VILNIUS UNIVERSITY

Paulius
RIMKEVIČIUS

The Opacity of Mind and Experiments on Intuition, Meditation, and Free Will

DOCTORAL DISSERTATION

Humanities,
Philosophy H 001

VILNIUS 2019

This dissertation was written in 2014–2019 at Vilnius University.
The research was supported by the Research Council of Lithuania.

Academic supervisor – prof. dr. Marius Povilas Šaulauskas
(Vilnius University, humanities, philosophy – H 001).
Academic consultant – dr. Renatas Berniūnas
(Vilnius University, social sciences, psychology – S 001).

VILNIAUS UNIVERSITETAS

Paulius
RIMKEVIČIUS

Sąmonės neskaidrumas ir intuicijos, meditacijos bei laisvos valios eksperimentai

DAKTARO DISERTACIJA

Humanitariniai mokslai,
filosofija H 001

VILNIUS 2019

Disertacija rengta 2014–2019 metais Vilniaus universitete.
Mokslinius tyrimus rėmė Lietuvos mokslo taryba.

Mokslinis vadovas – prof. dr. Marius Povilas Šaulauskas
(Vilniaus universitetas, humanitariniai mokslai, filosofija – H 001).
Mokslinis konsultantas – dr. Renatas Berniūnas
(Vilniaus universitetas, socialiniai mokslai, psichologija – S 006).

CONTENTS

INTRODUCTION

> ‘This problem is so important that I have studied it for much of my career.
> Or wait—is it the other way around? ’
>
> Wilson 2009: 384

In contemporary philosophy of cognitive science, a widely debated question is how one knows one’s own mind—the question of self-knowledge. A more particular question, which has received especially close attention in this debate, is how one knows one’s own propositional attitudes. Examples of attitudes, in the intended sense, are beliefs, desires, and intentions, as well as judgements and decisions. Two broad views of how one knows one’s own attitudes have been proposed in the literature. The first view claims that one knows one’s own attitudes in the same way that one knows others’ attitudes (Ryle 1949/2009, Wittgenstein 1953/2009, Sellars 1956, Bem 1967, Nisbett & Wilson 1977, Gopnik 1993, Gazzaniga 1998, Wegner 2002/2017, Wilson 2002, Carruthers 2011, Schwitzgebel 2011, Cassam 2014). The second view claims that one also knows one’s own attitudes in a way in which one does not know others’ attitudes (Armstrong 1968/2002, Lycan 1996, Moran 2001, Nichols & Stich 2003, Bar-On 2004, Frankish 2004, Bilgrami 2006, Goldman 2006, O'Brien 2007, Gertler 2011, Burge 2013, Fernández 2013, Proust 2013, Coliva 2016, Byrne 2018). One can call these two views ‘the symmetrical theory’ and ‘the asymmetrical theory’ (see Schwitzgebel 2019).

Relevance

The debate between the symmetrical and the asymmetrical theories of our knowledge of our own attitudes has implications for many classical and contemporary philosophical questions. Here are four of them. The first of these questions is how to know oneself better (Wilson 2009). Arguably, if the symmetrical theory is true, then ways to know one’s own attitudes better are roughly the same as ways to know others’ attitudes better. The second of these questions is whether self-knowledge is suited to serve as foundation for other kinds of knowledge (Gertler 2011). Arguably, if the symmetrical theory is true, then knowledge of one’s own attitudes is no better suited to serve as foundation for other kinds of knowledge than knowledge of others’ attitudes. The third of these questions is whether free will and moral responsibility exist (Carruthers 2011). If the symmetrical theory is true, then there is no free will or moral responsibility, at least not in a sense that presupposes conscious decisions. The fourth of these questions is whether philosophy has its own method, different from those of the special sciences, such as psychology and linguistics (Rey 2013b). Again, arguably, if the symmetrical theory is true, then there is no method that is special to philosophy, at least not of the kind that presupposes conscious philosophical judgements.

Method

The approach to self-knowledge adopted in this dissertation is the approach of the symmetrical theory. More specifically, it is that of the interpretive-sensory access theory of self-knowledge (Carruthers 2011). The ISA theory makes four main claims and six main empirical predictions that follow from those claims. All of its main rivals make claims that contrast with at least one of the main claims made by the ISA theory. All versions of the asymmetrical theory also make predictions that contrast with at least one of the six main predictions made by the ISA theory. Therefore, taken together, the following claims and predictions distinguish it from all of its rivals. Here they are in a nutshell (to be further explicated in the first chapter below).

Here are the main claims. (1) There is a single mental faculty underlying our attributions of attitudes, whether to ourselves or to others. (2) Its access to its own domain is sensory. (3) Its access to most kinds of attitudes is interpretive. (4) It evolved for attributing mental states to others. (See also Carruthers 2011: 1–2)

Here are the main predictions. (1) If there is no sensory basis for attributing an attitude, then there will be no attribution of it. (2) If a child is not yet capable of attributing an attitude to others, then it will not yet be capable of attributing it to itself; if it is already capable of attributing it to others, then it will already be capable of attributing it to itself. (3) If one's capacity for attributing an attitude to others is impaired, then one's capacity for attributing it to oneself will also be impaired, and vice versa; if a brain region is activated when attributing an attitude to others, then it will be activated when attributing it to oneself, and vice versa. (4) If one does not undergo effortful training in this domain, then one's capacity for monitoring one's attitudes will not be very reliable; moreover, one's capacity for controlling them will be broadly behavioural. (5) If one is presented with a misleading sensory basis for attributing an attitude, then one will misattribute the attitude. (6) If an animal is incapable of attributing an attitude to others, then it will be incapable of attributing it to itself; if it is capable of attributing it to others, then it will be capable of attributing it to itself. (See also Carruthers 2011: 370)

Aims

This dissertation has four main aims. The first is to review the available arguments for and against the ISA theory. The second is to reply to arguments against it. The third is to provide new arguments in favour of it. The fourth is to propose new empirical studies to further test it.

Previous work

The debate between the symmetrical and the asymmetrical theories of how we know our own attitudes has long historical roots. The asymmetrical theory appears in the work of modern philosophers René Descartes, John Locke, and Immanuel Kant (Descartes 1637/2006, Locke 1689/1975, Kant 1781/1997; see also Renz 2017). The appearance of the asymmetrical theory is primarily related to developments in epistemology. In particular, it is primarily related to the debate about the foundations of knowledge and to foundationalism. The symmetrical theory appears in the work of twentieth-century philosophers Ludwig Wittgenstein and Gilbert Ryle (Ryle 1949/2009, Wittgenstein 1953/2009, Bem 1967; see also Gertler 2011). The appearance of the symmetrical theory is primarily related to developments in psychology. In particular, it is primarily related to the debate about the method of psychology and to behaviourism. The ISA theory appears in the work of contemporary philosopher Peter Carruthers and is most thoroughly presented in the *The Opacity of Mind* (Carruthers 2011). Its appearance is primarily related to developments in psychology. In particular, it is primarily related to the debate about the development of children's capacities to attribute mental states and to the 'theory' theory (see also Carruthers 2009).

Carruthers presents four main arguments in favour of the ISA theory (Carruthers 2011). The first is that it receives support from the available empirical research. The second is that it makes new predictions that contribute to new empirical research. The third is that it is simpler than its rivals. The fourth is that it coheres with surrounding already well-established theories. A common argument against the theory is that it is counterintuitive. Carruthers counters it by arguing that it was adaptive for people to evolve the incorrect intuition about self-knowledge (Carruthers 2011: 39–45).

The ISA theory is controversial. Many researchers working in the field more or less tentatively endorse it, while many others more or less explicitly criticise it (*pro*: Carruthers 2006, 2007, 2008, 2009, 2010, 2011, 2013a, 2013b, 2015, 2017, 2018, Carruthers & Ritchie 2012, Cassam 2014, 2017, Frankish in conversion 2018, Guerini et al. 2015, King & Carruthers 2011, forthcoming, Knappik 2015, Levy 2012, 2014, Marraffa 2014, McGlynn 2012, Mercier & Sperber 2017, Nagel 2014, in conversation 2018, Nicholson et al. 2019, Vierkant 2015, in conversation 2018, Schwengerer, in conversation 2018, Westra & Carruthers 2017, Williams et al. 2018; *contra*:

Allen & May 2014, Antony & Rey 2016, Bar-On 2015, Bermúdez 2013, Byrne 2012, 2018, Dokic 2012, Doris 2015, Coliva 2016, Fernández 2013, Frankish 2009, 2012, 2016, Fricke 2014, Gertler 2011, 2015, Goldman & Jordan 2013, Hurlburt 2011, Keeling 2018, Moran 2017a, 2017b, Newen 2015, Nichols forthcoming, Peters 2014a, 2014b, 2014c, 2018, Proust 2012, 2013, 2016, Rey 2008, 2011, 2012, 2013a, 2013b, Schwengerer unpublished, 2019, Schwitzgebel 2011, 2012, 2013, Serban 2014, Shepherd 2013).

Novelty

The novelty of this dissertation principally consists of the following four contributions. The first is a review of arguments for and against the ISA theory. The second are replies to arguments against it. The third, and the most important, are three new arguments in favour of it. The fourth are three proposals for new empirical studies that to further test it. In addition, this dissertation connects the debate surrounding the ISA theory with three areas of empirical research that have previously received little or no attention in it: empirical research on intuition, meditation, and free will. As the Claims make evident, the view defended in this dissertation also diverges somewhat from the one defended by Carruthers, most notably with regard to intuition.

Claims

The central claim defended in this dissertation is that the ISA theory receives support from empirical research on intuition, meditation, and free will. The central claim is supported by three main arguments. The first is that the ISA theory receives support from empirical research on intuition because they both suggest that attribution of attitudes to oneself is either unconscious, or conscious and interpretive. Second, the ISA theory receives support from empirical research on meditation because both suggest that meditators misattribute attitudes to themselves and find their own thoughts difficult to control. Third, the ISA theory receives support from empirical research on free will because both suggest that self-attribution of intentions is based on external and internal evidence: perceptual feedback and mental imagery.

Structure

This dissertation has two parts. The first presents the current state of the debate surrounding the ISA theory. The second presents three new arguments in favour of the ISA theory. Each part has three chapters. Each chapter ends with a short concluding section.

The first chapter presents the ISA theory. The first section specifies what it aims to explain. The second specifies its four main claims and contrasts them with those of the main rivals. The third specifies its six main predictions and contrasts them with the corresponding predictions of its main rivals. The fourth specifies four of its wider implications.

The second chapter overviews the empirical evidence. The first section presents evidence on apparent non-sensory awareness. The second presents evidence on childhood development. The third presents evidence on apparent dissociations in impairment or activation. The fourth presents evidence on monitoring and control of cognitive states (metacognition). The fifth presents evidence on misattribution. The sixth presents evidence on other animal species (comparative evidence).

The third chapter overviews relevant explanatory considerations. The first section discusses scientific fruitfulness, or whether the theory makes new predictions that contribute to new empirical studies. The second discusses relative simplicity, or whether it is simpler than its rivals. The third discusses external coherence, or whether it coheres with surrounding already well-established theories. The fourth discusses intuitive appeal, or whether it is intuition-friendly.

The fourth chapter argues that the ISA theory receives support from empirical research on intuition. The first section defines intuition and argues that it is relevant to the present debate. The second argues that laypeople's intuition is in line with the theory. The third argues that the intuition of some experts is also in line with the theory. The fourth hypothesises about the motives behind the intuitions and proposes an empirical study to further test the theory.

The fifth chapter argues that the ISA theory receives support from empirical research on meditation. The first section defines meditation and argues that it is relevant to the present debate. The second argues that meditators misattribute attitudes to themselves. The third argues that meditators find their thoughts difficult to control. The fourth hypothesises about the mental mechanisms behind meditation and proposes an empirical study to further test the theory.

The sixth chapter argues that the ISA theory receives support from empirical research on free will. The first section presents the paradigm of a free will experiment and argues that they are relevant to the present debate. The second argues that self-attribution of intentions is based on external evidence (perceptual feedback). The third argues that it is based on internal evidence (mental imagery). Finally, the fourth hypothesises about the relation between feelings of agency and judgements of agency and proposes an empirical study to further test the theory.

I. FOUNDATIONS

1. THE THEORY

The ISA theory aims to explain self-knowledge. That is its target. However, for various reasons, its focus is much more narrow than that. It focuses on self-attributions of judgements and decisions. The ISA theory provides an explanation of these self-attributions that, in a nutshell, is given in its four main claims. Yet not all of these claims are equally central to the theory. At its very center are the three claims that refer to our faculties as we find them today. The ISA theory makes six main predictions that follow from the four main claims. Again, for various reasons, not all of these predictions stand in equally evident contrast with the predictions of its rivals. Its most distinctive prediction is the one that speaks about the pattern of misattributions. Finally, the ISA theory has many wide implications. Yet some of these implications will be much more contested than others. Its clearest wider implication, and one that has already given rise to considerable debate, concerns free will and moral responsibility.

1.1. Target

Here is why the ISA theory focuses on a very particular part of its target. The target is self-knowledge. However, knowledge of the body, such as of its height or weight, is relatively unproblematic. Therefore, the theory focuses on the mind. Knowledge of more complex features of the mind, such as of one's personality, depends on knowledge of its simpler features. Therefore, the theory focuses on knowledge of the more simple features of the mind. Some of its simpler features, such as particular memories, beliefs, desires, and intentions, are 'standing' mental states. You can have them even when you are asleep. Knowledge of standing mental states depends on knowledge of activated or 'occurrent' mental states. Consequently, the theory focuses on knowledge of one's own occurrent mental states.

One kind of occurrent mental states are sensations. Here, the term is used in the broad sense that includes perceptual, quasi-perceptual, interoceptive, and proprioceptive mental states. Perceptual mental states include visual, auditory, olfactory, gustatory, and tactile ones. An example of a perceptual mental state is the state of seeing a green field. Quasi-perceptual mental states include the imaginative counterparts of perceptual ones. An example of a quasi-perceptual mental state is the state of 'hearing' a sentence uttered in inner speech. Interoceptive mental states include those relating to stimuli produced within an organism, especially in the gut and other internal organs. An example of an interoceptive mental state is the state of feeling pain in the stomach. Finally, proprioceptive mental states include those relating to stimuli produced within the organism and, more specifically, to the position and movement of the body. An example of a proprioceptive mental state is the state of feeling of your legs being crossed. All of these mental states fall under the category of sensations.

