that one forms a self-ascription of a belief, the advocate of the ICC view would grant that one can reliably verbally convey to others that one believes *p* without conscious metacognition. The ICC proposal that self-ascriptions of beliefs have the function of enabling one to tell others that one believes *p* (or not-*p*) would remain unmotivated because self-ascriptions of beliefs wouldn't be required for this purpose.

(2) If alternatively the advocate of the ICC view of the function of self-ascriptions of beliefs doesn't concede that one's sincere answer to the question as to whether *p* reveals one's belief about *p*, then the ICC view is at odds with the plausible claim that we typically use TM to find out whether we believe *p*.

With the preceding discussion in mind, the tension between (1) and (2) can be resolved in the following way.

When the source of the question as to whether *p* is other people, then it is part of a safeguarding strategy to only draw on inter-subjective evidence when forming an answer. The result is that one's answer won't necessarily provide others with an insight into what one believes, and how one will act when, e.g., the inter-subjective evidence is inconclusive, or one's belief on the matter is based on intra-subjective evidence. So, in response to (1), one's sincere answer to the question as to whether *p* won't always be

sufficient to provide others with an insight into one's belief about *p*. The ICC proposal that self-ascriptions of belief are required for enabling the verbal communication of one's own beliefs to others remains thus intact.

But even though one's answer to the question as to whether *p* won't necessarily reveal what one believes when others ask the question, when one asks oneself the very same question then, in cases when one's inter-subjective evidence is inconclusive or one's belief on the matter is based on intra-subjective evidence, one will be able to find out what one believes about whether *p*. The reason is that when the source of the question as to whether *p* is oneself, the mentioned safeguarding strategy is disengaged from the outset. For in the case of an incorrect answer, there is no one present by whom one could be rebuked, as intra-subjective evidence on *p* is evidence for *p* that one already acknowledges as such. Furthermore, when the source of the question is oneself, then the use of intra-subjective evidence in one's search for an answer will in some cases have the additional and crucial benefit of terminating uncertainty as to whether *p*. This is why in the case of a self-directed query as to whether *p*, i.e. in the context of TM, the question 'Is *p* the case?' will be answered by reference to both inter-subjective and intra-subjective evidence. As a result, the answer to the first-order query will provide one with an insight into one's own belief about whether *p*. This deals with point (2).

Both (1) and (2) are hence compatible, and the initial dilemma is resolved. Supplemented with the two points just made, the ICC view offers an account of the function of self-ascriptions of beliefs that integrates TM as a method for forming these self-ascriptions. I shall refer to the resulting proposal on the function of self-ascriptions of beliefs as the *ICC-TM view*.

4.5 The ICC-TM view and IDR-ISA hybrid view

To further develop the ICC-TM view, I will now relate the proposal to the TM-based account of privileged self-knowledge of beliefs defended in previous chapters. It might seem that the former is in some respects incompatible with the latter. But I will argue that this is not the case.

For instance, the IDR-ISA hybrid view holds that in the context of TM one comes to know whether one believes *p* by applying to oneself a principle that one already uses to

work out other people's beliefs, namely (IDA): What subjects will say in response to the question of whether *p* is what they believe about *p*. It might seem that if the ICC-TM view were right, then one perhaps wouldn't assume (IDA). Rather, one would assume that what subjects will say in response to the question as to whether they *believe* that *p* reveals their belief about *p*.

However, there is no incompatibility here between the ICC-TM view and the IDR-ISA view, because on the ICC-TM view, other people's sincere response to, e.g., my question as to whether *p* will often still be sufficient for me to come to know whether they believe *p*. The initial principle for working out other people's beliefs (i.e. the principle that what subjects will say in response to whether *p* is what they believe about *p*) isn't undermined. It is only that another one is added to the repertoire of generalisations available to determine what people believe.

And, crucially, on the ICC-TM view, the need for me to ask another subject S the second-order question 'Do you *believe* that *p*?' to find out what she believes about *p* only arises because when I'm asking her whether *p*, then, for her, my probe is *other-sourced*. This is a relational fact that leads her to impose a constraint on her information processing (namely to only consult inter-subjective evidence) that has to be disengaged in order for me to be able to attain an insight into her belief when her belief is based on intra-subjective evidence.

The consideration also provides an answer to the question as to why I don't then ask myself similarly the *belief*-related question to find out my belief about *p* but rather the world-related one. I don't ask myself the belief-related question because I understand the following two things.

First, the only point of asking subjects the belief-related question, rather than the world-related one, is to disengage the mentioned constraint on their information processing and prompt them to settle whether *p* by also consulting intra-subjective evidence (it is to provide them with epistemic leeway, as it were). Second, when a subject asks *herself* the world-related question, then the constraint isn't, for the reasons mentioned above in response to (2), imposed to begin with. Note that an understanding of these two points doesn't require any kind of meta-representation of my own mental states or processes. It only requires theorising about other minds.

Once I understand the two points, then when I want to know my own belief about *p*, I won't ask the belief-related question but simply the world-related question. For in my own case the belief-related and the world-related questions would have the same effect. Asking the belief-related question would only require me to take an extra step, namely to move from the belief-related question back to the world-related one. Since this is redundant, when I want to find out whether I believe *p*, I will only ask myself the first-order question as to whether *p*, just as the IDR-ISA hybrid view holds. The ICC-TM view is thus compatible with the IDR-ISA hybrid view.

Still, it might be objected that the ICC-TM view assumes that subjects can draw on intra-subjective evidence to settle whether *p*, and this is at odds with the transparency method, as the IDR-ISA hybrid view describes it. For according to the IDR-ISA hybrid view, in the context of TM, one is supposed to be able to find out whether one believes *p* without attending to one's own mental processes or states simply by turning to facts of the external world. Yet, the argument continues, in accessing intra-subjective evidence one is accessing one's own, e.g., affective states, biases, intuitions etc., and therewith one is attending to one's own mind rather than to facts of the external world, or so the argument concludes.

The argument rests on a misunderstanding, however. There are two reasons for this.

First, the claim concerning TM is that I can determine whether I believe *p* "by putting into operation whatever procedure I have for answering the question whether *p*'" (Evans 1982: 225). And as mentioned, intra-subjective evidence is already part of the procedure that I use to answer whether *p*, because when I want to know whether *p* and inter-subjective evidence on the matter remains inconclusive, I will need to settle the issue somehow. It is at this point that intra-subjective evidence enters into my first-order belief formation. Since intra-subjective evidence already figures in first-order belief formation and is part of the procedure I have for answering the question whether *p*, there is nothing in the appeal to this kind of data that is at odds with transparency.

Second, it is not the case that when I access intra-subjective evidence in my first-order reasoning to settle a matter, I then attend to aspects of my own mind in the sense that I represent something mental as such – only this sense of 'attending to one's own mind' is in the context of TM problematic. For suppose that in my reflection on whether *p*, I

find that there is no inter-subjective support for or against *p* but it nonetheless strikes me that *p*, e.g., because I have the intuition that *p*. Suppose I then use *p* in my reasoning as support for *q*. When I do so, I don't represent my intuition that *p* as such. I simply represent *p*. As argued above, I'm able to keep *p* apart from inter-subjective evidence, and selectively consult it, because intra-subjective evidence is negatively differentiated from inter-subjective evidence. It is evidence for me for the obtaining of a state of affairs (e.g., *q*) that others would not acknowledge as such. This is the property of *p* that makes *p* for me intra-subjective evidence for *q*. It enables me to selectively draw upon intra-subjective evidence without representing aspects of my own mind as such as opposed to (for me) facts of the world. As a result, there is no conflict between the view that I can come to find out about my own beliefs by accessing intra-subjective evidence and the view that I can find out about my beliefs without attending to my own mind. For in accessing intra-subjective evidence I am attending to states of affairs of the external world that for me obtain but that I can't show to others to obtain in a way that they would accept as sufficient. There is thus again no incompatibility between the ICC-TM view of the function of self-ascriptions of beliefs and the IDR-ISA hybrid view's component pertaining to TM-based privileged self-knowledge.

5. The function of privileged self-knowledge

In the preceding section, I introduced an account of the function of self-ascriptions of beliefs, the ICC-TM view. I also argued that the ICC-TM view is compatible with the IDR-ISA hybrid view. Now, according to the IDR-ISA hybrid view, TM-based self-ascriptions lead to privileged self-knowledge of beliefs. With this in mind, it is time to turn from the question of the function of self-ascriptions of beliefs, in general, to the initial, more specific question of the function of TM-based privileged self-knowledge of beliefs. Before addressing the question, a clarification on the ICC-TM view is in order.

Above I proposed that the function of self-ascriptions of beliefs in general is to enable one to report one's own beliefs to others. There is one point about this proposal that I haven't yet made explicit but that will become relevant below. It is that, on the ICC-TM view, in competitive or hostile social environments, self-ascriptions of beliefs are unlikely to emerge. The reason is that there will be little selection pressure on oneself to make it easier for others to attain an insight into one's own beliefs. In fact, in competitive social environments, one would become vulnerable to exploitation,

deception, and manipulation by others if one made it easy for them to determine one's beliefs. On the ICC-TM view, in competitive social environments, self-ascriptions of beliefs are thus unlikely to come into being. A cooperative social environment is a prerequisite for the emergence of self-ascriptions of beliefs. For it is only when others are cooperative that the ability to report one's own beliefs has adaptive advantages, as this will then help improve inter-subjective control, which in turn leads to better joint decision-making and group coordination with mutual benefits for both self and others.

Suppose, then, that in cooperative environments there is selection pressure for self-ascriptions of beliefs, and self-ascriptions of beliefs come to acquire the function of enabling the report of one's own beliefs, just as the ICC-TM view proposes. Suppose further that we now adopt the point of view of evolution, and explore how best to produce the mechanism to implement the function at issue via natural selection.

Note first that self-ascriptions of beliefs can be formed in many different ways, and so the function at issue can be performed in many different ways. They might, for instance, be formed by interpreting oneself, by drawing inferences from evidence of one's belief, or non-inferentially. Clearly, some ways of forming self-ascriptions of beliefs are better than others when it comes to enabling one to report one's own beliefs. And some methods of forming self-ascriptions of belief will be easier to evolve from the already existing cognitive machinery and resources than others. The least computationally complex and most reliable method of forming self-ascriptions of beliefs that is also the easiest to design from pre-existing structures would be the best to serve the purpose. It is plausible to assume that it would be the one that evolution is most likely to design via natural selection.

Recall the formation of self-ascriptions of beliefs via TM, as the IDR-ISA hybrid view describes it, is non-inferential in nature, and based on one's direct access to one's own judgments in first-order thinking. Since that is so, as argued in Chapter 4, it will be a highly economical and reliable method for the formation of self-ascriptions of beliefs. Furthermore, the IDR-ISA hybrid view proposes that TM-based self-ascriptions of beliefs only recycle already existing structures and mechanisms. These structures, e.g., a grasp of (IDA), and mechanisms, e.g., the mindreading system, just need to be applied to oneself. Thus, among the different procedures for forming self-ascriptions of beliefs, the formation of self-ascriptions via TM, as the IDR-ISA hybrid view describes it, turns

out to be both the most efficient and reliable procedure as well as the easiest one to design and evolve. Since that is so, it becomes plausible to assume that in cooperative social environments the ability to form TM-based self-ascriptions will be selected. In such environments, TM-based self-ascriptions, formed in the way the IDR-ISA hybrid view proposes, and so instances of privileged self-knowledge of beliefs, will acquire the function to enable the most economical and reliable verbal communication of one's own beliefs to others.

The preceding offers an account of the evolutionary function of both self-ascriptions of beliefs, in general, and TM-based self-ascriptions, and so privileged self-knowledge of beliefs, in particular. I shall henceforth treat the just introduced proposal on the function of privileged self-knowledge of beliefs as a component of the ICC-TM-view.

6. Conclusion

In this chapter, my goal was to find an answer to the question of why there is such a thing as privileged self-knowledge of beliefs. I focused initially on the question of why we self-ascribe beliefs and then later returned to TM and the function of privileged self-knowledge. I first introduced and rejected three extant answers to the question of why we self-ascribe beliefs before developing my own account on the matter. The account I developed builds on a recent cognitive-scientific proposal on the function of conscious metacognition, which holds that the function of conscious metacognition consists in facilitating inter-subjective cognitive control (ICC). I argued that, as it stands, the ICC view leads to a dilemma when it is applied to self-ascriptions of beliefs formed via TM. Since the view struck me as otherwise highly plausible, I set out to solve the dilemma. The solution I offered manages to preserve both the ICC view on the function of self-ascriptions of beliefs, and the assumption that TM is often the method for forming them. According to the resulting proposal, the function of self-ascriptions of beliefs, in general, is to enable the report of one's own beliefs to others. The function of TM-based self-ascriptions, and so privileged self-knowledge of beliefs, in particular, is then derived from this. It is to enable the most efficient and safest verbal communication of one's own beliefs to others via belief reports.

CHAPTER 7

Conclusion – A four-component dual method theory

Over the course of the preceding chapters, I developed a theory of self-knowledge of beliefs that answers the three questions about privileged self-knowledge of beliefs with which this thesis started. I will now assemble the theory from its four components, which were introduced in the different chapters. I show how its last component, the ICC-TM view, connects the answers that the account offers to the three questions mentioned at the beginning of the thesis and also yields a final evolutionary argument for the overall account. I end by bringing out how the theory of self-knowledge of beliefs that emerges from this thesis supports the view that privileged self-knowledge of beliefs is grounded in the seemingly unrelated fact that other people's beliefs are opaque, i.e. only interpretively accessible, to us.

1. Putting it all together

The goal of this thesis was to answer three questions about privileged self-knowledge of beliefs: whether there is such a thing as privileged self-knowledge of beliefs, if there is, how it can be explained, and why it exists at all. To find answers to these three questions, which I called the *whether*-question, the *how*-question, and the *why*-question, my strategy was to take TM as a starting point.

One central reason for taking TM as a starting point was that the method suggests that privileged self-knowledge of beliefs can be explained without any need to postulate an 'inner sense' or some other extra mechanism in addition to the mechanisms already needed in other domains than self-knowledge. This made TM attractive, for the method promises to allow explaining privileged self-knowledge of beliefs in a highly theoretically and ontologically parsimonious way.

However, I noted in Chapter 1 that if a TM-based account of self-knowledge is to be developed that answers all three of the above questions about privileged self-knowledge of beliefs, then four problems need to be solved. In Chapters 2 to 6, I provided solutions to these problems. The solutions I offered form the individual components of the overall

theory of self-knowledge of beliefs that I introduced in the thesis. They were the following four proposals:

(1) the DA view (Chapter 2),
(2) the IDR theory (Chapters 3 and 4),
(2) the IDR-ISA hybrid view (Chapter 5), and
(4) the ICC-TM view (Chapter 6).

Since the overall account of self-knowledge of beliefs that consists of (1) to (4) holds that subjects have two different ways of acquiring self-knowledge of beliefs, one interpretive/inferential and the other non-inferential in nature, I shall refer to it as the *four-component dual method* (FDM) *theory* of self-knowledge of beliefs. Let me briefly summarise each of the four parts of the FDM theory and show how they are interconnected.

(1) The DA view (Chapter 2)

The first component of the FDM theory is the direct access (DA) view. It concerns first-order thinking and holds that we have non-inferential access to our own judgments in conscious first-order thinking. Recently, it has been argued that the DA view is false and that we have only indirect access (IA) to our own judgments via interpretation-dependent intermediaries (e.g., sensory-imagistic representations).

I contended that the case for the IA view and against the DA alternative doesn't in fact undermine the latter. Furthermore, I argued that considerations on conscious theoretical reasoning and empirical data on autism indicate that the IA view currently defended in the literature is false. While the falsity of the IA view that is currently advocated in the literature doesn't yet show that the DA proposal is correct, I maintained that the autism data in particular provides positive support for the claim that the DA view is correct.

The positive support for the DA view matters for the FDM theory, because the latter is in part a transparency account. It holds that one can come to know in a privileged way whether one believes *p* simply by determining whether *p*. Since that is so, the FDM theory depends on the DA view in the following two ways.

In the context of TM, one's determining whether *p* involves conscious first-order thinking. Suppose we had in conscious first-order thinking only indirect access to our own judgments via intermediaries (e.g., inner-speech tokens) that express them and first need to be interpreted, and their underlying attitudes self-ascribed, in order to acquire attitude-like functional roles.

If this were the case, then the conscious thinking involved in TM would similarly already require interpretative processing and self-ascriptions of attitudes. It would already require representing aspects of one's own mind as such. This would undermine a TM-based theory of privileged self-knowledge of beliefs. For such a theory holds that one can come to know one's own belief *p* without having to represent aspects of one's own mind prior to the formation of the self-ascription of the belief *p*. Since the FDM theory is in part a TM-based account, this means that it depends on the assumption that the IA view is mistaken and the DA view correct.

In addition, transparency theories hold that subjects come to self-ascribe a belief *p* on the basis of a judgment *p*. Since the proposition *p* might be judged, supposed, doubted etc., any such proposal needs to say how, in the context of TM, the judgment *p* can be detected and differentiated from these other attitudes without the involvement of any representation of the judgment itself.

The DA view provides an explanation, for it includes an argument to the effect that the direct access that we have to our own judgments in first-order thinking involves the operation of a sub-personal mechanism that detects and differentiates the attitude types of representations without representing the attitudes themselves. On the DA view, the mechanism at issue is able to do so by tracking the neural properties of the representations. The DA view thus solves the mentioned problem pertaining to judgment-detection without meta-representation. This is the second way in which the FDM account depends on the DA view.

(2) The IDR theory (Chapters 3 and 4)

The DA view is needed to prepare the ground for a transparency theory of privileged self-knowledge of beliefs. To further set the stage for such a theory, I then introduced

and motivated a set of conditions that an adequate transparency theory should meet. They were the following two conditions.

> (AC1) Any adequate account of self-knowledge of beliefs on which TM provides one with self-knowledge needs to solve the intelligibility puzzle, i.e. it needs to explain how a subject S can come to determine whether *p* in order to determine whether she believes *p* even though she understands that whether or not *p*, this won't tell her whether she believes *p*. Furthermore, the account shouldn't presuppose that S is already able to represent her own mental processes or states without explaining how she is able to do so.

> (AC2) Any adequate account of self-knowledge of beliefs on which TM provides one with self-knowledge needs to solve the knowledge puzzle, i.e. it needs to explain how the self-ascription *I believe p* that is thought to result from applications of TM can be justified and true.

I argued that a wide range of extant transparency theories fails to meet (AC1) and (AC2). In response to this, I then developed an alternative, what I called the *implicit dual-reasoning* (IDR) *theory*. The IDR theory is the second and also the central component of the FDM account. It holds that TM relies on resources used for the formation of other-ascriptions of beliefs. To work out whether other people believe *p*, one tends to ask them whether *p*. In doing so, one's underlying assumption is that when a subject S is asked whether *p*, her response will express what she believes on the matter. According to the IDR theory, one applies this basic idea to oneself. In order to find out whether one believes *p,* one asks oneself whether *p*, and on the basis of the response to the probe, one then self-ascribes a belief on the matter. Crucially, however, unlike when one wants to work out whether another subject S believes *p* via asking her whether *p*, one's own response to the question as to whether *p* needn't first become expressed (e.g., overtly or in inner speech), because one has already an insight into what one will say via one's decision to say it. According to the IDR theory, the decision to say *p* gives rise to an epistemic state that leads one to predict that one will say *p*, and to form non-conditional judgments that follow from the assumption that one will say *p.* On the basis of both (a) what one will say in response to the question of whether *p*, and (b) the principle that what a subject says in response to the question as to whether *p* is what she believes about *p*, one can then self-ascribe a belief about *p*. Thus, according to the

IDR theory, a combination of practical reasoning (to settle what to say in response to the probe about whether *p*) and theoretical reasoning (to infer what one believes on the basis of what one will say in response) makes the transition involved in TM intelligible. Furthermore, the reasoning at issue allows doing so without presupposing any kind of representation of one's own mental processes or states. The IDR theory hence meets (AC1).