Another kind of occurrent mental states are propositional attitudes. Call them ‘attitudes’, for short. Psychologists and philosophers tend to use the term differently (Carruthers 2011: xiii). Psychologists typically use the term to refer to a disposition to engage in evaluative behaviour. In this sense one’s attitude might be positive or negative. Philosophers tend to use the term to refer to a thought with propositional content that can usually be expressed in a that-clause. In this sense, one’s attitude can be that of believing, desiring, or intending. These three examples are examples of ‘standing’ attitudes. An attitude can also be that of wondering, supposing, judging, or deciding. The last four examples are examples of ‘occurrent’ attitudes. Some of them are probably more basic than others. Judgements and decisions are probably basic. Here are some working definitions of them. A judgement can be defined as the event that ends reasoning about what is the case and gives rise to a belief. A decision can be defined as the event that ends reasoning about what to do and gives rise to an intention.

Knowledge of one’s own judgements and decisions is at the point of focus of the ISA theory. Knowledge of one’s own sensations is discussed in a relatively peripheral way. This is because the explanation of it offered by the ISA theory is relatively simple and uncontroversial. Nevertheless, sensations themselves, as opposed to the process of coming to know them, do play a more central role in the discussion. This is because the theory claims that we come to know our own attitudes based on them.

The focus of the theory is also more narrow in another sense. It is the actual self-attribution of attitudes, whatever its features might be. For instance, if it is reliable, then it produces knowledge, at least on some theories of what knowledge is. Carruthers claims that it does produce ‘knowledge’, but he also claims that it is ‘not very reliable’ (Carruthers 2011: 67). Similarly, if it is more reliable than attribution of attitudes to others, then it is ‘privileged’, in philosophers’ parlance. Carruthers claims that this is ‘moot’ and ‘unresolved’ (Carruthers 2011: 70; see also Allen & May 2014). Likewise, if it uses a different method than the one used in attribution of attitudes to others, then it is ‘peculiar’. Carruthers claims that it is only peculiar in the sense that there are additional sources of evidence involved, namely: quasi-perceptual, interoceptive, and proprioceptive evidence (Carruthers 2011: 2–3). Finally, if it is not to be challenged by others, then it is ‘authoritative’. Carruthers claims that it is to be challenged by others and suggests that the reason why this is rarely done is the intuitive assumption of ‘transparency’ (Carruthers 2011: 12). To sum up, the theory primarily aims to explain the actual attribution of attitudes to oneself, not knowledge, and much less privileged, peculiar, or authoritative knowledge.

This means that its target is somewhat different from the targets of most of the other theories in the field. Its rivals often aim to explain why self-knowledge is privileged, peculiar, and authoritative (see Gertler 2015). This does not mean that the ISA theory does not challenge them. It does challenge them because it aims at a more basic, common target. If the theory turned out to be true, then many of the things that its rivals aimed to explain would turn out to be illusory. That is surely to challenge them.

1.2. Claims

Here is a more thorough explication of how the ISA theory aims to explain self-knowledge. As noted above, the core of the theory is a conjunction of four main claims (see also Carruthers 2011: 47–78). Let us first take them in turn and then contrast them with those of the alternative theories.

The first of the four main claims of the ISA theory is that there is a single mental faculty underlying our attributions of attitudes, whether to ourselves or to others. It means that the cognitive process of attributing an attitude is essentially the same in the cases of self and other. A cognitive process can be defined by its inputs, processing rules, and outputs. The claim is then that the inputs, processing rules, and outputs are the essentially same in both cases. It also says that a mental faculty underlies these attributions. A mental faculty can be defined as a part of the mind that functions in relative isolation from other parts. The claim is then that the process of attributing attitudes goes on in relative isolation from other cognitive processes.

A faculty dedicated to attributing attitudes to others is commonly referred to as ‘the mindreading faculty’. A faculty dedicated to attributing attitudes to oneself is commonly referred to as ‘the introspection faculty’. The claim is then that we have no introspection faculty. Instead, the mindreading faculty is repurposed for attributing attitudes to oneself as well.

The second of the four main claims of the ISA theory is that the faculty’s access to its own domain is sensory. It means that the input consists of sensory information. It includes perceptual, quasi-perceptual, interoceptive, and proprioceptive information. Carruthers also claims that some kinds of conceptual representations can be ‘bound into’ sensory representations and then used as input (Carruthers 2011: 72–73). An example of an experience that is a conceptual and sensory bundle is when one sees a cat as a cat, and not merely as a combination of shapes, colours and textures.

Carruthers claims that one ‘recognises’ these kinds of mental states, meaning that one does not need to interpret them (Carruthers 2011: 74). By ‘recognition’ he means that one attributes these kinds of experience to oneself without much further processing by the mindreading faculty.

The third of the four main claims of the ISA theory is that the faculty’s access to most kinds of attitudes is interpretive. It means that the processing rules are rules of interpretation. Here, ‘interpretation’ refers to the kind of cognitive process that takes place when one attributes attitudes to others. In particular, it refers to the process of inferring an attitude based on someone’s verbal and non-verbal behaviour and its context by using general rules of an intuitive theory of mind (usually unconsciously). For example, I might infer that you like this game more than that one based on observing that you spent more time playing it and by using the rule that: ‘people spend more time playing a game they like’. The inference might be wrong. It could be modified, for example, if I found out that you were asked to pretend. The claim is then that one attributes attitudes to oneself by inferring them using whatever rules one uses to infer the attitudes of others. In fact, there is some evidence that people easily misinterpret their own preferences using the rule from the example above (Bem 1967).

It also means that access to some attitudes is not interpretive. The claim is that some of them can also be bound into sensory representations. Carruthers notes two such exceptions. The first kind are so-called ‘perceptually-based judgements’ (Carruthers 2011: 75). An example of a perceptually-based judgement is when one sees that there is a cat on the mat, which directly gives rise to a belief that there is a cat on the mat. The reason why this is said to be an exception to the general rule that judgements are interpreted is that such experiences play a very similar role to that of a judgement. The second kind are so-called ‘context-bound affective attitudes’ (Carruthers 2011: 146).

An example of a context-bound affective attitude is when one wants to take this picture home and not that one, which one can reliably report. Here is the reason why the object of one’s affect and the valence of the affect are said to be an exception to the general rule that affective attitudes are interpreted. It is that they are not susceptible to mistakes coming from misinterpretation. Whereas attributions of the features that make the object desirable or attributions of long-term desire are susceptible those mistakes. In fact, there is some evidence that people easily misinterpret why they like a picture and whether they will like it in a few days (Wilson et al. 1989).

The last of the four main claims of the ISA theory is that the faculty evolved for attributing mental states to others. It means that attributions of attitudes to others came first in the course of evolution. This claim differs from the others because the others describe the current state of affairs and not how the current state of affairs came to be. For this reason, the first three claims might turn out to be true even if the fourth claim turned out to be false. For the same reason, the fourth claim is less central to the ISA theory (see also Carruthers 2011: 2). But if it turned out to be false, then the theory would seem less plausible from an evolutionary perspective. Then there would have to be an explanation of this strange turn in evolution: why is attribution of attitudes to others primary in humans, when this is not so in other species? It is conceivable that some creature could develop the introspection faculty instead of the mindreading faculty or that it could develop both. But if it turned out that no other animal is like this, or at least that none of our close relatives, who feel similar evolutionary pressures, are like this, then it would seem more likely that humans should only evolve the mindreading faculty.

These four main claims of the ISA theory contrast with those of its main rivals. All rival theories are incompatible with at least one of the four main claims made by the ISA theory. Here are the most important of them.

The first and the most important rival of the ISA theory is the inner sense theory. The inner sense theory claims that there is a mental faculty dedicated to the attribution of mental states to ourselves (Locke 1689/1975, Armstrong 1968/2002, Lycan 1996, Nichols & Stich 2003, Goldman 2006). It claims that the faculty is similar to the faculties underlying our external senses, hence the term ‘inner sense’. One version of the inner sense theory claims that the capacity to attribute mental states to oneself and the capacity to attribute them to oneself are independent (Nichols & Stich 2003). Another version of the inner sense theory, ‘the simulation theory’ (Goldman 2006), claims that the former capacity enables the latter.

In contrast to the ISA theory, the inner sense theory claims that there are two mental faculties underlying our attributions of attitudes: one underlying our attributions of attitudes to others (the mindreading faculty), and one underlying our attributions of attitudes to ourselves (the introspection faculty). Likewise, it claims that in the process of attributing an attitude to oneself input includes attitudinal information, not only sensory information, and processing rules are recognitional, not interpretive. Finally, it claims that the introspection faculty evolved in order to enable attributing mental states to oneself (see especially Nichols and Stich 2003: 150–165).

The second of the three most important rivals of the ISA theory is the transparency theory. The transparency theory claims that one attributes attitudes to oneself by attending to tracts of the outside world that one's attitudes are about (Evans 1982, Moran 2001, Fernández 2013, Byrne 2018). To take the famous example by Gareth Evans, when someone asks you whether you think there will be a third world war, you direct your attention to the geopolitical situation, rather than to your own mind (Evans 1982: 225). The term 'transparency 'refers to this supposed feature of attitudes that they themselves are as if seen through when one attributes them to oneself.

In contrast to the ISA theory, the transparency theory claims that general reasoning capacities underly the self-attributions of attitudes, not a dedicated faculty. Likewise, it claims that, in the process of attributing an attitude to oneself, input consists of information on the tracts of the outside world that the attitude is about, not of sensory cues relevant for interpretation. Yet these things partially overlap. Perhaps an example might elucidate the difference. A nod might serve as basis for interpreting that one decided to go out, as suggested: the nod is a sensory cue relevant for interpretation, but not what the decision is about. Similarly, it claims that the processing rules used are specific to the self case, as well as relatively simple (e.g., 'if *p*, believe that you believe that *p*'), not interpretive. This means that they are not sensitive to such sources of evidence as what one knows about one's own behaviour. Finally, it should claim that we primarily evolved to use 'the transparency procedure' for attributing mental states to ourselves and whatever further evolutionary purposes that might serve.

The last of the three most important rivals of the ISA theory is the constitutive theory. The constitutive theory claims that having an attitude and believing that you have it are parts of the same mental state (Shoemaker 1994, Bilgrami 2006, Boyle 2009, Roessler 2013, Coliva 2016). For example, the version proposed by Johanness Roessler claims that to have an attitude is also to believe that you have it, although that belief is usually only implicit and one might need to make it explicit for some purposes (Roessler 2013). The term 'constitutive' is meant to underline that the belief is supposed to be a part of what makes the attitude what it is and not merely a mental state that is in close causal connection with the attitude.

In contrast to the ISA theory, the constitutive theory claims that there is no need for a distinct faculty to enable attributions of attitudes to oneself. Similarly, it claims that there is no need for there to always be a basis for the attribution in one's sensations or any need to always rely on interpretations. Finally, it is in no way obviously committed to any story of how these dual mental states evolved. However, it does probably imply that most other species of animals do not have attitudes, at least not in the full sense of the term. A possibility to consider is that humans might have evolved them since it was especially advantageous for them to learn to make their mental states explicit, perhaps primarily in a cooperative context.

There are many more rivals. In their overviews of contemporary theories of self-knowledge, Brie Gertler notes eight and Eric Schwitzgebel notes nine (Gertler 2015, Schwitzgebel 2019). Each theory that they mention has been contrasted with the ISA theory elsewhere (see Carruthers 2011, especially: 7–8, 17–25). The reason why the following discussion focuses on only three of them is that they are the most widely endorsed and the strongest contenders. In particular, the inner sense theory is the strongest contender for empirical support. The transparency theory and the constitutive theory are strong contenders for some of the relevant theoretical virtues: simplicity, in the case of the transparency theory, and intuitiveness, in the case of the constitutive theory. Consequently, the following primarily focuses on the inner sense theory and, to a lesser extent, the transparency theory and the constitutive theory.

Nevertheless, most points, including the three new arguments presented below, will apply to all existing versions of the asymmetrical theory. This is because they are all incompatible with the claim that access to most of one's own attitudes is interpretive. By denying this, they are denying the third of the four main claims of the ISA theory. Theories that fall under this category also include the 'acquaintance' theory, the 'reasons' theory, the 'rationalist' theory, the 'agentialist' theory, and the 'expressivist' theory (Gertler 2015; see also Schwitzgebel 2019). These arguments are a challenge to all of them.

Finally, a note is due on why the discussion focuses on one particular version of the symmetrical theory, the ISA theory. The reasons are similar. This is because the most prominent other version of the symmetrical theory is neither widely endorsed nor a strong contender. Nevertheless, it remains a theoretical possibility that might shed some new light on the ISA theory. So here are its main claims.

The most important other version of the symmetrical theory is the behaviourist theory. The behaviourist theory claims that one attributes attitudes to oneself based on interpreting one's own external behaviour (Wittgenstein 1953/2009, Ryle 1949/2009, Bem 1967). The view is primarily associated with Ryle. Hence, other versions of the symmetrical theory are sometimes called 'neo-Rylean' (Byrne 2012). In spite of this, it is debatable, whether Ryle or anybody else really held this extreme version of the symmetrical theory (see Tanney 2009). All of these theories are sometimes called 'self/other parity accounts' (Schwitzgebel 2019). The reason why the behaviourist theory is more extreme than the ISA theory is that it insists on parity not only for attitudes, but for all mental states: 'in principle, as distinct from practice, John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe '(Ryle 1949/2009: 138).

The behaviourist theory is compatible with the claim that there is only one faculty underlying our attributions of attitudes, whether to ourselves or to others. However, a traditional behaviourist might want to refrain from talk of faculties and talk about capacities instead. It also claims that the process is interpretive. Similarly, it is compatible with the claim that the evolution of the capacity to attribute attitudes was primarily driven by the need to attribute them to others. In contrast to the ISA theory, the behaviourist theory claims that one interprets attitudes based on external behaviour and context only, so it only allows for a more narrow inferential basis (e.g., it excludes inner speech from it). It also allows no exception to the rule that all mental states are interpreted, not for sensory states, and not for any kinds of attitude (e.g., perceptually-based judgements are known by interpretation on this theory). These differences make the ISA theory a much stronger contender, so the following discussion focuses on it and its three main rivals.

1.3. Predictions

Here is a more thorough explication of the six main predictions that follow from the claims made by the ISA theory (see also Carruthers 2011: 3–7). Certain predictions contrast more clearly with those made by all of its rivals, while others contrast only with some of them. Nevertheless, the claims of the theory are sufficiently distinct to make it possible to pit it against any of the main rivals in empirical research.

The first of the six main predictions made by the ISA theory is this: if there is no sensory basis for attributing an attitude, then there will be no attribution. It follows from the claim that self-attributions of attitudes are always based on the interpretation of sensations. But not all sensations count as adequate basis for attributing an attitude. The sensory cues should be relevant in the sense that it would be reasonable to attribute an attitude on that basis. For example, feeling oneself nod while listening to a message is a reasonable basis for the interpretation that one believes what is being said. In fact, there is some evidence that people easily misinterpret nodding in this way (Briñol & Petty 2003). In contrast, hearing the noise of traffic outside the window is not a reasonable basis for the same interpretation. More generally, relevance is determined by the actual processing rules of the intuitive theory of mind.