In order to qualify as an account of self-*knowledge* of beliefs, however, the IDR theory needed to meet (AC2) also. I thus supplemented the proposal with an externalist account of epistemic justification, according to which S is justified in forming a belief *p* if she (i) forms the belief because she is undergoing a mental event *M*, and (ii) *M* reliably correlates in S with the type of state of affairs that makes her belief *p* true. I argued that in the context of TM, both conditions are met. For in applications of TM, self-ascriptions are formed on the basis of judgments, and there is a reliable correlation between these judgments and beliefs with the same content. I maintained that in cases when the judgment *p* correlates with the belief *p*, a TM-based self-ascription thus plausibly qualifies as self-knowledge of the belief *p*. Since that is so, the IDR theory also manages to satisfy (AC2) and so qualifies as an account of self-knowledge of beliefs.

Finally, I argued that even though the IDR theory appeals to various inferences to make the transition involved in TM intelligible, the account holds that typically, i.e. in unreflective self-attributions, most components of the reasoning (e.g., the transition from *p* to the decision to say *p*, or the transition from *I will say p* to *I believe p*) remain implicit. In its entirety, the dual reasoning that the IDR theory invokes need only be performed once, where the mentioned components might occur merely unconsciously. Once this has happened, the inferential links via which one can move from *p* to the judgment *I believe p* are established. This allows the system responsible for self-attributions of beliefs in subsequent applications of TM to suspend the transitions from *p* to the decision to say *p* and from the epistemic state with the content *I will say p* to the self-ascription *I believe p* and move directly from the outcome of one's determining whether *p* to the self-ascription. The assumption that the system does indeed refrain from executing the other components of the dual reasoning is according to the IDR theory supported by evidence on judgment and decision-making. The data at issue suggests that, in general, judgment-forming systems operate under resource constraints

(e.g., limits of attention) that force them to adopt mental short cuts (heuristics) to make the performance of tasks more computationally economical. I argued that the system responsible for self-attributions will be similarly constrained and hence, in the context of TM, reduce computations in the way the IDR theory proposes. All that then remains of the dual reasoning that was introduced to make TM intelligible is the probe as to whether *p*, the reasoning involved in settling whether *p*, and a non-inferential, purely causal transition (implemented by a simple detection mechanism) from the outcome of one's settling whether *p* to the self-ascription of what one believes on whether *p*.

I maintained that since settling whether *p* and the disposition to move directly from *p* to *I believe p* don't make TM on the IDR view inferential, TM-based self-ascriptions of beliefs turn out to be non-inferential in nature. The IDR account thus shows that and how TM-based self-ascriptions can be non-inferential. As a result, the account is not only able to capture the intuition that TM involves a direct transition. It also manages to explain the *privileged* nature of self-knowledge of beliefs, or so I argued.

The IDR theory is superior to a wide array of currently available transparency theories. For the latter typically don't recognise let alone offer a solution to the two problems[62] that the DA view addresses and that the IDR theory is able to tackle, as it contains the DA view as a component. Furthermore, while extant transparency theories don't manage to meet both (AC1) and (AC2) (see Chapter 3), the IDR theory is able to do so.

(3) The IDR-ISA hybrid view (Chapter 5)

Still, there are other views of self-knowledge of beliefs that don't rely on TM and might be preferable to the IDR theory. In addition, there is empirical evidence pertaining to self-interpretation and confabulation that suggests that we in fact lack privileged self-knowledge of attitudes. If this is right then the IDR view, which holds that we do have such knowledge at least of beliefs, will be wrong. To defend the IDR theory, the just-mentioned two points thus had to be taken into consideration.

In response to them, I equipped the overall account, which so far consisted of the DA view and the IDR theory, with a third component, a hybrid view of self-knowledge of

[62] Byrne (2012) addresses the second problem, but his response overlooks the first one and as a result remains unconvincing.

attitudes in general. The latter combines the IDR account and a revised version of Carruthers' (2011) *interpretive sensory-access* (ISA) *theory*, which holds that we come to know our own attitudes via interpretive processing on the basis of, e.g., our own behaviour, imagery, or circumstances. The reason for opting for this particular combination of proposals was that the ISA theory is currently perhaps the empirically and theoretically best-supported general account of self-knowledge of attitudes. It relies on data that is typically taken to undermine the existence of privileged self-knowledge of attitudes (e.g., findings on self-interpretation and confabulation). While this might seem at odds with the IDR theory, I argued that neither the data at issue nor the theoretical support for the ISA theory in fact threatens the IDR account. I maintained furthermore that the latter offers a preferable explanation of TM-based self-ascriptions of beliefs, in particular. This point led me to propose the IDR-ISA hybrid view.

The IDR-ISA hybrid view holds that the acquisition of self-knowledge of attitudes requires turning onto oneself the interpretive mechanism that evolved for other-knowledge of attitudes, i.e. the mindreading system. But even though the system typically operates interpretively, it can also produce privileged self-knowledge of beliefs. This is because when one has executed the dual reasoning once, the mindreading system is functionally re-structured so that it can employ the attitude-type detection mechanism involved in first-order reasoning (DA view) to also produce non-inferential TM-based unreflective self-attributions.

The support for holding that the mindreading system comes to suspend interpretive processing in producing TM-based self-ascriptions is the same as the support for holding that most of the dual reasoning that the IDR theory postulates isn't typically performed when one applies TM. It is the empirical finding that the human cognitive system tends to refrain from activating computationally complex processing (e.g., interpretive or inferential processing) if simpler methods for achieving the task at hand are available. That is, the fact that the human cognitive system operates, due to the limitations on its information-processing resources, with 'cognitive miserliness' is at the heart of the argument for the claim that TM-based self-ascriptions of beliefs are non-inferential in nature and amount to privileged self-knowledge.

The IDR-ISA hybrid view has a number of advantages, for it combines the theoretical and empirical support of both theories that it includes. Moreover, I argued that both of

the accounts that it brings together benefit from the integration. The ISA theory attains a more plausible explanation of TM-based self-ascriptions, and the IDR theory, which is only a belief-specific account, becomes embedded in a well-motivated general theory of self-knowledge of attitudes.

The resulting IDR-ISA hybrid view is distinct from any other account of self-knowledge of attitudes currently defended in the literature. For theorists who assume that self-knowledge of attitudes is the result of the mindreading system being turned inward onto oneself tend to think that this means that self-knowledge of (non-perceptual, non-affective) attitudes is always interpretive in nature (e.g., Dennett 1992; Gopnik 1993; Happé 2003; Cooper 2007; Williams and Happé 2010; Carruthers 2009a, 2011; Mandelbaum 2014). And theorists that assume that self-knowledge of attitudes is at least sometimes privileged in nature tend to postulate a special system in addition to the mindreading faculty to account for these cases of self-knowledge (see, e.g., Nichols and Stich 2003; Robbins 2004; Goldman 2006; Cassam 2010). The IDR-ISA hybrid view defended here occupies a middle ground between these two opposing positions.

(4) The ICC-TM view (Chapter 6)

What the IDR-ISA hybrid view leaves unexplained is why there is such a thing as privileged self-knowledge of beliefs at all, where this is understood as a question about the evolutionary function of such knowledge. I argued that, unlike it is often assumed, self-ascriptions of beliefs are not primarily required for one's own cognition, e.g., for one's beliefs or judgments to be conscious, for executive functioning, rationality, or moral responsibility. After that, I developed an alternative that is built on a recent cognitive-scientific proposal on the function of conscious metacognition, i.e. the *inter-subjective cognitive control* (ICC) *view*. More specifically, I argued that self-ascriptions of beliefs, in general, are for enabling us to report our own beliefs to others, which has in cooperative social environments adaptive advantages because it leads to improved joint decision-making and better group coordination with mutual benefits for both self and others. I maintained that the function of TM-based self-ascriptions and so privileged self-knowledge of beliefs, in particular, is then to enable the most efficient and most reliable verbal communication of one's own beliefs to others. This is what I called the *ICC-TM view*. It is the fourth and final component of the FDM theory.

The ICC-TM view fills an explanatory gap that the IDR-ISA hybrid view leaves open, for it says what the function of privileged self-knowledge of beliefs is. It also lends support to a crucial assumption underlying the FDM theory as a whole. The assumption is that to find out about our own beliefs, we redeploy a principle that we use to work out other people's beliefs, namely (IDA): What subjects will say in response to the question of whether *p* is what they believe about *p*. The FDM theory assumes that a grasp of (IDA) is in place prior to the ability to self-ascribe beliefs. But what supports this assumption?

The support is grounded in the ICC-TM view and its argument to the effect that self-ascriptions of beliefs via TM emerged for cooperative purposes. To see this, note first that competition and hostility is a fundamental aspect of evolution and cooperation is a later development in evolutionary history (Byrne and Whiten 1988, 1997; Sterelny 2003). Since that is so, prior to living in cooperative environments, people presumably lived in competitive or hostile social environments. In such environments, there was a clear need to determine other people's beliefs because knowing their beliefs helps predict their behaviour, and the prediction of other people's behaviour was vital to, e.g., avoid deception or manipulation. It also served one's own Machiavellian purposes. Since that is so, it is plausible to assume that the ability to other-ascribe beliefs developed prior to the ability to self-ascribe them. Once cooperative environments emerged, the ability to other-ascribe beliefs and the impetus to search for ways of working out other people's beliefs were thus already in place. That is, the disposition to use behavioural data to determine other people's beliefs and form principles that tie their behaviour to mental states such as beliefs was in cooperative environments already present before the tendency to form principles that would allow one to work out one's own beliefs, because it was carried over from non-cooperative times. In contrast, the tendency to search for ways of working out one's own beliefs had to arise anew. Since that is so, there is good reason to believe that the acquisition of a grasp of (IDA) happened prior to the acquisition of the ability to self-ascribe beliefs.

On this view, in the course of evolution, the human cognitive system acquired the disposition to recruit the resources of the mindreading faculty for the production of self-ascriptions of attitudes. It is in virtue of this disposition to redeploy and direct upon oneself principles and mechanisms that were originally only used for forming other-ascriptions of attitudes that, ontogenetically, a grasp of (IDA) will need to be in place

first for a subject to become able to produce TM-based self-ascriptions of beliefs.[63]

By supporting the claim that self-ascriptions of beliefs emerged in cooperative environments, the ICC-TM view thus yields support for a central assumption underlying the FDM theory, namely the assumption that to find out about our own beliefs, we recycle (IDA) which we already use to work out other people's beliefs. This makes the ICC-TM view an important fourth component of the overall account of self-knowledge of beliefs developed.

In virtue of combining the (1) DA view, the (2) IDR theory, the (3) IDR-ISA hybrid view, and the (4) ICC-TM view, the FDM theory is able to solve all four problems introduced in Chapter 1 and to offer answers to all three questions about privileged self-knowledge of beliefs with which the thesis began. It answers the *whether*-, and the *how*-questions with (1), (2) and (3). And it answers the *why*-question with (4).

2. An overarching evolutionary argument

The fourth component of the FDM theory, i.e. the ICC-TM view, ties all four parts together in a way that yields a tentative evolutionary argument for the resulting overall view. Aspects of it already figured in the discussion at the end of Chapter 6. But with all four components of the overall account of self-knowledge of beliefs in place, the argument can now be made fully explicit in the following eight points.

(1) In cooperative social environments, there was selection pressure for an efficient and reliable ability to report one's own beliefs. Since this ability requires the formation of self-ascriptions of beliefs, in cooperative social environments, there was selection pressure for an efficient and reliable ability to form self-ascriptions of beliefs (ICC-TM view).

(2) In general, "natural selection must work within pre-existing constraints, modifying pre-existing structures and pressing them into new uses" (Murphy 2003: 173).

(3) Prior to being able to self-ascribe beliefs, we already had direct access to our own judgments (and via them to our own beliefs) in conscious first-order thinking (DA

[63] This is compatible with a simultaneous emergence of the ability to self- and other-ascribe beliefs.

view). This made the formation of self-ascriptions of beliefs for the purpose of, e.g., controlling our own cognition redundant (ICC-TM view). We also already had the resources (e.g., a grasp of (IDA)) and the faculty (i.e. the mindreading system) to work out other people's beliefs (IDR-ISA hybrid view). For competitive social environments, which led to the emergence of them, were prior to cooperative ones, which were required for giving rise to self-ascriptions of beliefs (ICC-TM view).

(4) In cooperative social environments, these other-related resources and faculty only had to be applied to oneself to allow for the formation of TM-based self-ascriptions of beliefs. Designing the ability to use TM via natural selection would thus have been relatively straightforward (IDR-ISA hybrid view, ICC-TM view).

(5) TM-based self-ascriptions could have been formed in the same way as other-ascriptions of beliefs are formed, i.e. via interpretation, or more generally via inferences from evidence of one's holding the beliefs. But because of the way the human cognitive system is built and operates, when subjects applied to themselves the same procedure that they used to determine other (cooperative) people's beliefs (i.e. questioning them whether *p*), this initiated fundamentally different processing (IDR theory). For in virtue of the way in which the human cognitive system is set up, one has, e.g., (1) direct access to one's own judgments in first-order thinking, (2) a standing desire to correctly answer self-sourced questions as to whether *p*, (3) a tendency to predict, upon deciding to *φ*, that one will *φ*, (4) a strong truth bias towards self-sourced responses to self-generated questions as to whether *p*, and (5) a disposition to short-circuit complex information processing whenever possible (IDR-ISA hybrid view).

(6) Due to the coincidence of the preceding points, TM-based self-ascriptions of beliefs turned out to be non-inferential in nature. They turned out to be privileged self-knowledge of beliefs (IDR theory).

(7) This was precisely what would in cooperative social environments have met the selection pressure for an efficient and reliable ability to report one's own beliefs (ICC-TM view).

(8) Since the ability to form TM-based self-ascriptions in the way the FDM theory proposes was arguably also the easiest to evolve from pre-existing structures, there is

reason to believe that TM-based self-ascriptions, formed in the way the FDM theory suggests, were selected and did indeed evolve.

Points (1) to (8) capture a tentative evolutionary argument that suggests that the overall account of self-knowledge of beliefs developed in this thesis, comprising the DA view, the IDR theory, the IDR-ISA hybrid view, and the ICC-TM view, is on the right track.

3. From opacity to transparency

Finally, the FDM theory brings to the fore and corroborates a hitherto unexplored and somewhat surprising picture of the relation between the nature of self-knowledge of beliefs and the nature of other-knowledge of them. More specifically, the account supports the view that both the existence and privileged nature of self-knowledge of beliefs are grounded in and become explicable by appeal to the *prima facie* unrelated fact that other people's beliefs are *opaque*, i.e. only interpretively accessible, to us. To see this, recall that according to the FDM theory, privileged self-knowledge of beliefs is acquired by employing the same method that we often use to work out other people's beliefs. To determine whether other people believe *p*, we tend to ask them whether *p* and then take their response to the probe to capture what they believe. According to the FDM theory, this procedure is applied to oneself and becomes the basis for the acquisition of privileged self-knowledge of beliefs. For it is via this practice that our beliefs become *transparent* in that we can find out whether we believe *p* simply by determining whether *p*. The transparency of our own beliefs thus hinges on our practice of finding out about other people's beliefs concerning *p* by asking them whether *p*. For if this procedure never came to our minds in the first place, then we also couldn't subsequently apply it ourselves.

Suppose now we had direct access to other people's beliefs in that we could find out about their beliefs without observing and interpreting any expression of the beliefs in people's behaviour. If that were so, then to find out whether others believe *p*, we wouldn't need to first ask them whether *p*, await their response, and on the basis of the response ascribe a belief about *p* to them. We wouldn't need to access their belief about *p* in this roundabout way via an interpretation-dependent intermediary, i.e. their assertion about *p*. We could instead employ the envisaged direct access to their minds. As a result, we wouldn't acquire the practice of determining other people's belief

concerning *p* by asking them whether *p* to start with. This procedure only arises, because we *lack* direct access to their beliefs and rely on the expressions of them in behaviour. That is, the fact that other people's beliefs are opaque to us, i.e. only accessible indirectly, interpretively via intermediaries, gives rise to our practice of asking them whether *p* when we want to find out whether they believe *p*. Since we need to have acquired this practice to be able to apply it to ourselves in order to form TM-based self-ascriptions and since we only acquire the procedure at issue if other people's beliefs are opaque to us, it follows that the formation of TM-based self-ascriptions itself requires the opacity of other people's beliefs. That is, without our lack of direct access to other people's beliefs, there would be no privileged self-knowledge acquired via TM in the way the IDR theory proposes either.

There is a second way in which, according to the account of self-knowledge of beliefs introduced in this thesis, our lack of direct access to other people's beliefs is critical for the emergence of self-knowledge of beliefs in general and TM-based privileged self-knowledge in particular. The FDM theory argues that self-knowledge of beliefs emerged because one sometimes holds beliefs that are based on intra-subjective evidence, and these beliefs can't typically be verbally conveyed to others unless one reports them, which requires self-ascriptions of them first. According to the FDM theory, the selection pressure for the report of one's own beliefs to others drove the emergence of self-ascriptions and self-knowledge of beliefs. This selection pressure evidently only arises if other people's beliefs are not directly accessible to us. If they were directly accessible to us, i.e. without any intermediaries standing in our way, then other people wouldn't need to self-ascribe and report their beliefs in order for us to know about them. Hence, according to the FDM theory, our lack of direct access to other people's beliefs is crucial for creating the conditions in which self-knowledge of belief evolves.

Thus, according to the overall account of self-knowledge of beliefs developed in this thesis, the opacity of other people's beliefs to each of us leads to both the emergence of self-knowledge of beliefs and its privileged nature, for it provides each of us with the resources for forming non-inferential TM-based self-ascriptions. This, then, is how, when it comes to knowledge of beliefs, the opacity of other people's minds spawns the transparency of our own minds.

REFERENCES

Achtziger, A., Martiny, S., Oettingen, G., and Gollwitzer, P., 2012. "Metacognitive Processes in the Self-Regulation of Goal Pursuit". In: P. Briñol, P., and DeMarree, K., (eds.), *Social Metacognition*, New York, NY: Psychology Press, 121–140.

Alais, D., and Burr, D., 2004. "The ventriloquist effect results from near optimal bimodal integration". *Current Biology*, 14, 3, 257–262.

Anscombe, G. E. M., 1976. *Intention.* Ithaca, New York: Cornell University Press.

Armstrong, D., 1968. *A Materialist Theory of the Mind*. New York: Humanities.

Armstrong, D., 1999. *The Mind-Body Problem*. Boulder: Westview Press.

Baars, B., 1988. *A Cognitive Theory of Consciousness*. New York: Cambridge University Press.