The prediction implies that there should be no self-attributions of what is known in the literature as ‘unsymbolised thinking’. Unsymbolised thinking is defined as ‘an explicit, differentiated thought that does not include the experience of words, images, or any other symbols’ (Hurlburt & Akhter 2008: 1364). The prediction implies that one should only attribute an attitude to oneself when such symbols or other relevant sensory cues are present.

In contrast, most of the asymmetrical theories should predict attributions of attitudes to oneself in absence of relevant sensations. They should predict that since they claim that the presence of sensations relevant to interpretation is not necessary for an attribution of an attitude to oneself to be made. For instance, the inner sense theory should predict that one will attribute a mental state to oneself whenever one makes use of the introspection faculty. There is no reason why this should always coincide with the presence of sensations that are relevant for interpretation. Nor does the constitutivist theory have any reason to claim that one should only make one’s beliefs about one’s own attitudes explicit when such sensory cues are present.

The transparency theory is a more complicated case. It should predict that one will attribute attitudes to oneself only when one is attending to what the attitudes are about. Attending to what an attitude is about is always a relevant cue for attributing that attitude. If one can only attend to sensory cues, then one will self-attribute an attitude only in presence of sensations relevant to interpretation. However, it should also predict that one will not attribute an attitude to oneself if the sensation is not related in the right way to what the attitude is about. For instance, it should predict that one will not attribute fear to oneself based on the feeling that one’s legs are shaking if that is not what one is afraid of. So the prediction is still different from the one made by the ISA theory.

The second of the six main predictions made by the ISA theory is this: if a child is not yet capable of attributing an attitude to others, then it will not yet be capable of attributing it to itself; if it already is capable of attributing it to others, then it will already be capable of attributing it to itself. It follows from the claims that there is a single faculty underlying these capacities and that it evolved for attributing attitudes to others. It predicts, for example, that children will only be capable of attributing a false belief to themselves when they are already capable of attributing a false belief to others. However, if there is a difference in when children pass the corresponding tests that is explained by, say, linguistic demands, then this does not challenge the ISA theory. The prediction only concerns underlying conceptual capacities for understanding one’s own and other minds.

In contrast, the simulation theory should predict that if a child that is capable of attributing an attitude to itself, it will be capable of attributing it to others; and if it is incapable of attributing it to itself, then it will be incapable of attributing it to others. This follows from its claim that other-attributions depend on self-attributions. It is unclear what the other versions of the inner sense theory, the transparency theory, or the constitutive theory should predict in this respect. But it is at last clear that there is no obvious reason for any of the asymmetry theories to predict that mindreading will come first in development.

At this point it is important to note that making no prediction for a given domain is not an advantage of a theory. It is not an advantage because it means that the theory has a more narrow field of application and is less amenable to empirical testing. These are not virtues.

The third of the six main predictions made by the ISA theory is this: if the capacity for attributing an attitude to others is impaired, then the capacity for attributing it to oneself will also be impaired, and vice versa: if the capacity for attributing an attitude to oneself is impaired, then one's capacity for attributing it to others will be impaired; relatedly, if a brain region is activated when attributing an attitude to others, then it will be activated when attributing it to oneself, and vice versa: if a brain region is activated when attributing an attitude to oneself, then it will be activated when attributing an attitude to others. It follows from the claim that the same faculty underlies both capacities. This implies that literally the same mechanism in the brain is responsible for them. It predicts, for example, that autism, which is known to affect attitude attribution to others, will also affect attitude attributions to oneself. Similarly, it predicts that the medial prefrontal cortex, which is known to be activated when attributing attitudes to others, will also be activated when attributing attitudes to oneself. It primarily concerns systematic impairments and general patterns of activation: it allows for minor differences.

In contrast, all the other theories should predict dissociations in impairment and brain activation related to the two capacities. It follows from their claim that the two capacities are not underlain by the same faculty. There is no reason why the introspection faculty, one's general reasoning capacities, or the ability to make one's attitude explicit should always be affected when the mindreading faculty is affected. Likewise, it is implausible that the introspection faculty, one's general reasoning capacities, or the ability to make one's attitude explicit should have their base in the same areas of the brain as the mindreading faculty.

The fourth of the six main predictions made by the ISA theory is this: if one does not undergo effortful training in this domain, then one's capacity for monitoring one’s own attitudes will not be very reliable; moreover, one’s capacity for controlling one’s own attitudes will be broadly behavioural. In other words, in absence of training, people should lack any deep or well-developed metacognitive capacities. It follows from the claim that access to one’s own attitudes is of the same kind as access to others’ attitudes and two other plausible assumptions: that monitoring of others’ attitudes is not very reliable and that control of others ’attitudes is behavioural. It does allow for some differences, however. It claims that the evidence base is wider in one’s own case, so it is not committed to saying that the reliability will be exactly the same. Likewise, it really claims that attitude control in one’s own case is behavioural in a rather broad sense. In case of self, one can also use inner speech and other internal promptings. For instance, repeating an encouraging phrase in inner speech in order to stick to a decision is something one cannot do for anyone else even though that is analogous to saying it in outer speech and therefore can be said to be broadly behavioural.

In contrast, the inner sense theory should predict deep and well-developed metacognitive competence in absence of training. It follows from the claim that one has an introspection faculty and the additional assumption that the faculty evolved to increase reliability and facilitate control. Again, it is not that clear what the other alternatives should predict in this domain, since they do not postulate a distinct mental faculty. However, all version of the asymmetrical theory have to explain why one developed a distinct way for knowing one’s own attitudes. It is not clear how the explanation would go if it did not appeal to supposed improvements in reliability and control.

The fifth of the six main predictions made by the ISA theory is this: if one is presented with a misleading sensory basis for attributing an attitude to oneself, then one will misattribute that attitude. It follows from the claim that attribution of attitudes to oneself is based on interpreting sensations. This means that one should be misled about one’s own attitudes in cases that are analogous to those where one would be misled about others’ attitudes. It predicts that, for example, physical arousal will, in some circumstances, mislead one into thinking that one feels attraction when in fact one feels fear. In fact, there is some evidence suggesting people easily misinterpret their physical arousal in this way (see the ‘love on the bridge ’study (Dutton & Aron 1974)).

In contrast, all the main rivals should predict that a misleading sensory cue will not be sufficient to derail one’s attribution of an attitude to oneself. It follows from their claim that the normal way of attributing an attitude to oneself does not rely on interpreting sensations. There is no reason why the normal way should systematically become unavailable when misleading sensory cues are present. By default, they should predict that there will be no mistakes, much less a pattern of them. This is because, on the face of it, none of their main claims implies any particular pattern of mistakes. They could then try to explain away apparent mistakes case by case. They could also adopt an additional assumption that would imply a particular patter of mistakes. For example, they could assume that people turn to interpretation when others suggest it. An experimenter might do this inadvertently, a psychotherapist might do it on purpose. Making this additional assumption would result in a prediction of a different pattern of mistakes from the one predicted by the ISA theory.

The last of the six main predictions made by the ISA theory is this: if an animal is incapable of attributing an attitude to others, then it will be incapable of attributing it to itself; if an animal is capable of attributing an attitude to others, then it will be capable of attributing it to itself. It follows from the claim that the faculty for mental state attribution evolved for the purpose of attributing mental states to others. It predicts, for example, that if capuchin monkeys turned out to be incapable of attributing false beliefs to others, then they will also be incapable of attributing it to themselves. This prediction is less central to the theory than the others, just like the last of the four main claims of the theory, from which it primarily follows. This is because, again, the theory could be right about the current state of affairs and wrong in its explanation of how the current state of affairs came to be.

In contrast, the simulation theory should predict that if an animal is capable of attributing an attitude to itself, then it will be capable of attributing it to others; and if it is incapable of attributing it to itself, then it will be incapable of attributing it to others. This follows from its claim that introspection grounds mindreading. The other main rivals should probably at least predict that there should be an animal that is capable of attributing an attitude to itself but not to others. Again, it is less clear what the others should predict in this domain, since they do not postulate a mental faculty dedicated to mental state self-attribution.

1.4. Implications

Here is a somewhat more thorough explication of four of the ISA theory's wider implications. Since they follow from claims that, as argued above, are distinctive, the implications themselves are most likely distinguish the ISA theory of its main rivals as well.

The first wider implication of the ISA theory is for the question how to know oneself better (Wilson 2009). This question concerns 'substantive', as opposed to ‘trivial’, self-knowledge (Cassam 2014). An example of trivial self-knowledge is knowing that you believe it is raining. An example of substantive self-knowledge is knowing that you are an agreeable person. Although substantive self-knowledge is probably more relevant to the imperative to ‘Know thyself’, which was of central concern for Socrates, modern philosophers have tended to focus on trivial self-knowledge, leaving the investigation of substantive self-knowledge to psychologists. The ISA theory might be better placed to give suggestion in this respect, since it has been developed with an eye for empirical research on metacognition and its more practical applications. The main implication, as noted above, is that the ways to know oneself better should be analogous to the ways to know others better.

The second wider implication of the ISA theory is for the question whether self-knowledge is suited to serve as foundation for other kinds of knowledge (Gertler 2011). Perhaps the main reason why philosophers, especially since Descartes, have focused on trivial self-knowledge is that it is more likely to be especially secure. If it were privileged, peculiar, and authoritative, as Descartes thought that it is, then it could probably serve as foundation for other kinds of knowledge (Descartes 1637/2006). Substantial self-knowledge is rarely thought to be special in this respect. However, if the ISA theory is true, then the epistemic credentials of trivial self-knowledge (of attitudes) are probably no better than those of most other kinds of knowledge. In that case, it is hardly suited to serve the foundational role that was laid out for it.

The third wider implication of the ISA theory is for the question whether philosophy has a method different from the methods of the special sciences, such as psychology or linguistics (Rey 2013). One might argue that the theory claiming that there is such a method presupposes that one’s access to one’s own mind is special. This assumption is clearly made by Descartes. It is, arguably, also made by many contemporary theories of philosophical method. Since the ISA theory denies that one’s access to one’s own mind is special in the way that philosophers typically assume, it could probably be used to argue against such theories.

The fourth wider implication of the ISA theory is for the question whether people have free will and moral responsibility (Carruthers 2011: 379–383). Some theories of free will and moral responsibility claim that free will presupposes conscious intentions, hence the intense debate on whether science has disproved their existence (see Sinnott-Armstrong & Nadel 2011). If they do presuppose conscious intentions, then the ISA theory implies that we have no free will and moral responsibility. It follows from the claim that access to decisions is interpretive. However, there are theories that claim they do not presuppose conscious intentions. For instance, one theory defines a free and responsible action as an action that expresses one's values, and claims that the subject does not need to be conscious of it (Doris 2015). In that case, the ISA theory would not imply that we have no free will and moral responsibility. A lively debate is already taking place whether it does imply it (see Carruthers 2011, King & Carruthers 2011, forthcoming, Levy 2012, 2014, Marraffa 2014, Peters 2014a).

1.5. Conclusion

The primary target of the ISA theory is self-attribution of judgements and decisions. This is because that target is more basic. Its core claim is that we self-attribute judgements and decisions by interpreting sensations. This distinguish it from all of its main rivals, including the inner sense theory, the transparency theory, and the constitutive theory. Its most distinctive prediction is that there will be misattributions of judgements and decisions when sensations are misleading. This also distinguish it from all of its main rivals because they should all make contrasting predictions. Finally, its clearest wider implication is for free will or moral responsibility. This is because if they presuppose conscious decisions and the ISA theory is true, then they do not exist.

2. EMPIRICAL EVIDENCE

This chapter overviews the available empirical evidence relevant to each of the six main predictions of the ISA theory in turn. As noted above, it is always an advantage of a theory if its predictions are supported, rather than challenged, but the advantage is much greater if the evidence also challenges rivals' predictions. In the following, the evaluation of rivals 'predictions is mostly left implicit. However, it should be fairly obvious from the discussion above of the rivals' claims and predictions. The reason for mostly leaving the evaluation of rivals' predictions implicit in this overview is that the relevant literature is vast. For the same reason, the discussion will focus on evidence that was previously not discussed in relation to the ISA theory, only briefly recapitulating the earlier debate.

2.1. No Non-Sensory Awareness

The ISA theory predicts that if there is no sensory basis for attributing an attitude, then there will be no attribution of it. Carruthers argued that this is the case (Carruthers 2011: 214–221). The main challenge he responded to concerned reports of unsymbolised thinking. They came from descriptive experience sampling studies by Russell Hurlburt and colleagues (Hurlburt 2011: 291–308). In these studies, participants wore a beeper signalling at unpredictable intervals that they should report their immediately past inner experience. Some participants reported episodes of unsymbolised thinking.

In brief, Carruthers' response was that people fail to report the sensory cues that are present and, in any case, the reports are not very consistent. First, there are at least two reasons to think that some sensory cues should go unreported. One is that some of them probably appear before or after the moment that the participant is asked to report. Another is that some of them are probably forgotten: partly because they are 'backward masked' by the auditory signal to report and partly because they are fleeting or fragmentary. Second, reports of unsymbolised thinking are not very consistent: some people never report it, others report it only rarely, and yet others later retract their reports.

One significant development in this area concerns the revival of the debate about the existence of sui generis cognitive phenomenology (Bayne & Montague 2011, Breyer & Gutland 2015). Some philosophers claim that what it is like to entertain an amodal thought is irreducible to other kinds of phenomenology, such as the quasi-perceptual experience of hearing oneself speak in inner speech. To the above discussion this debate adds historical perspective and expert reports. It raged in philosophy and psychology roughly a hundred years ago, with many of the same claims being advanced. It is a source of reports from experts in describing one's inner experience, i.e., psychologists and philosophers, who have had many hours of practice and optimal reporting conditions. However, many of the same issues that affect lay reports of unsymbolised thinking also affect experts' reports of sui generis cognitive phenomenology.

First, even with experts, some sensory cues should go unreported, because they come before or after the reported moment of inner experience: either since they are forgotten, or since they are fleeting or fragmentary. Second, even with experts, the reports are not very consistent: some deny that there is such a thing, others disagree which amodal thoughts that have it, and yet others disagree on the fineness of grain. Moreover, the proportion of those experts who do think that there is such a thing is probably smaller than it seems from the philosophical discussion. For this is a self-selected sample, and those who think that they found something rather than nothing are more likely to volunteer their report.