Bach, K., 1984. "Default Reasoning: Jumping to Conclusions and Knowing When to Think Twice". *Pacific Philosophical Quarterly*, 65, 37–58

Baddeley, A., 1986. *Working Memory*. Oxford: Oxford University Press.

Baddeley, A., 2006. *Working Memory, Thought, and Action*. Oxford: Oxford University Press.

Baddeley, A., 2010. "Working memory". *Current Biology*, 20, R136–R140.

Bahrami, B., Olsen, K., Latham, P., Roepstorff, A., Rees, G., and Frith, C., 2010. "Optimally interacting minds". *Science*, 329, 1081–85.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., and Frith, C., 2011. "Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making". *Journal of Experimental Psychology: Human Perception and Performance*, 38, 3–8.

Barnett, D., 2015. “Inferential Justification and the Transparency of Belief”. *Nous*, online first.

Baron-Cohen, S., 1995. *Mindblindness: An essay on autism and theory of mind.* Cambridge, MA: MIT Press.

Baron-Cohen, S., 1989. “The autistic child’s theory of mind: A case of specific developmental delay”. *Journal of Child Psychology and Psychiatry*, 30, 285–297.

Bayne, T., 2008. “The Unity of Consciousness and the Split-Brain Syndrome”. *The Journal of Philosophy*, 105, 6, 277–300.

Bayne, T., and Montague, M., (eds.), 2011. *Cognitive Phenomenology*. Oxford: Oxford University Press.

Bayne, T., and Hattiangadi, A., 2013. “Belief and its bedfellows”. In: Nottelmann, N., (ed.), *New Essays on Belief: Constitution, Content and Structure*. London: Palgrave, 124–145

Bilgrami, A., 1998. “Self-Knowledge and Resentment”. In: Wright, C., Smith, B., and Macdonald, C., (eds.), *Knowing Our Own Minds*. Oxford: Oxford University Press, 207–241.

Bilgrami, A., 2006. *Self-Knowledge and Resentment*. Cambridge, MA: Harvard University Press.

Bilgrami, A., 2012. “The unique status of self-knowledge”. In: Coliva, A., (ed.), *The Self and Self-Knowledge*. Oxford: Oxford University Press, 263–270.

Blakemore, S., Wolpert, D., and Frith, C., 2002. “Abnormalities in the awareness of action”. *Trends in Cognitive Sciences*, 6, 6, 237–242.

Block, N., 1983. “Mental pictures and cognitive science”. *The Philosophical Review*, 92, 499–541.

Block, N., 1995. “On a confusion about a function of consciousness”. *Behavioural and Brain Sciences*, 18, 2, 227–47.

Block, N., 2002. “Some concepts of consciousness”. In: Chalmers, D., (ed.), *Philosophy of Mind*. Oxford: Oxford University Press.

Block, N., 2007. “Consciousness, accessibility, and the mesh between psychology and neuroscience”. *Behavioural and Brain Sciences*, 30, 481–548.

Boghossian, P., 2008. *Content and Justification: Philosophical Papers*. Oxford: Oxford University Press.

Boghossian, P., 2014. “What is inference?” *Philosophical Studies*, 169, 1, 1–18.

Bond, C., and DePaulo, B., 2006. “Accuracy of deception judgments”. *Personality and Social Psychology Review*, 10, 3, 214–34.

Bonjour, L., and Sosa, E., (eds.), 2003, *Epistemic Justification: Internalism vs. Externalism, Foundations vs. Virtues.* Oxford: Blackwell.

Boyle, M., 2011. “Transparent Self-Knowledge”. *Proceedings of the Aristotelian Society*, 85, 1, 223–41.

Brasil-Neto, J., et al., 1992. “Focal transcranial magnetic stimulation and response bias in a forced choice task”. *Journal of Neurology, Neurosurgery, and Psychiatry*, 55, 964–966.

Bratman, M., 1987. *Intention, Plans, and Practical Reason*. Stanford: CSLI Publications.

Bratman, M., 2009. “Intention, Belief and Instrumental Rationality”. In: Sobel, S., and Wall, S., (eds.), *Reasons for Action*. Cambridge: Cambridge University Press, 13–36.

Briñol, P., and Petty, R., 2003. “Overt head movements and persuasion: a self-validation analysis”. *Journal of Personality and Social Psychology*, 84, 1123–1139.

Brown, R., and McNeill, D., 1966. "The 'tip of the tongue' phenomenon". *Journal of Verbal Learning and Verbal Behavior*, 5, 325–337.

Brüne, M., Lissek, S., Fuchs, N., Witthaus, H., Peters, S., Nicolas, V., Juckel, G., and Tegenthoff, M., 2008. "An fMRI study of theory of mind in schizophrenic patients with 'passivity' symptoms". *Neuropsychologia*, 46, 1992–2001.

Buller, D., Strzyzewski, K., and Comstock, J., 1991. "Interpersonal deception: I. Deceivers' reactions to receivers' suspicions and probing". *Communication Monographs*, 58, 1–24.

Burge, T. 1996. "Our Entitlement to Self-Knowledge". *Proceedings of the Aristotelian Society*, 96, 91–116.

Burgoon, J., Buller, D., Ebesu, A., and Rockwell, P., 1994. "Interpersonal deception: V. Accuracy in deception detection". *Communication Monographs*, 61, 303–325.

Burgoon, J., Blair, P., Storm, R., 2008. "Cognitive Biases and Nonverbal Cue Availability in Detecting Deception". *Human Communication Research*, 34, 4, 572–599.

Byrne, A., 2005. "Introspection". *Philosophical Topics*, 33, 79–104.

Byrne, A., 2011. "Transparency, Belief, Intention". *Aristotelian Society Supplementary*, 85, 1, 201–221.

Byrne, A., 2012. "Review: The Opacity of Mind: An Integrative Theory of Self-Knowledge". *Notre Dame Philosophical Reviews*.

Byrne, R., and Whiten, A., (eds.), 1988. *Machiavellian Intelligence*. Oxford: Oxford University Press.

Byrne, R., and Whiten, A., (eds.), 1997. *Machiavellian Intelligence II.* Cambridge: Cambridge University Press.

Carruthers, P., 2000. *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge: Cambridge University Press.

Carruthers, P., 2009a. "How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition". *Behavioural and Brain Sciences*, 32, 2, 121-138.

Carruthers, P., 2009b. "Action-awareness and the active mind". *Philosophical Papers*, 38, 133–56.

Carruthers, P., 2010. "Introspection: Divided and partly eliminated". *Philosophy and Phenomenological Research*, 80, 76–111.

Carruthers, P., 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge.* Oxford: Oxford University Press.

Carruther, P., 2014. "On central cognition". *Philosophical Studies*, 170, 1, 143–162

Carruthers, P., 2015. *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought.* Oxford: Oxford University Press.

Cassam, Q., 2010. "Judging, Believing and Thinking". *Philosophical Issues*, 20, 80–95.

Cassam, Q., 2011. "Knowing What You Believe". *Proceedings of the Aristotelian Society*, 111, 1, 1–23.

Cassam, Q., 2014. *Self-Knowledge for Humans*. Oxford: Oxford University Press.

Cassam, Q., 2015. "What asymmetry? Knowledge of self, knowledge of others, and the inferentialist challenge". *Synthese*, 1–19.

Chun, M., and Jiang, Y., 1998. "Contextual cueing: Implicit learning and memory of visual context guides spatial attention". *Cognitive Psychology*, 36, 28–71.

Cooper, J., 2007. *Cognitive Dissonance: 50 Years of a Classic Theory*. London: Sage.

Corcoran, R., Frith, C., and Mercer, G., 1995. “Schizophrenia, symptomatology, and social inference: Investigating ‘theory of mind’ in people with schizophrenia”. *Schizophrenia Research*, 17, 5–13.

Corcoran, R., and Frith, C., 1996. “Conversational conduct and the symptoms of schizophrenia”. *Cognitive Neuropsychiatry*, 1, 305–318.

Corcoran, R., Cahill, C., and Frith, C., 1997. “The appreciation of visual jokes in people with schizophrenia: A study of ‘mentalizing’ ability”. *Schizophrenia Research*, 24, 319–327.

Cowan, N., 2008. “What are the differences between long-term, short-term, and working memory?” In: Sossin, W., Jacaille, J., Castellucci, V., and Belleville, S., (eds.), *The essence of memory*. Amsterdam: Elsevier/Academic Press, 323–338.

Craig, A., 2015. *How Do You Feel? An Interoceptive Moment with Your Neurobiological Self.* Princeton: Princeton University Press.

Crane, T., 2013. “Unconscious belief and conscious thought”. In: Kriegel, U., (ed.), *Phenomenal Intentionality*. Oxford: Oxford University Press, 156–173.

Cummins, R., 1996. *Representations, Targets, and Attitudes*. Cambridge, MA: MIT Press.

D’Agata, M., and Jacobson, J., 2014. “Cognitive heuristics”. In: Levine, T., (ed.), *Encyclopedia of Deception*. Thousand Oaks, CA: SAGE Publications, 157–159.

Davidson, D., 1994. “Knowing one’s own mind”. In: Cassam, Q., (ed.), *Self-Knowledge*. Oxford: Oxford University Press, 43–64.

Dehaene, S., and Naccache, L., 2001. “Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework”. *Cognition*, 79, 1–37.

De Neys, W., Rossi, S., and Houdé, O., 2013. “Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools”. *Psychonomic Bulletin and Review*, 20, 269–273.

Dennett, D., 1978. “Beliefs About Beliefs”. *Behavioral and Brain Sciences* 1 (4): 568.

Dennett, D., 1991. *The Intentional Stance*. Cambridge, MA: MIT.

Dennett, D., 1992. “The self as a center of narrative gravity”. In: Kessel, F., Cole, P., and Johnson, D., (eds.), *Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Erlbaum.