Another significant development in this area concerns meditative cases: cases where the subject has minimal behavioural or contextual cues for interpreting their own state of mind. It has been argued that if there would be no non-sensory awareness of attitudes, then our reports of them would be unreliable in meditative cases (Rey 2013a: 273–274). Following this line of thought, one might predict that, if there is such a thing as unsymbolised thought or sui generis cognitive phenomenology, then one should expect it to be reported more frequently in meditative circumstances.

However, the existing evidence does not support this prediction. A pilot study of descriptive experience sampling with people in a resting state in a scanner, who were non-expert meditators, suggests that people in meditative states have no less sensory cues than usual (Hurlburt et al. 2015). The crucial contribution here is probably made by quasi-perceptual cues, such as mental imagery that is related, for instance, to inner speech. This corroborates the results of an earlier case study of an expert meditator, also in the descriptive experience sampling paradigm, who primarily diverged from the norm in that his reported experiences were more sensory in character (Hurlburt & Heavey 2006: 246). However, since the samples in these studies were very small, one should be very cautious when making conclusions at this point.

2.2. Childhood Development

The ISA theory predicts that if a child is not yet capable of attributing an attitude to others, then it will not yet be capable of attributing it to itself; if it already is capable of attributing it to others, then it will already be capable of attributing it to itself. Again, Carruthers argued that it is the case (Carruthers

2011: 203–209, 240–248). In particular, he argued that children know their own and other people's percepts and goals by one year of age, know others' false beliefs by one or one and a half years of age, and know their own false beliefs by four years of age.

The first of the two main challenges that he responded to concerned reports of knowledge of one's own knowledge, pretence, and perspective, in children who are as yet incapable to attribute those mental states to others (Nichols & Stich 2003: 174–176). In brief, Carruthers' response to this challenge was that the Self and Other conditions were poorly matched in the experiments where such discrepancies were found. The children in the Knowledge and Perspective experiments were better at answering questions about themselves, most probably because they remembered what they themselves saw, whereas they had to infer what the other agent had seen. As for pretence, since the Self and Other conditions were taken from separate experiments, which were not originally meant to be compared, the groups of participants were poorly matched. Moreover, one group was asked what they themselves pretend, while the other group was asked what another pretending agent thinks (the latter group might have inferred that they have to contrast pretending with thinking). Finally, other studies suggest that children who are younger than those in any of these experiments (two years of age) already know when they themselves or other people pretend.

The second of the two main challenges that he responded to concerned an alternative explanation, based on behavioural rules, of reports of knowledge of other people's false beliefs, in children who are as yet incapable of attributing false beliefs to themselves (Nichols & Stich 2003: 170–174). In brief, Carruthers' response was that the most plausible of these explanations have already been ruled out. In particular, children do not seem to use the behavioural rule 'people look where they last saw something', or the behavioural rule 'ignorance leads to error'. Moreover, although one can always think of a behavioural rule that would explain the results, explaining all of them would require many and complex rules, and the more complex they are, the harder they are to test empirically. The ISA theory's explanation is simpler and more amenable to empirical testing.

One significant development in this area concerns replication issues affecting non-verbal studies of children's early knowledge of others' false beliefs. A recent review found over thirty published reports of it in children from six to thirty-six months of age (Scott & Baillargeon 2017). However, there have also been some only partly successful replications as well as failed replications (Baillargeon et al. 2018).

In response to this one might note, following René Baillargeon and colleagues, that there are some procedural differences between some of the original and replication studies, and there are still many original paradigms that remain unaffected by the replication crisis. Moreover, even if all of the non-verbal false belief studies failed to replicate, this would not show that knowledge of one's own false beliefs develops earlier than knowledge of other people's false beliefs. This is because knowledge of one's own false beliefs only develops by four years of age. That is the same age when knowledge of other people's false beliefs develops, according to verbal false belief studies (Wellman et al. 2001). These studies have not been affected by the replication crisis.

Another significant development in this area concerns non-verbal studies of early knowledge of one's own false beliefs. Louise Goupil and colleagues report that infants as young as 12 and 18 months of age respond differently depending on how uncertain the infant is (Goupil et al. 2016, Goupil & Kouider 2016, 2019). The authors interpret these results as suggesting that core metacognitive capacities are already in place in infancy.

One could respond to this in at least three different ways. One is to insist, despite the replication crisis, that there is evidence of knowledge of others' false beliefs at the same age or even earlier. Another way to respond is to argue that in the experiments by Goupil and colleagues the infants merely had to respond based on their certainty, not to monitor their certainty as such or their knowledge as such. In fact, the authors make no explicit suggestion that the abilities they found are meta-representational (involving representations of representations). One could also note that there is some evidence (to be discussed below) suggesting that the implicit metacognition tasks that they used actually measure something quite different from explicit metacognition tasks that clearly measure of meta-representational ability. This also suggests that the infants in the studies by Goupil and colleagues did not monitor attitudes as such.

2.3. Dissociations

The ISA theory predicts that if the capacity for attributing an attitude to others is impaired, then the capacity for attributing it to oneself will also be impaired, and vice versa; also, if a brain region is activated when attributing an attitude to others, then it will be activated when attributing it to oneself, and vice versa. In particular, Carruthers argued that both capacities are systematically impaired in schizophrenia and autism, but not in alexithymia (Carruthers 2011: 293–324). Also, he argued that the brain areas that are activated when using either of the two capacities are these: the medial prefrontal cortex, the posterior cingulate cortex, the temporal pole, the temporo-parietal junction, and the superior temporal sulcus (see also Figure 10.1, from Carruthers 2015: 314).

One of the two main challenges that he responded to concerned purported dissociations of the two capacities in these pathological cases: schizophrenia with passivity symptoms, paranoid schizophrenia, autism, and alexithymia (Nichols & Stich 2003: 178–192). In brief, his response was that it remains highly controversial whether in schizophrenia and autism generally the dissociation exists. As for the cases of schizophrenia with passivity symptoms and alexithymia, the impairment is probably best explained in terms of faults in first-level mechanisms.

More specifically, schizophrenia with passivity symptoms probably involves a fault in the mechanism that compares predicted feedback with actual feedback, concerning either actual, or mentally simulated movement. This mechanism does not represent the mental states as such, and therefore does not engage the person's meta-representational capacities. Alexithymia probably involves a fault in the mechanism that makes the valence of one's affect directly accessible to the person. As noted above, Carruthers claimed that the valence of one's own affect is simply recognised. So the fault lies not with one's meta-representation abilities.

The second of the two main challenges that he responded to concerned purported dissociations between the level of activation and the brain areas involved when exercising the capacities for mindreading and metacognition (Lombardo et al. 2009). In brief, his response was that the results are not consistent across studies and that the tasks testing these capacities were not very well matched. In particular, some of the discussed studies compared remembering one's own mental state with inferring another person's mental state. What is more, since information about oneself is generally more familiar, more emotionally charged, and more deeply processed, it should be expected that activation levels and patterns will be somewhat different in Self and Other conditions.

One important development in this area concerns autism. Together with his colleagues, Carruthers recently conducted three empirical studies testing the ISA theory's predictions. In the first study, they found that healthy people who had more autistic traits performed worse at detecting a lie but just as well on other mindreading tasks as people who had less autistic traits. In the second study, they found that autistic people performed worse at detecting a lie than healthy people (Williams et al. 2018). The authors interpret these results as suggesting that mindreading is significantly impaired in people with autism.

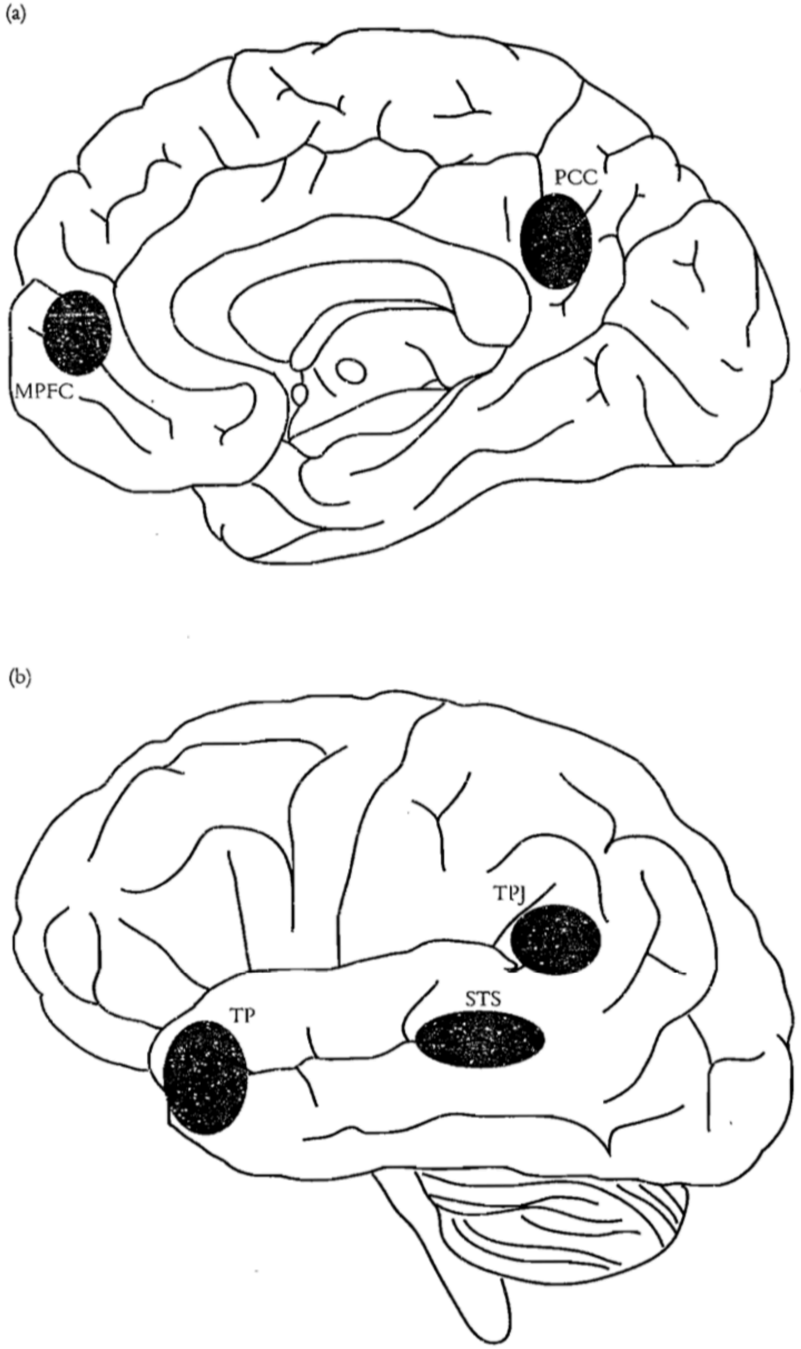

Figure 10.1 The Mindreading Network in the Brain.
Top Figure: Right hemisphere, inside view. MPFC = Medial prefrontal cortex. PCC = Posterior cingulate cortex. *Lower Figure*: Left hemisphere, outside view. TP = Temporal pole. STS = Superior temporal sulcus. TPJ = Temporo-parietal junction.

In the third study, they found that performance on explicit metacognition tasks was associated with performance on mindreading tasks in both healthy and autistic people (Nicholson et al. 2019). Notably, they also found that performance on implicit metacognition tasks was not associated with performance on mindreading tasks or explicit metacognition tasks in either of the groups. Moreover, performance on implicit metacognition tasks was unimpaired in people with autism. According to the authors, this suggests that metacognition and mindreading are equally impaired in autism, while implicit metacognition tasks track first-order decision-making capacities, rather than meta-representational capacities.

Another important development in this area concerns working memory. Carruthers recently put forward a book-length argument to the effect that working memory and consciousness are sensory-based (Carruthers 2015). If this were so, then it would further support the case for the ISA theory, since the cited evidence is quite different. He supported his claim in part by citing empirical studies that show consistent involvement of sensory brain areas when performing working memory tasks.

However, as noted by Wayne Wu, the evidence is ambiguous between merely causal and computational involvement of these sensory areas (Wu 2014). If the ISA theory is right, when performing tasks that engage one's working memory or consciousness and especially when performing tasks that require metacognition, sensory areas should be computationally involved, not merely causally involved.

Yet another important development in this area concerns two other pathologies: anarchic hand syndrome and utilisation behaviour. Uwe Peters noted that there is a case where both of these impairments are present in the same person and argued that the ISA theory cannot explain how this is possible (Peters 2014b). In particular, the person in question denies intending the (inappropriate) movements of their left hand (anarchic hand syndrome), but confirms intending the (inappropriate) movements of their right hand (utilisation behaviour).

If the person's mindreading mechanism is unimpaired, then it is not clear why they incorrectly deny intending the movements of their left hand. If the person's mindreading mechanism is impaired, then it is not clear why they correctly confirm intending the movements of their right hand. If both the mindreading mechanism and the comparator mechanism are impaired, as in schizophrenia with passivity symptoms, then it is not clear why this does not affect both hands equally or why the person does not have an impaired sense of ownership of their own thoughts, like people with schizophrenia with passivity symptoms.

Conceding that this is a difficult case, one could try to respond by noting that the rival theories would probably have difficulties explaining this case as well. For if the supposed introspective mechanism is impaired, then it is not clear why the person correctly self-attributes intentions concerning the right hand. If this is because the task is taken over by the mindreading mechanism, then it is not clear why the mindreading mechanism does not do the same for the left hand. If both mechanisms are impaired, then it is not clear why the person continues to correctly self-attribute intentions that are not related to the hands. Finally, one could also note that caution is needed when making inferences from a single pathological case.

2.4. Metacognition

The ISA theory predicts that if one does not undergo effortful training in this domain, then one's capacity for monitoring one's attitudes will not be very reliable; moreover, one's capacity for controlling them will be broadly behavioural. Carruthers argued this is the case (Carruthers 2011: 263–278). In particular, he argued that one controls one's own learning by initiating behaviour or intervening on it (e.g., rehearsing what one wants to memorise). Likewise, he argued that one relies on heuristics to judge whether one has learned something (e.g., on retrieval fluency). He also argued that one rarely exercises control over one's own attitudes as such, and when one does so, one operates on symbols rather than the thoughts themselves. Moreover, he argued that one uses metacognition primarily for the purpose of improving arguments one would later put to others (this last suggestion is in accordance with the argumentative theory of reason (Sperber & Mercier 2010)).

An important challenge that he had to respond to concerned the link between direct knowledge and direct control. In principle, one could have direct control without direct knowledge, and one could have direct knowledge without direct control. If that is so, then showing that one does not have direct control over one's attitudes does not yet show that one does not have direct knowledge of one's attitudes.

In brief, his response was that theorists in the field usually assume that the capacity for direct knowledge of one's own attitudes evolved primarily for the purpose of enhancing control over one's own attitudes (Carruthers 2011: 66–67). If one only controls one's attitudes indirectly, it is unclear what evolutionary function is the supposed capacity for direct knowledge should serve. This in turn throws doubt on the claim that the capacity did evolve.

One important development in this area concerns situational self-control.

Citing a wide range of evidence, Angela Duckworth and colleagues argue that what is central to self-control is the ability to use situational strategies (Duckworth et al. 2016). This contrasts with the idea that the most important thing in self-control is effort. Situational strategies are ways to circumvent the need to engage in direct encounters with temptations, which would require effort or 'willpower'. These situational strategies are indirect ways of controlling one's own attitudes. Therefore, the suggestion is in line with the prediction made by the ISA theory that people will control their own attitudes by broadly behavioural means.

Another important development in this area concerns being alone with one's thoughts. In a series of eleven studies, Timothy Wilson and colleagues found that people generally perceive the task to entertain themselves with their own thoughts as unpleasant, find it difficult to concentrate on their own thoughts, and prefer to engage in almost any other mundane activity instead, or even to take mild electric shocks that they had earlier said they would pay to avoid (Wilson et al. 2014). These results support the ISA theory's prediction that people will find it difficult to control their own minds in absence of effortful training. Perhaps the results would be somewhat different if the participants were specially trained in the task.

Yet another important development in this area concerns the distinction between evaluative and voluntary processes. If conscious processes are voluntary (as most cognitive scientists agree), and if judgements are not voluntary (as most philosophers agree), then conscious judgements do no exist (Vierkant forthcoming). This line of reasoning could probably be extended to other kinds of attitude. If mental processes cannot be both evaluative and voluntary, and if conscious processes are voluntary, then evaluative processes are not conscious. This would mean that people do not control or access any of their evaluative mental processes directly.

2.5. Misattribution

The ISA theory predicts that if one is presented with a misleading sensory basis for attributing an attitude, then one will misattribute the attitude. In particular, Carruthers argued that misleading sensory cues lead people to mistakes about the causes of their own attitudes as well as the occurrence of particular decisions and judgements. He also argued that people are likewise misled about the more fine-grained properties of the objects of their own affects (as opposed to the valence of those affects or the identity of the objects of those affects) (Carruthers 2011: 147–154, 325–367).

The main challenge that he responded to concerned an alternative explanation of confabulation, which appealed to pragmatic pressures (Rey 2008). In brief, his response was that pragmatic pressures are unlikely to account for all the known cases. This is because in studies on misattribution experimenters usually go to great lengths in order to ensure that pragmatic pressures are minimal (see also Wilson et al. 1989). If one were to concede this point and insist on some other explanation for cases where pragmatic pressures are minimal, then the overall explanation would become more complex. Again, the ISA theory offers a simpler explanation.

An important development in this area concerns one such proposal of an alternative way to explain confabulation. Sophie Keeling has recently argued that instead of appealing to interpretation and misleading sensory cues misattribution should be explained by appeal to a certain desire: the desire to fulfil the obligation to explain one's attitude knowledgeably, with reference to motivating reasons (Keeling 2018). However, there are to things to note about the explanation Keeling offers.

First, although this is offered as an alternative account, the existence of the desire and its influence on mindreading are consistent with the ISA theory. Second, it may be questioned whether this desire alone can account for all the known cases. One consideration for thinking that it cannot is the fact that people misattribute not only reasons for their attitudes or actions, they misattribute mere causes of their attitudes or actions (Olson et al. 2015, Schlegel et al. 2015). Moreover, misattributions are not always first-personal in character. A person's misattribution of their own attitude often parallels their explanation of other people's attitudes or actions in similar circumstances (Bem 1967).

Another important development in this area concerns misattribution of decisions. In order to directly support his claim that people misattribute decisions, Carruthers heavily relied on two empirical studies (Brasil-Neto et al. 1992, Wegner & Wheatley 1999). The problem is that there might be too many methodological issues with these particular studies to warrant the strong conclusions that he draws from them (Shepherd 2013, Peters 2014c, Walter 2014). However, there are now methodologically stronger studies that support at least some of the conclusions drawn from the earlier ones. In particular, people sometimes deny that they have made a decision, when they did in fact make a decision (Olson et al. 2015, Schlegel et al. 2015).

These studies do not support the other conclusion drawn from the earlier experiments: that people confirm having made a decision, when they did not make a decision. However, there are studies that support a very similar conclusion to that one: empirical research suggests that people sometimes confirm having made one decision, when they in fact have made another decision (Johansson et al. 2005, Hall et al. 2010, 2012).

2.6. Comparative Evidence

Finally, the ISA theory predicts if an animal is incapable of attributing an attitude to others, then it will be incapable of attributing it to itself; if it is capable of attributing it to others, then it will be capable of attributing it to itself. In particular, Carruthers argued that some non-human animals can attribute percepts and goals to others and to themselves, but no non-human animals are able to attribute false beliefs, whether to themselves or to others (Carruthers 2011: 254–259, 278–287).

The first of the three main challenges that he responded to concerned the purported capacity of some non-human animals to understand misleading appearances (Krachun et al. 2009). The suggestion was that in such cases the animal is contrasting an object with its appearance as such. In brief, his response was that the animal might be doing something entirely different. In particular, it might think that there are two objects or that there is only one object that undergoes magical transformations. He also noted that one could respond to this challenge by conceding that the animal really thinks about misleading appearances as such but still insisting that this is more primitive than thinking about false beliefs.

The second of the three main challenges that he responded to concerned the purported capacity of some non-human animals to monitor their own uncertainty (Couchman et al. 2009). The suggestion was that the animal thinks about its lack of knowledge as such. In brief, his response was that the animal might think about the choices presented to it as more or less likely to lead to success, which would generate different levels of anxiety and lead the animal to act accordingly; it would be a first-order decision making process rather than an instance of thinking about one’s own lack of knowledge.

The third of the three main challenges that he responded to concerned the purported capacity of some non-human animals to seek information while thinking about it as such (Kornell et al. 2007). Again, the suggestion was that the animal is thinking about knowledge as such. In brief, his response was that the animals might think in terms of a first-order question directed at the world, such as: ’Where is the food?’ or ‘What symbol will appear next?’; they do not need to think about knowledge as such.

An important development in this area concerns questioning attitudes, such as curiosity. Carruthers has proposed an account of questioning as a sui generis attitude, which takes a question rather than a proposition as its content and directly motivates one to act, in a similar way that fear directly motivates to act, without the need for the agent to represent fear as such (Carruthers 2018). The explanation of the developmental evidence proposed above and of the comparative evidence presented here probably hinges on the success of this new account.

Another important recent development concerns new evidence that great apes can pass certain false-belief tasks. Cristopher Krupenye and colleagues report that great apes were able to pass the anticipatory looking task (Krupenye et al. 2016). Buttelmann and colleagues report that great apes were able to pass the interactive helping task (Buttelmann et al. 2017). There is no evidence that great apes are able to attribute false-beliefs to themselves, so this presents a problem for the ISA theory.

In response to the findings by Krupenye et al. one could say that the anticipatory looking task is probably not a good measure of false-belief understanding. This is the kind of response given by Baillargeon et al. in the face of the replication crisis affecting this particular experimental paradigm as applied to human infants (Baillargeon et al. 2018). However, there is no similar response available for the interactive helping task. Here, two other responses are possible.

One response is to note that the core of the prediction about other animals is that metacognition should not evolve earlier than mindreading. The other part of the prediction, which says that metacognition should evolve no later than mindreading, is less central. There is some leeway here because one could argue that the repurposing of the mindreading faculty did not happen immediately. After all, the theory claims that there were fewer evolutionary pressures for learning mindreading than for learning metacognition.

Another response is to say that the fourth claim and the sixth prediction should be given relatively less weight (Carruthers 2011: 7). The claims and predictions about humans might be true and supported even if the claim and prediction about other animal species is false. This would deprive the theory of an evolutionary argument in its favour, but would not yet yield a very strong argument against it.

2.7. Conclusion

The ISA theory is supported by empirical evidence. At the same time, its main rivals are challenged by it. During the first decade since the theory was introduced its predictions have received significant further support, and none of the challenges have yet been such that it could not provide a reasonable response to them. Perhaps the most significant new source of support comes from studies suggesting that impairment in explicit metacognition is related to impairment in explicit mindreading, but not implicit metacognition, in autism.

Perhaps the most significant new challenge comes from studies suggesting that great apes only attribute false beliefs to others.

3. EXPLANATORY CONSIDERATIONS

In choosing a theory, one should also consider other virtues that a theory could have besides being supported by the empirical evidence. Theoretical virtues are such features of a theory by which it is rational to guide one's theory choice. They are generally relied upon as steady guides because throughout the history of science they have tended to reliably indicate if a research programme is progressing or degenerating (Lakatos 1970: 116; see also Newton-Smith 1981: 225). Although being empirically supported is the most important theoretical virtue, sometimes rival theories can possess it to a similar degree. Then one might turn to look for other theoretical virtues.

These other theoretical virtues are now widely agreed to include the following three: scientific fruitfulness, relative simplicity, and external coherence (Newton-Smith 1981: 223–232). In addition, at least in the debate on self-knowledge, another often invoked theoretical virtue is intuitive appeal (Carruthers 2011: 16). Since a theory might be buying one of them at the expense of another, these theoretical virtues should be considered in light of each other.

More specifically, the four virtues to be considered in this chapter can be defined as follows. A theory is scientifically fruitful if it makes predictions that contribute to new empirical research. A theory is relatively simple if it postulates fewer new entities than other theories with the same target and the same scope. A theory is externally coherent if surrounding theories that are already relatively well-established lend it support, in the sense that assuming these older theories are true makes it more likely that the new theory is true. A theory has intuitive appeal, is intuition-friendly, if it coheres with people's pre-reflective attitudes regarding the question that it aims to answer.

Originally, it was argued that the ISA theory has all of these theoretical virtues, except for the last one, and to a higher degree than its rivals. Even in the case of the last virtue, intuitive appeal, an explanation has been proposed why intuition should not be in its favour even if the theory is true. During the last decade, the ISA theory has been attacked on most of these frontiers. It is the task of this chapter to review these new explanatory considerations, and to argue that the foundations of the theory remain firm, even fortified.

3.1. Scientific Fruitfulness

Here are the reasons to think that the ISA theory possesses the theoretical virtue of scientific fruitfulness. First of all, it is now generally agreed to be a theoretical virtue. This is because it proves to be a reliable sign that a scientific theory is progressing rather than degenerating if, in addition to explaining the already available evidence, it also continues to make new predictions that provide a framework for new empirical research (Newton-Smith 1981: 223–232). Carruthers argued that the ISA theory makes such new predictions, and that it makes more of them than its rivals (Carruthers 2011: 370).

In particular, he argued that the framework for most of the empirical research on self-knowledge in the past was provided by predictions that were drawn from theories of self-knowledge similar to the ISA theory (Bem 1967, Nisbett & Wilson 1977, Gazzaniga 1998, Wegner 2002/2017, Wilson 2002). More generally, one could argue that predictions about the illusions people will have about their own minds were always the primary driving force behind empirical research on self-knowledge, they remain the primary driving force today (Wilson 2009, Vazire & Carlson 2010). As noted above, the ISA theory predicts misattribution, while most of its main rivals merely accommodate it.

He also argued that the ISA theory itself makes new predictions that could potentially contribute to new empirical research. First of all, it gives a clear new set of six main predictions that are specific to it (Carruthers 2011: 202). In contrast, as noted above, it is not always clear what some of its rivals should predict. As a matter of fact, the predictions are clearly spelled out by the authors themselves. Only the inner sense theory is an exception. Moreover, in addition to these six main predictions, the ISA theory explicitly makes many more specific predictions.[1] Some of these predictions went on to be tested in several experiments by Carruthers and colleagues, and some of the experiments were followed-up by independent researchers.

One area where the predictions of the ISA theory contributed to new empirical studies is empirical research on people's intuitions about self-knowledge. Carruthers predicted that a belief in 'transparency' (non-interpretive access) should be a human universal. As mentioned above, although false, this assumption was said to simplify processing with no significant loss of accuracy. This prediction led to a pilot study by Carruthers and Clark Barrett on the intuitions about self-knowledge in the Shuar of Ecuadorian Amazonia (reported in Carruthers 2008). It also led to five follow-up studies by Benjamin Kozuch and Shaun Nichols on intuitions about self-knowledge in people living in the United States of America (Kozuch & Nichols 2011).

Another area where the theory's predictions contributed to new empirical studies is empirical research on autism. As noted above, together with his experimentalist colleagues, Carruthers conducted three empirical studies to

[1] Examples of more specific predictions can be found in Carruthers 2011: 207, 217, 221, 235, 237, 262, 269, 274–276, 284–285, 296, 303–304, 309–310, 317, 322–323, 339, 341, 343, 354.

test the predictions that mindreading capacities are impaired in autism, that this impairment matches impairments in explicit metacognition, and that this impairment is unrelated to impairments in implicit metacognition (Williams et al. 2018, Nicholson et al. 2019).

All of this supports the claim that the ISA theory makes new predictions that contribute to new empirical research. A modest contribution to this is also made by the new predictions and suggested outlines of empirical studies that are presented at the end of each of the last three chapters below.

3.2. Relative Simplicity

Here are the reasons to think that the ISA theory possesses the theoretical virtue of relative simplicity. Note that although Carruthers originally argued that the ISA theory possesses this theoretical virtue to a higher degree than its rivals (Carruthers 2011: 6, 369), the claim has been challenged. It has been challenged by two recent developments in philosophy and cognitive science. One is the emergence of unified transparency theories of self-knowledge. Another such development is the relative establishment of a new framework in cognitive science, the predictive processing theory, which has been said to have a complicated relationship with the ISA theory.

Before considering the new challenges posed by recent developments, it is worthwhile to review the original argument that the ISA theory is the simplest of the available alternatives. Carruthers gave two main reasons to think that this is the case. One concerned a comparison between knowledge of one's own mind and knowledge of other minds. The other concerned a comparison between one's typical way of knowing one's own mind and the interpretive way of knowing one's own mind.

The first reason out of those which were originally offered as to why the ISA theory is simpler than its rivals is that it gives a unified explanation of one's knowledge of one's own and other minds. In fact, all versions of the symmetrical theory claim that one knows one's own attitudes and others' attitudes in the same way. In contrast, all versions of the asymmetrical theory claim that one also knows them in different ways. The relative complexity of the asymmetrical theory is perhaps most evident in the case of the inner sense theory since it goes further than the others in postulating a dedicated mental faculty. But it should be evident that, in this sense, the symmetrical theory provides a simpler explanation than any version of the asymmetrical theory.