Dennett, D., 2001. “Are We Explaining Consciousness Yet?” *Cognition*, 79, 221–237.

DePaulo, B., Lindsay, J., Malone, B., Muhlenbruck, L., Charlton, K., and Cooper, H., 2003. “Cues to deception”. *Psychological Bulletin*, 129, 1, 74–118.

Dewey, M., 1991. “Living with Asperger’s syndrome”. In: Frith, U., (ed.), *Autism and Asperger Syndrome*. Cambridge: Cambridge University Press.

Dretske, F., 1988. *Explaining Behavior: Reasons in a World of Causes.* Cambridge, MA: MIT Press.

Dretske, F., 2003. “Externalism and Self Knowledge”. In: Nuccetelli, S., (ed.) *Semanatic Externalism, Skepticism and Self-Knowledge*. Cambridge, MA: MIT Press, 131–143.

Ernst, M., and Banks, M., 2002. “Humans integrate visual and haptic information in a statistically optimal fashion”. *Nature*, 415, 6870, 429–433.

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

Evans, J., 2005. “Deductive reasoning”. In: Holyoak, K., and Morrison, R., (eds.), *Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press, 169–185.

Evans, J., 2008. “Dual-processing accounts of reasoning, judgment, and social cognition”. *Annual Review of Psychology*, 59, 255–278.

Evans, J., and Stanovich, K., 2013. “Dual-process theories of higher cognition advancing the debate”. *Perspectives on Psychological Science*, 8, 223–241.

Farrant, A., Boucher, J., and Blades, M., 1999. “Metamemory in children with autism”. *Child Development*, 70, 107–131.

Feldman, R., and Conee, E., 2001. “Internalism defended”. *American Philosophical Quarterly*, 381, 1–18.

Fernández, J., 2013. *Transparent Minds: A Study of Self-Knowledge*. Oxford: Oxford University Press.

Finkelstein, D., 2012. “From transparency to expressivism”. In: Abel, G., and Conant, J., (eds.), *Rethinking Epistemology*. Berlin: De Gruyter, 101–118.

Fiske, S., and Taylor, S., 2013. *Social cognition*. London: Sage.

Fodor, J., 1975. *The Language of Thought*. Cambridge, Massachusetts: Harvard University Press.

Fodor, J., 1990. *A Theory of Content and Other Essays*. Cambridge, Massachusetts: MIT Press.

Frankish, K., 2009. “Systems and levels: Dual-system theories and the personal-sub-personal distinction”. In: Frankish, K., and Evans, J., (eds.), *In Two Minds: Dual Processes and Beyond.* Oxford: Oxford University Press.

Frankish, K., 2012. “Dual systems and dual attitudes.” *Mind and Society*, 11, 1, 41–51.

Frith, C., and Done, D., 1989. “Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action”. *Psychological Medicine*, 19, 359–363.

Frith, C., and Corcoran, R., 1996. “Exploring theory of mind in people with schizophrenia”. *Psychological Medicine*, 26, 521–530.

Frith, U., and Happé, F., 1999. “Theory of mind and self-consciousness: what is it like to be autistic?” *Mind and Language*, 14, 1–22.

Frith, C., Blakemore, S., and Wolpert, D., 2000. “Abnormalities in the awareness and control of action”. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 355,1771-88.

Frith, C., 2010. “What is consciousness for?” *Pragmatics and Cognition*, 18, 3, 497–551.

Frith, C., 2012. “The role of metacognition in human social interactions”. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 2213–2223.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., and Tylén, K., 2012. “Coming to terms quantifying the benefits of linguistic coordination”. *Psychological Science*, 23, 931–939.

Ganea, P., Lillard, A., and Turkheimer, E., 2004. “Preschooler’s understanding of the role of mental states and action in pretense”. *Journal of Cognitive and Development*, 5, 213–238.

Gazzaniga, M., 1995. “Consciousness and the cerebral hemispheres”. In: Gazzaniga, M., (ed.), *The Cognitive Neurosciences*. Cambridge, Massachusetts: MIT Press.

Gazzaniga, M., 2000. “Cerebral specialization and inter-hemispheric communication”. *Brain*, 123, 1293–1326.

Gennaro, R., 2012. *The Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts.* Cambridge, Massachusetts: MIT Press.

Gertler, B., 2011. *Self-Knowledge*. New York: Routledge Press.

Gertler, B., 2015. “Self-Knowledge”. *The Stanford Encyclopedia of Philosophy*, Zalta, E., (ed.), URL: <http://plato.stanford.edu/archives/sum2015/entries/self-knowledge/>.

Gettier, E., 1963. “Is justified true belief knowledge?” *Analysis*, 23, 6, 121–123.

Goldman, A., 1993. “The psychology of folk psychology”. *Behavioral and Brain Sciences* 16, 15–28.

Goldman, A., 2006. *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Goldman, A., 2009. “Internalism, externalism, and the architecture of justification”. *Journal of Philosophy*, 106, 6, 309–338.

Goldman, A., 2012. “Theory of mind”. In: Margolis, E., Samuels, R., and Stich, S., (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science.* Oxford: Oxford University Press.

Gollwitzer, P., 1990. “Action phases and mind-sets”. In: Higgins, E., and Sorrentino, R., (eds.), *The Handbook of Motivation and Cognition: Foundations of Social Behaviour*. New York: Guilford Press, 53–92.

Gollwitzer, P., and Bayer, U., 1999. “Deliberative versus implemental mindsets in the control of action”. In: Chaiken, S., and Trope, Y., (eds.), *Dual-Process Theories in Social Psychology.* New York: Guilford, 403–422.

Gopnik, A., 1993. “The illusion of first-person knowledge of intentionality”. *Behavioural and Brain Sciences*, 16, 1–14.

Gopnik, A. and Slaughter, V., 1991. “Young children’s understanding of changes in their mental states”. *Child Development*, 62, 98–110.

Gordon, R., 2007. “Ascent routines for propositional attitudes”. *Synthese*, 159, 2, 151–165.

Gould, S., and Lewontin, R., 1979. “The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme”. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 205, 581–598.

Grainger, C., Williams, D., and Lind, S., 2014. “Metacognition in high-functioning adults with autism spectrum disorder: Diminished feelings of knowing and mindreading ability”. *Journal of Abnormal Psychology*, 123, 3, 650–659.

Grandin, T., 1984. “My experiences as an autistic child and review of selected literature”. *Journal of Orthomolecular Psychiatry*, 13, 144–75.

Hall, L., Johansson, P., Tärning, B., Sikström, S., and Deutgen, T., 2010. “Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea”. *Cognition*, 117, 54–61.

Hampshire, S., 1965. *Freedom of the Individual*. London: Chatto and Windus.

Happé, F., 2003. “Theory of mind and the self”. *Annals of the New York Academy of Sciences*, 1001, 134–144.

Hart, J., 1965. “Memory and the feeling-of-knowing experience”. *Journal of Educational Psychology*, 56, 208–216.

Harman, G., 1973. *Thought*. Princeton: Princeton University Press.

Harman, G., 1976. “Practical reasoning”. *Review of Metaphysics*, 29, 431–463.

Harman, G., 1978. “Studying the Chimpanzee’s Theory of Mind”. *Behavioral and Brain Sciences* 1, 4, 576.

Hartwig, M., Granhag, P., Stromwall, L. and Vrij, A., 2004. “Police officers’ lie detection accuracy: Interrogating freely versus observing video”. *Police Quarterly*, 7, 429–456.

Hlobil, U., 2014. “Against Boghossian, Wright and Broome on inference”. *Philosophical Studies*, 167, 2, 419–429.

Holton, R., 2009. *Willing, Wanting, Waiting*. Oxford: Clarendon Press.

Hrdy, S., 2009. *Mothers and Others.* Harvard University Press.

Hurlburt, R., Happé, F., and Frith, U., 1994. "Sampling the form of inner experience in 3 adults with Asperger syndrome". *Psychological Medicine*, 24, 2, 385–395.

Hurlburt, R., 2009. "Unsymbolized Thinking, Sensory Awareness, and Mindreading". *Behavioral and Brain Sciences*, 32, 2, 149–150.

Ichikawa, J., and Steup, M., 2014. "The analysis of knowledge". *The Stanford Encyclopedia of Philosophy*, Zalta, E., (ed.), URL = <http://plato.stanford.edu/archives/spr2014/entries/knowledge-analysis/>.

Johansson, P., Hall, L., Sikstrom, S., Olsson, A., 2005. "Failure to detect mismatches between intention and outcome in a simple decision task". *Science*, 310, 116–119.

Kahneman, D., and Frederick, S., 2002. "Representativeness revisited: Attribute substitution in intuitive judgment". In: Gilovich, T., Griffin, D., and Kahneman, D., (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press, 49–81.

Kahneman, D., 2011. *Thinking, Fast and Slow.* New York: Farrar, Strauss, Giroux.

Kazak, S., Collis, G., and Lewis, V., 1997. "Can young people with autism refer to knowledge states? Evidence from their understanding of 'know' and 'guess'". *Journal of Child Psychology and Psychiatry*, 38, 1001–1009.

Keefe, R., Poe, M., McEvoy, J., and Vaughan, A., 2003. "Source monitoring improvement in patients with schizophrenia receiving antipsychotic medications". *Psychopharmacology*, 169, 383–389.

Koriat, A., 2007. “Metacognition and consciousness”. In: Zelazo, P., Moscovitch, M., and Thompson, E., *Cambridge handbook of consciousness.* Cambridge, UK: Cambridge University Press, 289–325.

Kornblith, H., (ed.), 2001. *Epistemology: Internalism and Externalism.* Oxford: Blackwell Publishers.

Kosslyn, S., Thompson, W., and Ganis, G., 2006. *The case for mental imagery.* New York: Oxford University Press.

Kovács, Á., Téglás, E., and Endress, A., 2010. “The social sense: Susceptibility to others’ beliefs in human infants and adults”. *Science*, 330, 1830–1834.

Kriegel, U., 2013. “Brentano’s Most Striking Thesis: No Representation without Self-Representation”. In: Fisette, D., and Fréchette, G., (eds.), *Themes from Brentano*, Amsterdam: Rodopi, 22–40.

LaBerge, S., 1990. “Lucid dreaming: Psychophysiological studies of consciousness during REM sleep”. In: Bootzen, R., Kihlstrom, J., and Schacter, D., (eds.), *Sleep and Cognition.* Washington, D.C.: American Psychological Association, 109–126.

Lawlor, K., 2009. “Knowing what one wants”. *Philosophy and Phenomenological Research*, 79, 47–75.

Leslie A., and Thaiss L., 1992. “Domain specificity in conceptual development: neuropsychological evidence from autism”. *Cognition*, 43, 3, 225–51.

Levine, T., Park, H., and McCornack, S., 1999. “Accuracy in detecting truths and lies: Documenting the ‘veracity effect.’” *Communication Monographs*, 66, 125–144.

Levine, T., Anders, L., Banas, J., Baum, K., Endo, K., Hu, A., and Wong, N., 2000. “Norms, expectations, and deception: A norm violation model of veracity judgments”. *Communication Monographs*, 67, 123–137.

Levine, T., and Kim, R., 2010. “Some considerations for a new theory of deceptive communication”. In: Knapp, M., and McGlone, M., (eds.). *The Interplay of Truth and Deception*. Routledge, 16–34.

Levine, T., 2014. “Truth-default theory (TDT): A theory of human deception and deception detection”. *Journal of Language and Social Psychology*, 33, 378–392.

Logan, G., and Crump, M., 2010. “Cognitive illusions of authorship reveal hierarchical error detection in skilled typists”. *Science*, 330, 683–686.

Lombardo, M., and Baron-Cohen, S., 2010. “Unraveling the paradox of the autistic self”. *Wiley Interscience Reviews* (WIRES), 1, 393–403.

Low, J., and Perner, J., 2012. “Implicit and explicit theory of mind: State of the art”. *British Journal of Developmental Psychology*, 30(1), 1–13.

Lurz, R., 2006. “Conscious beliefs and desires: Same-order approach”. In: Kriegel, U., and Williford, K., (eds.), *Self-Representational Approaches to Consciousness*. Cambridge, MA: MIT Press, 321–351.

Lycan, W., 1996. *Consciousness and Experience.* Cambridge, MA: MIT Press.

Lycan, W., 2008. “Phenomenal intentionalities”. *American Philosophical Quarterly*, 45, 3, 233–52.

Lyons T., and Fitzgerald M., 2013. *Recent Advances in Autism Spectrum Disorders*. Rijeka: In Tech.

Mandelbaum, E., 2014. “Thinking is believing”. *Inquiry*, 57, 1, 55–96.

Martin, M., 1998. “An eye directed outward”. In: Wright, C., Smith, B., and Macdonald, C., (eds.), *Knowing Our Own Minds*. Oxford: Clarendon Press, 99–121.

McCornack, S., and Parks, M., 1986. “Deception detection and relationship development: The other side of trust”. In: McLaughlin, M., (ed.), *Communication Yearbook*. Beverly Hills, CA: Sage, 377–389.

McCornack, S., and Levine, T., 1990. “When lovers become leery: The relationship between suspicion and accuracy in detecting deception”. *Communication Monographs*, 57, 219–230.

McGeer, V., 1996. “Is ‘Self-knowledge’ an empirical problem? Renegotiating the space of philosophical explanation”. *Journal of Philosophy*, 93, 483–515.

McGeer, V., 2008. “The moral development of first-person authority”. *European Journal of Philosophy*, 16, 1, 81–108.

Mele, A., 2009. *Effective intentions*. Oxford: Oxford University Press.

Mercier, H., 2013. “Our pigheaded core: How we became smarter to be influenced by other people”. In: Calcott, B., Joyce, R. and Sterelny, K. (eds.) *Evolution, Cooperation, and Complexity*. Cambridge, Massachusetts: MIT Press, 373–399.

Millar, M., and Millar, K., 1995. “Detection of deception in familiar and unfamiliar persons: The effects of information restriction”. *Journal of Nonverbal Behaviour*, 19, 2, 69–84.

Millar, M., and Millar, K., 1997. “The effect of cognitive capacity and suspicion on truth bias”. *Communication Research*, 24, 5, 556–570.

Millikan, R., 1993. *White Queen Psychology and Other Essays for Alice*. Cambridge, Massachusetts: MIT Press.

Mlakar, J., Jensterle, J., and Frith, C., 1994. “Central monitoring deficiency and schizophrenic symptoms”. *Psychological Medicine*, 24, 557–564.

Moore, G., 1942. “A reply to my critics”. In: Schlipp, P., (ed.), *The Philosophy of G. E. Moore*. Evanston: Tudor, 543–667.

Moran, R., 2001. *Authority and Estrangement*. Princeton: Princeton University Press.

Moran, R., 2003. "Responses to O'Brien and Shoemaker". *European Journal of Philosophy*, 11, 402–419.

Moran, R., 2004. "Review: Replies to Heal, Reginster, Wilson, and Lear". *Philosophy and Phenomenological Research*, 69, 2, 455–472.

Moran, R., 2012. "Self-Knowledge, 'Transparency' and the Forms of Activity". In: Smithies, D., and Stoljar, D., (eds.), *Introspection and Consciousness*. Oxford: Oxford University Press, 211–239.

Murphy, D., 1998. "Theory of mind in a sample of men with schizophrenia detained in a special hospital: its relationship to symptom profiles and neuropsychological tests". *Criminal Behaviour and Mental Health*, 8, 13–26.

Musholt, K., 2012. "Self-consciousness and intersubjectivity". *Grazer Philosophische Studien*, 84, 63–89.

Naccache, L., Dehaene S., Cohen, L., Habert, M., Guichart-Gomez, E., Galanaud, D., and Willer, J., 2005. "Effortless control: executive attention and conscious feeling of mental effort are dissociable". *Neuropsychologia*, 43, 1318–1328.

Nagel, T., 1974. "What is it like to be a bat?" *The Philosophical Review* 83, 4, 435–50.

Nichols, S., and Stich, S., 2003. *Mindreading*. New York: Oxford University Press.

Oettingen, G., and Gollwitzer, P., 2010. "Strategies of setting and implementing goals: Mental contrasting and implementation intentions". In: Maddux, J., and Tangney, J., (eds.), *Social psychological foundations of clinical psychology*. New York: Guilford, 114–135.

Onishi, K., and Baillargeon, R. 2005. "Do 15-month-old infants understand false beliefs?" *Science*, 308(8), 255–258.

Pacherie, E., 2007. “The anarchic hand syndrome and utilization behaviour: A window onto agentive self-Awareness”. *Functional Neurology*, 22, 4, 211–217.

Paul, S., 2014. “The transparency of mind”. *Philosophy Compass*, 9, 295–303.

Paulus, M., Proust, J., and Sodian, B., 2013. “Examining implicit metacognition in 3.5-year-old children: An eye-tracking and pupillometric study”. *Frontiers in Psychology*, 4, 145, 1–7.

Peacocke, C., 1996. “Our entitlement to self-Knowledge: Entitlement, self-knowledge, and conceptual redeployment”. *Proceedings of the Aristotelian Society*, 96, 117–58.

Peacocke, C., 2003. “Conscious attitudes, attention and self-knowledge”. In: Gertler, B., (ed.), *Priviledged Access*. Aldershot: Ashgate, 83–110.

Peacocke, C., 2008. *Truly Understood*. Oxford: Oxford University Press.

Perner, J., Frith, U., Leslie, A., and Leekam, S., 1989. “Exploration of the autistic child’s theory of mind: Knowledge, belief, and communication”. *Child Development*, 60, 689–700.

Perner, J., 1991. *Understanding the representational mind.* Cambridge, MA: MIT Press.

Perner, J., and Roessler, J., 2010. “Teleology and causal reasoning in children’s theory of mind”. In: Aguilar, J., and Buckareff, A., (eds.) *Causing Human Action: New Perspectives on the Causal Theory of Action*. Cambridge, MA: MIT Press, 199–228.

Peters, U., 2014. “Interpretive sensory-access theory and conscious intentions”. *Philosophical Psychology*, 27, 4, 583–595.

Phillips, W., Baron-Cohen, S., and Rutter, M., 1998. “Understanding intention in normal development and in autism”. *British Journal of Developmental Psychology*, 16, 337–348.

Pickup, G., and Frith, C., 2001. “Theory of mind impairments in schizophrenia: Symptomatology, severity and specificity”. *Psychological Medicine*, 31, 207–220.

Pollock, J., 1982. *Language and Thought*. Princeton: Princeton University Press.

Pratt, C., and Bryant, P., 1990. “Young children understand that looking leads to knowing (so long as they are looking into a single barrel)”. *Child Development*, 61, 973–982.

Prinz, J., 2011. “Is Attention Necessary and Sufficient for Consciousness?” In: Christopher Mole, C., Smithies, D., and Wu, W. (eds.), *Attention: Philosophical and Psychological Essays*. Oxford: Oxford University Press, 174–204.

Prinz, J., 2012. *The Conscious Brain: How Attention Engenders Experience*. Oxford: Oxford University Press.

Proust, J., 2013. *The Philosophy of Metacognition: Mental Agency and Self-Awareness.* Oxford: Oxford University Press.

Pylyshyn, Z., 2003. “Return of the mental image: are there really pictures in the brain?” *Trends in Cognitive Sciences*, 7, 3, 113-118.