The second reason of those that were originally taken to support the claim that ISA theory is simpler than its rivals is that it gives a unified account of standard self-knowledge and interpretive self-knowledge. All its contemporary rivals agree that one sometimes interprets one's own attitudes.

For instance, they agree that one might be led to interpret one’s own desires in such circumstances as a psychotherapy session. Since the ISA theory claims that the access is always interpretive, it does not need to postulate an additional means of access to one’s own attitudes in order to explain the interpretations. In contrast, since the asymmetrical theory claims that normally the access is not interpretive, it needs to postulate an additional means of access to one’s own attitudes in order to explain the interpretations. Therefore, the symmetrical theory also provides a simpler overall explanation of standard self-knowledge and interpretive self-knowledge.

These two reasons why the ISA theory is simpler than its rivals are now widely acknowledged by the theory’s opponents. For instance, Alex Byrne acknowledges the last point when he writes that: ‘all accounts of self-knowledge have to acknowledge a helping hand from Ryle’, that is—from some form of the symmetrical theory (Byrne 2018: 177). However, they question its relative simplicity on other grounds.

In particular, Byrne suggests that it provides a less unified account of knowledge of one’s own attitudes and sensations (Byrne 2012, 2018: 16). According to him, most ‘neo-Ryleans’, and perhaps even Ryle himself, would agree that knowledge of sensations is not always interpretive. As noted above, Carruthers is fairly explicit on this when he writes that access to one’s own sensory mental states might be more like recognition than interpretation. He even goes further to suggest that the process might be close to how the transparency theory describes how one comes to know one’s own attitudes (Carruthers 2011: 81). Byrne concludes that the ISA theory is in that sense a complex theory of self-knowledge.

Byrne himself defends a version of the transparency theory. As noted above, the transparency theory claims that one comes to know one’s own attitudes by attending to the relevant tracts of the outside world, as opposed to the mind itself, so the attitudes themselves are in that sense transparent (Evans 1982, Moran 2001, Fernández 2013, Byrne 2018, Schwengerer unpublished). It is widely agreed that earlier versions of the transparency theory were likewise complex in the sense that Byrne says the ISA theory is complex. For they only seemed to apply to knowledge of certain kinds of one’s own attitudes, such as beliefs, but not to other kinds of mental states.

However, unified versions of the transparency theory have now emerged, which account for knowledge of all kinds of one’s own mental states in the same way (Byrne 2018, Schwengerer unpublished). Byrne’s own updated version of the transparency theory claims that one infers conclusions about one’s own mental states from premises about corresponding tracts of the outside world. According to this theory, one normally comes to know that one believes that *p* by applying the inference rule 'If *p*, believe that you believe that *p'* (Byrne 2018: 102). Likewise, one comes to know that one feels a pain by applying the inference rule 'If you seem to (nociceptively) perceive a disturbance in your body, believe that you feel a pain '(Byrne 2018: 149). Crucially, his theory maintains that applying these rules of inference requires only ordinary reasoning capacities, not a mental faculty dedicated to self-knowledge. Byrne concludes that the new version of the transparency theory gives a more unified account of self-knowledge.

One thing that merits emphasising here is that a unified transparency theory still gives a less unified overall account of one's knowledge of one's own and other minds, and a less unified account of one's standard and interpretive self-knowledge. That is to say that the original reasons to think that the ISA theory is simpler in those respects would still stand even if Byrne's suggestion were also left to stand beside them. If Byrne is right, one then has to concede that the ISA theory is simpler in some respects while the transparency theory is simpler in another respect. However, it is unclear whether Byrne is right. There are at least two possible responses to his suggestion that a proponent of the ISA theory could make.

The first response is to say that the ISA theory is in fact compatible with the claim that all self-knowledge is interpretive. As it is, the four main claims of the ISA theory do not specify whether access to sensations is interpretive or not. They only say that access to most attitudes is interpretive. If one were to add a fifth claim that says access to one's own sensations is interpretive, then one would get a unified interpretive theory of self-knowledge. At some points, Quassim Cassam seems to suggest that the ISA theory should make this fifth claim (Cassam 2014: ch. 12). As noted above, at some points Carruthers himself seems to suggest that one should reject this fifth claim (Carruthers 2011: 81). At other points, he seems to suggest that one should remain neutral (Carruthers 2011: xi). Note that to say that the ISA theory is compatible with this fifth claim is not to say that it implies it. This seems to be one of the options that are open to the theorist.

Suppose one rejected the fifth claim and conceded Byrne's point about simplicity. Then one could still argue that the ISA theory is as simple as any theory of self-knowledge should be. One could argue that the transparency theory oversimplifies things. For the simplicity or complexity of a theory should reflect the simplicity or complexity of reality. One might argue that knowledge of one's own attitudes and sensations are really different in the relevant respect, and a theory of self-knowledge should reflect this.

For instance, one might suggest that self-attributions of attitudes and self-attributions of sensations differ in their reliability and the kinds of mistakes that they are susceptible to. This would not be an ad hoc assumption either, since a convincing case has already been made that one often misinterprets one's own attitudes (Carruthers 2011: 325–367). No similar case has been made regarding sensations. And the default view seems to be that there is a difference in reliability. Thus, one might argue that the transparency theory buys simplicity at the cost of empirical adequacy.

But perhaps the proponent of the ISA theory does not need to concede Byrne's point about simplicity even if they do reject the fifth claim. There is another way to respond.

The second response to Byrne's suggestion is to stress that none of the entities postulated by the ISA theory are new. Every entity that the ISA theory postulates is already postulated by surrounding theories that are already relatively well-established. In particular, the entities are postulated by theories of mindreading. Mindreading is a process in which sensory input is fed into a mental mechanism that processes that input according to the inference rules of an intuitive theory of mind and then produces beliefs about mental states as output. According to the ISA theory, the same process takes place when one attributes mental states to oneself.

Here, another note is due on where the process is meant to be exactly the same in cases of self and other and where the cases are supposed to be different. What 'essentially the same' means. In both cases, input is sensory. However, there are kinds of sensory input that are related used in the case of self. As an example, one's own proprioceptive sensation of pleasure is used primarily to attribute a mental state to oneself. Similarly, in both cases, processing rules are interpretive, they are rules of inference of one's intuitive theory of mind. However, different rules may be applied to processing information about different individuals. For instance, the same kind of sensory input might be processed more deeply if it is related to the self. Finally, in both cases, the outputs are all beliefs about mental states. However, these beliefs may be stored somewhat differently. We might have different 'mental files' or 'person models' for different people (Newen 2015). The crucial point to note is that this is something that one already assumes to exist when explaining knowledge of other minds.

In contrast, some of the entities postulated by the transparency theory are new. The transparency theory does not go as far as postulating an entire new mental faculty. However, Byrne's version of the theory does postulate a new set of processing rules such as 'If *p*, believe that you believe that *p*'. It postulates these rules for the sole purpose of explaining self-knowledge. Moreover, it claims that the process of applying these rules is relatively insulated from other mental processes. This makes it resemble the workings of a separate mental faculty, at least to some degree. The theory claims that the rules in question are applied unconsciously because if the process were made conscious then the rules would strike their user as irrational. The reason they would strike one as irrational is that, for instance, in the case of belief, the fact that *p* is generally not a good reason to believe that someone believes that *p*. For example, if in fact it is now snowing at the North pole, it is not be a good reason to believe that someone believes it. To conclude, the entities that the transparency theory postulates are new, even if they are less weighty than those postulated by other versions of the asymmetrical theory.

3.3. External Coherence

The theoretical virtue of relative simplicity is closely linked with the theoretical virtue of external coherence. In the above, it was argued that the ISA theory is simpler than the transparency theory because it only postulates entities that well-established theories of mindreading already postulate. In the same manner, it might be argued that better coherence with some other well-established theory gives an advantage to the transparency theory. In particular, it has been argued that the transparency theory fits the framework provided by the predictive processing theory better, so the transparency theory promises to be a part of a simpler overall account of the mind. The argument concedes the point that one should take surrounding theories into account and uses it against the ISA theory. The idea pursued in the following is that it fails to note an important link that makes the ISA theory fit the framework just as well.

Before moving to this challenge, which is posed by recent developments, it might be worthwhile to review the original argument for thinking that the ISA theory is externally coherent. The suggestion was that it got indirect support from three surrounding theories that were already relatively well-established at the time (Carruthers 2011: 47–68). These were the global workspace, working memory, and Machiavellian intelligence theories. Here is why the ISA theory seemed to receive support from them.

The global workspace theory claims that the mind consists of many specialised systems communicating through a central system, consciousness, by means of sensory information (Baars 1988). Since the ISA theory claims that the attribution of mental states is subserved by one such specialised system feeding on sensory information, it seems to cohere with the global workspace theory. The working memory theory claims that there is a kind of relatively short-term memory that allows one to simultaneously keep in mind different pieces of sensory information and to consciously operate on them (Baddeley & Hitch 1974). Since the ISA theory claims that the mental faculty charged with the attribution of mental states is largely dependant on manipulation of sensory information, it seems to cohere with the working memory theory. Finally, the Machiavellian intelligence theory claims that the adaptive challenge of living in a social group was a major driving force in the evolution of intelligence (Byrne & Whiten 1988). Since the ISA theory claims that a specialised cognitive system for understanding other minds evolved early and was only later repurposed for understanding one's own mind, it seems to cohere with the Machiavellian intelligence theory.

Crucially, there is no suggestion in either the global workspace theory or the working memory theory that a system charged with attributing mental states would have non-sensory access to its domain. Likewise, there is no suggestion in the Machiavellian intelligence theory of comparable pressures for evolving a specialised procedure for attributing mental states to oneself. These theories make it seem natural that one should have evolved a specialised cognitive system for understanding other minds that feeds on sensory input and might be repurposed for understanding one's own mind. Therefore, they provide indirect support for the ISA theory. It would still receive support from them, at least, even if it did not fit other frameworks.

In addition to this early argument in favour of the ISA theory, there have been early arguments against it that implied that it does not cohere with some of the more general surrounding theories. In particular, it has been suggested that it does not fit the dual-process framework. The dual-process theory claims that the human mind generally processes information in two different ways: intuitively and reflectively (Evans & Stanovich 2013). Keith Frankish and Joëlle Proust have both expressed worries about the ISA theory that are related to the dual-process framework. Proust argues that one knows one's own mind in a special way by means of intuitive processing, through what she calls 'meta-cognitive feelings' (Proust 2013: 293–307). Frankish argues that one knows one's own mind in a special way by means of reflective processing, through what he calls 'explicit belief' (Frankish 2016: 32). Proust might be taken to suggest that the ISA theory only explains reflective self-knowledge, while Frankish might be taken to suggest that it only explains intuitive self-knowledge.

However, at least some of the disagreement here seems terminological. In a recent response to Proust, Carruthers notes that he agrees with her that the feelings in question, such as the feeling of confidence, are directly accessible but not meta-representational. But he disagrees that these feelings should then be called ‘meta-cognitive’ (Carruthers 2017b). In another recent paper, he seems to agree with Frankish that the events in question, such as one’s saying to oneself in inner speech ‘Men and women are equal’, are directly accessible but do not constitute an attitude, such as a belief, on their own. They only do so in conjunction with other things that are not directly accessible, such as a commitment to what one says. Yet he seems to disagree, that the directly accessible event and those conjoined with it should then together be called a kind of attitude, an ‘explicit belief’ (Carruthers 2018, Frankish in conversation 2018). There might well be deeper disagreements lurking beneath these terminological ones, but on the face of it, the theory seems to cohere with most of what Proust and Frankish say about intuitive and reflective processes related to self-knowledge.

The idea to be pursued in the following is that perhaps one might explain away the apparent disagreement between Carruthers and the proponents of the predictive processing theory. The predictive processing theory claims that the mind's function is to reduce prediction error (Clark 2013). Since it is a very general theory, what it aims to explain inevitably overlaps with what the ISA theory aims to explain. If it turned out that the ISA theory is not readily compatible with the predictive processing theory, then one of two unwanted implications would seem to follow. Either one grafts the ISA theory onto the predictive processing framework after all by making additional assumptions significantly complicate the picture. Or one sees the theories as competitors and rejects one of them. The idea to be pursued in the following is that neither needs to be done because a third theory provides the link that makes the connection between the first two theories rather simple.

But first, here are the reasons for thinking that the transparency theory fits the predictive processing framework better (Schwengerer 2019). In a more traditional framework, one would define a piece of self-knowledge roughly as a reliably formed true belief about one's own mental states. Since the predictive processing theory substitutes talk of attitudes, such as beliefs, with talk of sub-personal predictions and error-correction, Schwengerer argues that it should define a piece of self-knowledge as a pattern of higher-level predictions accurately predicting a pattern of lower-level predictions. He suggests that this way of defining self-knowledge is in line with the transparency theory since the predictions are ultimately about the external world and the transparency theory stresses that this is what we attend to. He also suggests a prediction that might distinguish a theory of self-knowledge that is couched in terms of predictive processing from others. It should predict occasional surprise at the workings of one’s own mind. This would be explained which in terms of an error being registered in the higher levels of prediction. In response to this, there are at least two responses that a proponent of the ISA theory could make.

The first response to Schwengerer's suggestion is to note that the ISA theory also predicts occasional surprise at the workings of one's own mind. In fact, any existing version of the symmetrical theory should predict this, since they all claim that knowledge of attitudes is interpretive. Naturally, interpretation can lead to making an error, then to the realisation that one made it, and eventually to the surprise at discovering it. If this is the only prediction that is specific to theories of self-knowledge that are embedded in the predictive processing framework, then our theory fits this framework.

The second response to Schwengerer's suggestion is to say that the predictive processing theory comes out to be readily compatible with most theories of self-knowledge when one takes into account the supporting theories that the predictive processing theory must itself rely on. The need for such additional support is made evident by the famous Darkened Room problem (Clark 2016: 262–265). The problem can be stated roughly as follows: if the mind simply seeks to minimise prediction error, why does one not to stay forever in such especially predictable environments as an empty and silent darkened room? To solve this problem, the predictive processing theorist assumes that one evolved to have certain rigid prediction patterns, such as the one predicting that one will get food: one would not correct the prediction into a prediction that one will never get food even if it would be a more simple way to reduce prediction error. Crucially, on its own, the theory cannot say what rigid prediction patterns humans should evolve.

Therefore, for all one knows, the rigid predictions patterns that humans did evolve to have might turn out to correspond to what any of the theories of self-knowledge that are currently on offer need. In particular, the patterns might correspond to the mental architecture postulated by the ISA theory. In fact, the ISA theory already receives indirect support from a relatively well-established theory explaining the driving forces behind the evolution of intelligence. The Machiavellian intelligence theory provides the link that connects the ISA theory with the predictive processing theory.

3.4. Intuitive Appeal

Here are the considerations relevant to intuitive appeal. Carruthers argued that the assumption of the mind's transparency to itself is a human universal (Carruthers 2008, Carruthers 2011: 25–32). In support of this claim he provided three arguments. The first was that a review of claims about self-knowledge in the history of philosophy in the West and the East showed no examples of anything like the ISA theory and plenty examples of claims that the mind is transparent to itself. The second argument was that, in a pilot study conducted with the Shuar of the Amazonia, lay participants tended to say that it is impossible to be mistaken about one's own occurrent attitudes. The third argument was that an informal quantitative study of contemporary views on self-knowledge found that more than nine out of ten philosophers who work on it say that the mind is transparent to itself.

Of course, Carruthers does not think that intuition should be trusted in this case. However, he does subscribe to the opinion that is widely shared in the field that, other things being equal, it is rational to choose the theory that is more intuitive (Carruthers 2011: 16). One might argue the rationale for this is partly similar to the rationale for appealing to external coherence. If for no other reason, one might stick to what one is already inclined to think for pragmatical purposes. Accordingly, he needs an argument why it should not be considered an argument in favour of the rival theories if they preserve the transparency assumption.

First, he argued that everybody should accept that the transparency assumption is clearly too strong: people do make mistakes about their own attitudes. So intuition should not be trusted in this case. Second, he argued that people might have evolved the intuition not because it is true, but for other reasons. They evolved it because it is conductive to survival in spite of being false. More specifically, it simplifies the attribution of attitudes without making it less accurate. To see the point, suppose someone tells you what they think. It is much easier to just assume that the person knows what they think than to try to work it out by yourself. If people generally know what they think, as almost everybody in the self-knowledge debate agrees, then that the assumption is safe enough. The universality of the assumption and the explanation are questionable, however.

Regarding the universality, note that there is at least one expert on non-Western philosophy that disagrees with him. The expert in question is Jay Garfield, perhaps the foremost expert on the relevance of Buddhism to contemporary philosophy. He writes: 'Carruthers mistakenly claims <…> that all Buddhist traditions regard consciousness as transparent to itself' (Garfield 2015: 184). That does not yet amount to a counterargument, but it suggests that there is some room for doubting cultural universality.

Regarding the explanation, note that the ISA theory is at least compatible with the claim that people usually know their attitudes. One might argue that this assumption is enough to simplify attitude attribution. If it is enough, and if it is compatible with the theory, then it is not clear why one should find the theory counterintuitive. In that case, a proponent of the theory should either concede that it is counterintuitive and it is an argument against it, or explain why one should make the stronger assumption. Another response, which is developed in the following chapter, is to question whether the ISA theory is really counterintuitive.

3.5. Conclusion

The ISA theory has the theoretical virtues of being scientifically fruitful, relatively simple, and externally coherent. After the first decade, it can still claim to possess them to a higher degree than its rivals. It contributed to new empirical research on intuitions and autism. It postulates fewer new entities than the new unified transparency theory. It coheres with the new framework provided by the predictive processing theory. What is less clear, is whether there is an argument against the theory that it lacks intuitive appeal.

II. ELABORATIONS

4. SILENT INTERPRETERS

'…a silent or non-conscious partner to much else in the mind…'

O'Shaughnessy 2000: 106

The task of this chapter is to argue that the ISA theory receives support from empirical research on intuition. The first step is to show that having intuition on its side is an asset of a theory. The second is to argue that the ISA theory actually has intuition on its side. One idea pursued in the following is that laypeople are actually quite indecisive on this matter, as they probably are one many other deeply theoretical matters. However, insofar as they do have relevant intuitions, they favour the ISA theory. Another idea pursued in the following is that even some of the experts who go on to argue that the intuition is quite different do sometimes acknowledge the intuitive pull of the ISA theory indirectly. Finally, the task for this chapter is to suggest, at least in outline, a better test of what the intuition is, as well as a test of the motives that drive different intuitions about self-knowledge.

4.1. Pristine Attitudes

Before considering which of the rival theories of self-knowledge is more in line with intuition, it is worthwhile to examine its nature and role in this debate in more detail. It could be said that intuition is often treated as a 'pristine attitude' in the sense of being both original and unspoilt: the earliest attitude taken towards a state of affairs and also unmarred by development. If that is the nature of intuition, then it is clear why philosophers should want them to be in their theory's favour. However, the nature of intuition is more complicated than that. Although relying on them has something to be said for it, they are not altogether free from affliction. It can be defined as an attitude that a person feels justified in holding without having justifying reasons (see Mercier & Sperber 2017: 63–67). The feeling might mislead or it might not. Sometimes the attitude is ultimately justified, sometimes it is not.

If that is the case, then one might doubt whether calling them 'pristine' is appropriate. To answer this question, one should perhaps first define the attitude in a way that is more amenable to testing. One could define intuition as a pre-reflective attitude: an attitude one holds towards something without reflecting on it, at least without reflecting too much. To give an operational definition, an intuition is the attitude which is discerned in someone's quick answers, in their first impressions, or in laypersons' opinions.

Here is why one might expect intuitions to come to surface in those circumstances. If you think that a person's attitude might change after they reflect on your question, you can ask them to answer faster. If you think that the person already reflected too much on your question, you can ask them to go back to their first impression. Finally, if you think that the person already went too far in reflecting on your question to be able to see clearly the place where they started from, you can ask someone who is not used to thinking about it, a non-specialist. These are the ways to elicit intuition, and they can clearly be combined for the strongest effect. What this does not yet explain is why one should care about it when trying to answer a theoretical question.

Intuition is important in the debate about self-knowledge mainly for two reasons. The first is that they are part of what needs to be explained. For theorists of self-knowledge, the general question to be answered is this: how do people get reliably formed true beliefs about their own minds? Beliefs about self-knowledge count as beliefs about one's own mind. If they are also reliably formed and true, then they are part of self-knowledge. This would make beliefs about self-knowledge part of what needs to be explained by theories of self-knowledge.

One could say that intuitions are beliefs in their nascent form, which acquire their mature form after reflection, perhaps only in the theoretician's mind. If so, then intuitions about self-knowledge are, to borrow a phrase by Jérôme Dokic, 'seeds of self-knowledge' (Dokic 2012). To think that intuitions come earlier than any reasoning that might influence them is perhaps unwarranted, but they do at least come earlier in the sense that they come prior to (prolonged) explicit reflection.

Moreover, it is a fact about current philosophical discourse about self-knowledge that intuitions are considered to be important. Some researchers even define the desiderata for a theory of self-knowledge primarily by reference to intuitions (e.g., Bar-On 2004: 20, Bilgrami 2006: x, Fernández 2013: 38). In this debate, it is quite common for a theorist to proceed roughly as follows. One starts with a description of the intuitive view of self-knowledge. The claims made about it might be taken for granted or they might be justified by an analysis of what 'we think', 'would say', or how 'it seems'. It is commonplace to suggest that self-knowledge seems different from other-knowledge, and is privileged, peculiar, and authoritative. The theorist might then set themselves the task of providing a theory that explains how such knowledge is possible, preserving as many of those intuitive features as possible. This approach explains why the intuition is there by arguing that it is there because it is (mostly) accurate.

There are other reasons why the intuition might be there. Generally, people evolve to have intuitions that help them in the business of surviving and producing offspring. Accuracy is only one of the features that can make an attitude conductive to that end. Sometimes, intuitions are selected for in spite of their inaccuracy, as is the case with self-serving bias. For this reason, instead of explaining our intuitions, a theorist might as well set themselves the task of explaining them away. As noted above, Carruthers takes this path.

It is evident, then, that researchers of quite different ilk—those who try to explain intuitions about self-knowledge and those who try to explain them away—agree that in the debate on self-knowledge intuitions carry weight. They hold that there is a prima facie reason to stick to one's intuition and the burden of proof lies on the shoulders of those who suggest abandoning it. Yet one could grant that this is in fact the consensus and still question whether relying on intuition to any degree is acceptable in philosophy. Should we continue to rely on them, or should we strive to eradicate the practice?

Here are two reasons one could give for choosing the more intuitive when the theories are in other respects equal. The first reason is epistemic. One might think that intuition is generally reliable. Some researchers argue that this assumption is not warranted in philosophy (Machery 2017). The second reason is pragmatic. One might think that choosing the more intuitive view requires less effort. As defined above, the more intuitive view is the one that you already find oneself with. It obviously requires less effort to just plump for the attitude you already find yourself with than to try to change it. In any case, it is at least clear that having intuition on your side gives a dialectical advantage. That much can be assumed without going much deeper into the admittedly vexed question of what role intuitions should play in philosophy.

4.2. Dabblers' Indecision

Here are the reasons for thinking that the intuition of laypeople is in favour of the ISA theory. Of course, no theory is completely intuitive in all of its minute particulars. If one already believed in everything that it says, there would probably be little interest in hearing out the theorist. However, some might be more intuitive than others. It is this relative intuitiveness that is at question. The suggestion is that the ISA theory is relatively intuitive. As with the other theoretical virtues, the theory's intuitive appeal is primarily to be seen in light of other theories.

As noted above, people's intuitions about self-knowledge were probed experimentally for the first time in a pilot study by Carruthers and Barrett (reported in Carruthers 2008). The participants of this study were the Shuar of the Ecuadorian Amazonia. The rationale for the peculiar choice of participants was that the researchers aimed at testing the hypothesis that the transparency intuition is a human universal. If one assumes that the intuition undoubtedly exists in the United States of America, where the researchers themselves reside, then the choice of a more exotic culture seems a natural candidate for falsifying the hypothesis. It is probably less influenced by the Western philosophical tradition.

The researchers report that their participants were presented with two vignettes of the following form (Carruthers 2008: 48) [variations in brackets]:

> Suppose that Mary is sitting in the next room. She is just now deciding to go to the well for water, but [she/John] doesn't know that she is deciding to go to the well for water. Is that possible?

The researchers report that their participants had no difficulty with John being ignorant about Mary's decision but found the suggestion that Mary is ignorant about her own decision ‘well-nigh unintelligible ’(Carruthers 2008: 48).

Carruthers interprets these results, together with a wealth of culturally diverse textual evidence about scholars' opinions (Carruthers 2011: 25–32), as suggesting that the transparency assumption is a human universal. In particular, he claims people intuitively follow these two particular rules:

> (1) If one thinks one is in mental state M → one is in mental state M;
>
> (2) If one thinks one isn't in mental state M → one isn't in mental state M.

These two rules constitute ‘the transparency assumption’ (Carruthers 2011: 12). He claims that these rules are either innate, or learned and habitual.

The only follow up on this study that has ever been published is a paper by Benjamin Kozuch and Shaun Nichols in which they report a series of five experiments (Kozuch & Nichols 2011). Their participants were of less exotic descent: in most of the studies, they are reported to be undergraduates from the University of Arizona, and in others they were workers employed on the online platform Amazon Mechanical Turk. In the first experiment, their participants received several vignettes of the following form (Kozuch & Nichols 2011: 10-11) [variations in brackets]:

> John is just now deciding to go outside, and even though he's paying close attention to his thoughts and feelings, he doesn't know [that/why] he is [deciding/thinking/feeling happy/feeling an urge] to go outside.
>
> It's possible that this really could happen.
>
> 1 (strongly disagree) – 7 (strongly agree)

The main finding is this: participants tended to say that the agent can be ignorant of their current decisions, thoughts, feelings, and urges, and especially of the causes of these mental states (for states: *M* = 4.09, *SD* = 2.0; for causes: *M* = 5.34, *SD* = 0.99; the difference between the two being significant: *t* (31) = -3.147, *p* < 0.01, two-tailed). In the second experiment, participants again received two vignettes, which were of the following form (Kozuch & Nichols 2011: 12) [variations in brackets]:

> When I am making a decision about what to do (for example, deciding whether to go swimming), if I pay attention to my thought processes, I can usually see what leads me to [make the decision/feel the urge] I do.

The main finding was this: participants tended to say that they can usually see what leads to their own decisions or urges, but more so for decisions (for decisions: *M* = 5.47, *SD* = 1.36; for urges: *M* = 4.69, *SD* = 1.35; the difference between the two being significant: *t* (98) = 2.883, *p* < 0.01).

In the third experiment, participants first read re-descriptions of actual experiments. In one of the described experiments, people who gulped a placebo-pill 'producing 'typical symptoms of electric shock—such as palpitations, irregular breathing and 'butterflies' in the stomach—endured ever stronger electric shocks for longer, presumably because they attributed some of their physical symptoms to the effects of the pill (Nisbett & Schachter 1966). In the other of the described experiments, people who memorised the word pair 'ocean-moon' (among other word pairs) were more likely to say 'Tide' when asked to name a laundry detergent, presumably because of the association in their minds between oceans, the moon, and tides (Nisbett & Wilson 1977).

In this third experiment by Kozuch and Nichols, the participants were then asked if people in the Placebo-Pill experiment would have been aware that they attributed some of their physical symptoms to the pill and whether the people in the Tide experiment would have been aware of the influence of the memorised word pair on their choice to name that particular laundry detergent. Participants said they would have been aware of the attribution of physical symptoms, but not of the effect of the association of words (for symptoms: *M* = 5.23, *SD* = 0.973; for associations: *M* = 3.19, *SD* = 1.52; the difference between the two being significant: *t* (43) = 5.59, *p* < 0.01) (which is wrong: people in the Placebo-Pill experiment, when interviewed, said they did not attribute any of their physical symptoms to the effects of the pill).

The last two experiments by Kozuch and Nichols were similar to the third and only differed in that they described different past experiments. In the fourth, two such experiments were described. In one of them, people anonymously chose to give more of their money, which they got by chance, to a player who did not get any money if an eye was drawn at the top of the computer that they were using to allocate the funds (Haley & Fessler 2005; slightly modified) (incidentally, the results of the actual experiment have replication issues, see Matsugasaki et al. 2015). The difference in the other described experiment was that it was said that a circle was drawn around the camera at the top of the computer, instead of a picture of an eye. Participants tended to say that people would have been aware of the effect of the camera, but not of the drawing of the eye (for the camera: $M = 3.91$, $SD = 1.31$; for the eye: $M = 2.71$, $SD = 1.64$; the difference between the two being significant: t (95.139) = 4.026, $p < 0.001$; note that, for experiments 4-5, they used a six-point scale).

In the fifth experiment, participants read that a person's movements were predicted from signals in the brain half a second before the person became aware of their own intention to move (Libet et al. 1985). For one group of participants, the movement was described as spontaneous, while for the other group it was described as a reaction of withdrawal when the person in the experiment was presented with a picture of a spider. Participants tended to say that they would have been aware of their decision to move earlier than it could be seen from the brain signal, but only in the spontaneous movement condition, not in the withdrawal-reaction condition (for spontaneous: 68%; for reflexive: 19%; the difference between the two being significant: χ^2 (1, N = 65) = 13.931, $p < 0.001$).