Quine, W., 1953. *From a Logical Point of View*. Cambridge, MA: Harvard University Press

Rakoczy, H., 2010. “Executive function and the development of belief-desire psychology”. *Developmental Science*, 13, 4, 648–661.

Reed, B., 2010. “Self-knowledge and rationality”. *Philosophy and Phenomenological Research*, 80, 164–181.

Revonsuo, A., 2006. *Inner Presence. Consciousness as a biological phenomenon.* Cambridge MA: MIT Press.

Richerson, P., and Boyd, R., 2005. *Not By Genes Alone: How Culture Transformed Human Evolution.* University of Chicago Press.

Richter, T., Schroeder, S., and Wohrmann, B., 2009. "You don't have to believe everything you read: background knowledge permits fast and efficient validation of information". *Journal of Personality and Social Psychology*, 96, 538–58.

Robbins, P., 2004. "Knowing Me, Knowing You: Theory of Mind and the Machinery of Introspection". *Journal of Consciousness Studies*. 11, 7-8, 129-143.

Roessler, J., 2013. "The silence of self-knowledge". *Philosophical Explorations*, 16, 1, 1–17.

Roessler, R., 2015. "Self-knowledge and communication". *Philosophical Explorations* 18, 2, 153–168.

Rolls, E.T., 2014. *Emotion and Decision-Making Explained.* Oxford: Oxford University Press.

Rosen, C., Schwebel, D., and Singer, J., 1997. "Preschoolers' attributions of mental states in pretense". *Child Development*, 68, 1133–1142.

Rosenthal, D., 2002. "Explaining consciousness". In: Chalmers, D., (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press, 109–131.

Rosenthal, D., 2005. *Consciousness and Mind*. Oxford: Oxford University Press.

Rosenthal, D., 2008. "Consciousness and its function". *Neuropsychologia*, 46, 3, 829–840.

Ross, J., 2009. "How to be a cognitivist about practical reason". *Oxford Studies in Metaethics*, 4, 243–281.

Ryle, G., 1949. *The Concept of Mind*. London: Hutchinson.

Schenck, C., and Mahowald, M., 2002. “REM sleep behaviour disorder: Clinical, developmental, and neuroscience perspectives 16 years after its formal identification in sleep”. *Sleep*, 25, 2, 120–138.

Schneider, D., Slaughter, V., and Dux, P., 2015. “What do we know about implicit false-belief tracking?” *Psychonomic Bulletin and Review*, 1–12.

Scott, F., and Baron-Cohen, S., 1996. “Logical, analogical, and psychological reasoning in autism: a test of the Cosmides theory”. *Development and Psychopathology*, 8, 235-246.

Scott, F., Baron-Cohen, S., and Leslie, A., 1999. “‘If pigs could fly’: A test of counterfactual reasoning and pretence in children with autism”. *British Journal of Developmental Psychology*, 17, 349–362.

Searle, J., 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Setiya, K., 2007. “Cognitivism about instrumental reason”. *Ethics*, 117, 647–673.

Setiya, K., 2012. “Transparency and inference”. *Proceedings of the Aristotelian Society*, 112, 263–268.

Shallice, T., 1988. *From Neuropsychology to Mental Structure.* Cambridge: Cambridge University Press.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., and Frith, C., 2014. “Supra-personal cognitive control and metacognition”. *Trends in Cognitive Sciences*, 18, 4, 186–193.

Shoemaker, S., 1996. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

Shoemaker, S., 2012. “Self-intimation and second-order belief”. In: Smithies, D., and Stoljar, D., (eds.), *Introspection and Consciousness*. Oxford: Oxford University Press, 239–259.

Silins, N., 2012. “Judgment as a guide to belief”. In: Smithies, D., and Stoljar, D., (eds.), *Introspection and Consciousness*. Oxford: Oxford University Press, 295–328.

Smithies, D., 2012. “A simple theory of introspection”. In: Smithies, D., and Stoljar, D., (eds.), *Introspection and Consciousness*. Oxford: Oxford University Press, 259–294.

Sobel, D., 2007. “Children knowledge of the relation between intentional action and pretending”. *Cognitive Development*, 22, 130–141.

Sodian, B., Schuwerk, T., and Kristen, S., 2015. “Implicit and Spontaneous Theory of Mind Reasoning in Autism Spectrum Disorders”. In: Fitzgerald, M., (ed.), *Autism Spectrum Disorder – Recent Advances*. InTech, 114–135

Soteriou, M., 2005. “Mental action and the epistemology of mind”. *Noûs*, 39, 1, 83–105.

Soteriou, M., 2013. *The Mind’s Construction: The Ontology of Mind and Mental Action.* Oxford: Oxford University Press.

Spehn, M., and Reder, L., 2000. “The unconscious feeling of knowing: a commentary on Koriat’s paper”. *Consciousness and Cognition*, 9, 187–192.

Sperber, D., 2001. “An evolutionary perspective on testimony and argumentation”. *Philosophical Topics*, 29, 401–13.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al., 2010. “Epistemic vigilance”. *Mind and Language*, 25, 4, 359–393.

Stephens, G., and Graham, G., 2000. *When self-consciousness breaks: Alien voices and inserted thoughts.* Cambridge, Massachusetts: MIT Press.

Sterelny, K., 2003. *Thought in a Hostile World: the Evolution of Human Cognition.* Oxford: Blackwell Publishing.

Stoljar, D., and Smithies, D., 2012. "Introspection and consciousness: An overview". In: Smithies, D., and Stoljar, D., (eds.), *Introspection and Consciousness*. Oxford: Oxford University Press, 3–25.

Surian, L., Caldi, S., and Sperber, D. 2007. "Attribution of beliefs by 13-month-old infants". *Psychological Science*, 18, 580–586.

Synofzik, M., Vosgerau, G., and Voss, M., 2013. "The experience of agency: an interplay between prediction and postdiction". *Frontiers in Psychology*, 4, 127, 1–8.

Torralva, T., Gleichgerrcht, E., Roca, M., AIbanez, A., Marenco, V., Rattazzi, A., and Manes, F., 2013. "Impaired theory of mind but intact decision-making in Asperger syndrome: Implications for the relationship between these cognitive domains". *Psychiatry Research*, 205, 282-284.

Tsakiris, M., and Frith, C., 2009. "The Self: neurocognitive approaches". In: Bayne, T., Cleeremans, A., and Wilken, P., (eds.), *The Oxford Companion to Consciousness*. Oxford: Oxford University Press, 585–588.

Tye, M., 1995. *Ten Problems of Consciousness*. Cambridge, Massachusetts: MIT Press.

Tye, M., 2000. *Consciousness, Color, and Content.* Cambridge, MA: MIT Press.

Valaris, M., 2011. "Transparency as inference: Reply to Alex Byrne". *Proceedings of the Aristotelian Society* 111, 2, 319–324.

Valaris, M., 2014a. "Self-Knowledge and the Phenomenological Transparency of Belief". *Philosophers' Imprint* 14, 8, 1–17.

Valaris, M., 2014b. "Reasoning and regress". *Mind*, 123, 489, 101–127.

Velleman, D., 1989. *Practical Reflection*. Princeton: Princeton University Press.

Velleman, D., 2007. “What good is a will?” In: Leist, A., (ed.), *Action in Context*. Berlin: Walter de Gruyter, 193–215.

Vrij, A., 2008. *Detecting Lies and Deceit: Pitfalls and Opportunities.* West-Sussex: Wiley.

Wallace, R., 2001. “Normativity, commitment, and instrumental reason”. *Philosophers’ Imprint*, 1, 4, 1–26.

Wegner, D., and Wheatley T., 1999. “Apparent mental causation: Sources of the experience of will”. *American Psychologist*, 54, 480–491.

Wellman, H., Cross, D., and Watson, J., 2001. “Meta-analysis of theory-of-mind-development: The truth about false belief”. *Child Development*, 72, 665–684.

Williams, D., and Happé, F., 2009. “What did I say? versus What did I think? Attributing false beliefs to self amongst children with and without autism”. *Journal of Autism and Developmental Disorders*, 39, 6, 865–873.

Williams, D., 2010. “Theory of own mind in autism”. *Autism*, 14, 5, 474–494.

Williams, D., and Happé, F., 2010. “Representing intentions in self and other: studies of autism and typical development”. *Developmental Science*, 13, 2, 307–319.

Wilson, T., 2002. *Strangers to Ourselves*. Cambridge, MA: Harvard University Press.

Wilson, R., 2003. “Intentionality and phenomenology”. *Pacific Philosophical Quarterly*, 84, 4, 413–431.

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, 103–128.

Wimmer, H., Hogrefe, G., and Perner, J., 1988. “Children’s understanding of informational access as a source of knowledge”. *Child Development*, 59, 386–396.

Wimmer, H., 1989. “Common-sense Mentalismus und Emotion: Entwicklungspsychologische Implikationen”. In: Roth, E., (ed.), *Denken und Fühlen*. Berlin: Springer, 56–66.

Wright, C., 2014. “Comment on Paul Boghossian, ‘What is inference’”. *Philosophical Studies*, 169, 1, 27–37.

Wu, W., 2014. “Being in the Workspace, From a Neural Point of View: Comments on Peter Carruthers, ‘On Central Cognition’”. *Philosophical Studies*, 170, 1, 163–174.

Yeung, N., and Summerfield, C., 2012. “Metacognition in human decision-making: confidence and error monitoring”. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 1310–1321.

Zalla, T., Miele, D., Leboyer, M., and Metcalfe, J., 2015. “Metacognition of agency and theory of mind in adults with high functioning autism”. *Consciousness and Cognition*, 31, 126–138.

Zuckerman, M., DePaulo, B., and Rosenthal, R., 1981. “Verbal and nonverbal communication of deception”. In: Berkowitz, L., (ed.), *Advances in Experimental Social Psychology*. New York: Academic Press, 1–59.