Here is what the studies by Kozuch and Nichols, taken as a whole, suggest about laypeople's intuitions regarding self-knowledge (or, at least, American undergraduates' intuitions about self-knowledge). The results suggest that people think it is possible to be unaware of one's current decision, thought, feeling, or urge, and even more so of their causes. This is the opposite of what Carruthers found. They think people are usually aware of the causes of their decisions and urges but especially of the causes of their decisions. They think that people are aware of their reasons for thinking that they feel as they do but not of their associations or how they influence their behaviour. They think people are aware of causes for one's action that are readily interpreted as rational, such as being more generous when possibly being filmed by a video camera, but not irrational, such as being more generous when one sees a drawing of an eye. They think that people are aware of their decision before brain signals predict it in case of spontaneous movement but not reactive movement.

Kozuch and Nichols interpret these results as indicating that people do not assume transparent access without restriction (Kozuch & Nichols 2011: 24). However, they also say that people still overestimate the amount of transparent access they do have, particularly concerning decisions. One might wonder what makes decisions special. Here is a suggestion: people think they have special access to decisions because decisions are supposed to be responsive to reasons. If people are thinking along these lines, then they should also think that they have more transparent access to other reasons-responsive attitudes as well, such as judgements.

What these results imply for theories of self-knowledge is not that clear, however. As noted above, Carruthers interprets his own results as indicating that the transparency assumption is intuitive. However, even if this were the right interpretation of his own results, there is the obvious problem that his results clash with those from the first experiment by Kozuch and Nichols. Their findings were directly opposite. Moreover, the latter study had a sample of participants who Carruthers should predict will have the strongest transparency intuition. This makes the results all the more surprising, from his point of view. Kozuch and Nichols also interpret their results as implying that some laypeople's intuition is more in line with some version of the asymmetrical theory. There are several problems with the proposed interpretations.

First, every theory of self-knowledge which is nowadays defended in the literature allows that one sometimes makes mistakes about one's own mind and about one's own attitudes. Therefore, if people were to say that it is impossible to be mistaken about one's own attitudes, that is, if they turned out to be intuitive infallibilists about self-knowledge of attitudes, then this would mean that each and every one of the currently competing theories of self-knowledge is counterintuitive. However, the results from the experiment by Carruthers and Barrett and the first experiment by Kozuch and Nichols do not force this conclusion upon us. This is because many of the participants said that mistakes in the domain of self-knowledge are in fact possible. In fact, the majority did so in the experiment by Kozuch and Nichols. The results from their study actually speak against the intuitiveness of infallibilist theories of self-knowledge and in favour of fallibilist theories of self-knowledge. However, it means: in favour of every theory of self-knowledge currently on the market.

Second, every theory of self-knowledge which is nowadays defended in the literature allows that you usually know your own attitudes and even their causes. If they denied reliable access, then they would be denying self-knowledge. That is, of course, an option. However, Schwitzgebel is probably the only theorist in the field that gravitates towards it (Schwitzgebel 2011). The results would speak against the intuitiveness of that kind of theory, and he would likely be the first to admit that it is counterintuitive. The second experiment by Kozuch and Nichols probed only whether people think they usually have access to their own attitudes. If people form beliefs about their own attitudes in a reliable way, that is, if they have self-knowledge of attitudes, then trivially they usually have access to their own attitudes. The question they were asked was about reliability and not about the mechanism that underlies that reliability. However, the latter and not the former is the central point of contention among contemporary theories of self-knowledge.

What distinguishes current theories of self-knowledge is not that some allow mistakes and others do not, and not that some allow that people usually know their own attitudes and others do not, but that they tell different stories about the way people are led to those mistakes or to that knowledge. None of the experiments above probed lay intuitions about the process by which one acquires self-knowledge or remains self-ignorant, although that is where the theories differ (even the constitutive theory might be read as offering an answer to this question, although it is negative: there is no intervening process). One should probe intuitions about the process and not the result. However, for that very reason the prospects of discerning people's intuitions about self-knowledge might look dim. They are unlikely to be very decisive on such matters as particular mechanisms of belief formation. Believing in such mechanisms is a strong sort of theoretical commitment.

In any case, the results from the experiments conducted on lay intuitions so far do not show that the ISA theory is any less intuitive than its rivals. On the contrary, from the results by Kozuch and Nichols one can glean some support for thinking it is more intuitive than the others. Admittedly, the argument here is rather impressionistic. Still, one could note that the ISA theory is primarily aimed at explaining the pattern of mistakes that people make about their own attitudes. These experiments indicate that people are quite willing to admit that such mistakes happen. If that is so, then one might venture to predict that they would also readily understand the motivation behind the theory.

4.3. Experts' Lapse

Here are the reasons for thinking that some experts' intuition is also in line with the ISA theory. One way of finding out whether it has any intuitive appeal in a population of experts is to look at how many of them end up endorsing it after reflection. The assumption here is that, other things being equal, researchers end up endorsing the more intuitive theory. Another way of finding out what the experts intuitively think is to look at what they say about their own intuition and about other people's intuition concerning self-knowledge. Sometimes, they straddle the line between what is intuitive to them and what is intuitive to everyone without argument, perhaps taking themselves to be representative examples. Experimental philosophy has shown time and again that this assumption is problematic, and the case of self-knowledge, as already indicated above, is no exception. Finally, one might also question philosophers' explicit descriptions of what they find intuitive based on what they themselves say elsewhere, that is, by looking at the internal consistency of their expressed views.

There is no denying that the asymmetrical theory is widely endorsed among contemporary philosophers. To argue the point, Carruthers has conducted an informal experiment (Carruthers 2011: 17–18). He took all the articles on self-knowledge published after 1970 from the largest database of philosophical publications, *PhilPapers* (N = 334). He then removed the articles that concerned irrelevant topics, such as self-identity or knowledge of sensations rather than attitudes. Finally, he classified each author into one of two categories. In the first category, he put together all the scholars who assumed any of the following: authority regarding one's own attitudes, a principled contrast between self-knowledge and other-knowledge of attitudes, a self-presenting character for attitudes (meaning that if one has them, one believes that one has them), or a high degree of certainty regarding self-knowledge of attitudes. In the second category, he put all the scholars who assumed neither of those things. What he found was that the overwhelming majority of authors turned out to belong to the first category, making those assumptions, while only a few were opposed, undecided, or could not be classified (75 out of 80 (or 94%) were in the first category).

If the survey were conducted today, there would certainly be a larger number of dissenters. One can observe this by simply counting the authors in the list of those who more or less tentatively endorse the ISA theory, which was given above, in the introduction. However, the list of authors who more or less explicitly criticise it is obviously longer and it would certainly get a lot longer if one included those who assume a version of the asymmetrical theory but do not, in an explicit way, respond to the ISA theory. So perhaps, if professional philosophers were now questioned on the topic, most of them would still endorse the asymmetrical theory. If philosophers specialising in the philosophy of cognitive science or cognitive scientists more broadly were surveyed, then perhaps the tendency would be less pronounced, or even opposite.

As it happens, the only large scale survey of professional philosophers' beliefs regarding core issues in philosophy that has been published to date, the one by David Bourget and David Chalmers (Bourget & Chalmers 2014), did not include any questions about self-knowledge. Perhaps this will be rectified in the future. All in all, a cursory glance at the literature on self-knowledge from the last decade or so will likely suffice to give one the strong impression that the ISA theorist remains a rare animal.

As noted above, philosophers who endorse the asymmetrical theory tend to take their view to be intuitive, at the very least to themselves. Likewise, they often take it that the way they themselves describe the acquisition of self-knowledge is also the way that it intuitively seems to people in general that one acquires self-knowledge. And it is customary to take it to be an asset of the theory.

In order to see why all of this is problematic, it will be worthwhile to examine in more detail a particular kind of dissenter. These happen to be widely esteemed experts on self-knowledge and on describing one's own inner experience. They claim, contrary to the official view of the multitude, that the acquisition of self-knowledge is 'silent'. By this they mean that acquiring self-knowledge, from the perspective of the acquirer, does not intuitively seem like anything at all. You 'just know'. There are at least two theorists who explicitly say that acquiring self-knowledge is like that : Brian O'Shaughnessy (O'Shaughnessy 2000) and Johannes Roessler (Roessler 2013). The slight differences between their respective treatments of the topic are illuminating.

O'Shaughnessy says that the acquisition of self-knowledge: 'must arise otherwise than in the mode of experience, it must so to say be a *silent or non-conscious partner* to much else in the mind, and leave no residue in event-memory' (O'Shaughnessy 2000: 106; emphasis in the original). This means that the process of acquiring self-knowledge, as opposed to the result, leaves no mark in consciousness and in that sense is not an experience at all. It is important to note here that the process and not the result is precisely what the theorists of self-knowledge are debating. If O'Shaughnessy is right, then any theory which says that one acquires self-knowledge by way of an unconscious process is true to one's conscious experience. Since the ISA theorist says just this, one should conclude it is true to conscious experience.

What is more, if he is right, any theory that says there is something more to be gleaned from the inner experience of acquiring self-knowledge is untrue to that experience. The ISA theory does not say that the acquisition of self-knowledge is experienced as interpreting. More precisely, it does not say that this is normally so. One usually interprets oneself unconsciously. Moreover, even O'Shaughnessy allows that the process is experienced as interpreting sometimes. For instance, he says that this happens when one takes seriously a friend's suggestion that one's current motives are not what one might have supposed (O'Shaughnessy 2000: 105).

On this account, typically one either fails to notice how one arrived at self-knowledge, or one experiences it in something like the following way: one hears the question, one stays silent for a moment, and then an answer pops into one's head. For what it is worth, I myself subscribe to this description of my inner experience and hold that this is the one that is the most accurate. Why are the multitude of researchers who study self-knowledge not with us on this point?

Roessler's case is instructive in this respect. First, it is notable that he begins his article 'The Silence of Self-Knowledge' by acknowledging that some researchers who are widely regarded as experts in describing inner experience (O'Shaughnessy, in particular) say that acquiring self-knowledge does not seem like anything at all from the subject's point of view (Roessler 2013: 1). He also notes that other researchers disagree with this and contend that it does seem like something, namely: like going from believing that *p* to believing that you believe that *p* (i.e., as the transparency theory describes the process). As noted above, researchers of the second kind follow the famous suggestion by Evans that the process by which you come to believe that you believe there will be a third world war is the same as the process by which you come to believe that there will be a third world war; that is, you look at the geopolitical situation, etc. (Evans 1983; see also Moran 2001, Fernández 2013, Byrne 2018, Schwengerer unpublished).

Roessler says that the suggested inference—'*p*, therefore I believe that *p*'—would strike one as irrational if one were aware of it. He says that this is because something's being the case by no means logically implies that it is believed to be the case. For that reason, Roessler begins by taking the side of O'Shaughnessy in this debate. He quotes the latter as saying that there is no 'cognitive path via which this knowledge is reached' (Roessler 2013: 2). However, then he goes on to argue that getting at one's higher order belief does seem like something from the first-person point of view: like making explicit what was implicit. On Roessler's considered view, believing that *p* and believing that you believe that *p* are aspects of the same mental state. The idea is that from the first person point of view getting at what you believe about our own mind seems not like going from one mental state to another but like explicating the same mental state. However, if that were so, then the process would not be entirely silent. What seems to be happening here is that the theorist briefly acknowledges the intuitive appeal of the view that the process is silent, but then leaves this austere conception in favour of a more elaborate one.

Roessler also notes that even if laypeople do have some stake in this debate 'the finer points of an explanation of self-knowledge will be reserved to philosophers' (Roessler 2013: 1). This is probably meant to include the finer points of the conscious experience of getting at one's higher-order beliefs as well, thus the reference to esteemed experts on describing it. However, at some points, he seems to suggest that his own story, the one where you experience believing that you believe that *p* as an aspect of believing that *p*, or as something that has been implicit and was then made explicit, is the intuitive story. Yet it seems likely that people who merely dabble in these issues should be just as indecisive when asked about the structure and explication of their beliefs as they are about the process of forming those beliefs. Of course, this is something that should be addressed empirically.

Whatever the results of such future studies might be, it is clear that the currently available evidence lends no support to Roessler's view. There is even some evidence to challenge it. For his view is a version of the constitutive theory of self-knowledge, which claims that to be in a mental state and to believe that one is in that mental state are, in some sense, aspects of the same mental state (see also Shoemaker 1994, Boyle 2009, Coliva 2016). It is well-known that the theory finds it more difficult to explain how mistakes in this domain are possible, let alone to explain the pattern of those mistakes. If knowledge of a belief is a constitutive part of that belief, then how can one be wrong about that belief? In contrast, the evidence discussed above suggests that laypeople intuitively think that one does make such mistakes.

The view that one acquires self-knowledge silently seems to be at least implicitly acknowledged by many more than the two philosophers discussed above. As noted above, it is common to introduce the topic by suggesting that self-knowledge is different from other-knowledge. The most obvious difference is supposed to be that the latter is gained by means of interpreting that person's behaviour, whereas the former is not, it is gained in some other way. This other way is often described as 'direct' or 'immediate'. These expressions suggest that the silent nature of self-knowledge is acknowledged more widely.

If silence is the only thing there is to the normal conscious experience of acquiring self-knowledge, then any theory that says the acquisition is usually an unconscious process is in line with the intuition. Moreover, if that is true, then any theory that says there is something more to it, is not in line with the intuition. That is arguing that self-knowledge is loud, not silent. Since the ISA theory is clearly of the first variety, it is in line with the intuition. And since it is experts who have argued that self-knowledge is silent, the ISA theory is in line with the intuition of at least some of the experts.

4.4. Sceptics' Leniency

Here are some suggestions for further empirical research on intuitions about self-knowledge. The idea is that one might learn from empirical studies on intuitions about free will and moral responsibility and adopt similar tools. The first step is to see what these empirical studies have to teach. They are the invoked here partly for the reason that research on intuitions about free will is probably the most developed sub-field of empirical research on philosophical intuitions. Also, because intuitions about free will and moral responsibility are, most likely, what motivates intuitions about self-knowledge.

Let us start with the underlying motivation. Setting aside the desire to support a particular theory of self-knowledge, why would one be motivated to believe that the mind is transparent or that the mind is opaque? The history of philosophy and, in particular, the history of the debate on free will offer some suggestions. Some are also to be found in the few empirical studies that have already been done. These include empirical studies on laypeople's intuitions about self-knowledge, which were discussed above, as well as a study on the beliefs of professional philosophers concerning closely related issues. An empirical study that would address the above question directly has not yet been done, but an outline will be proposed at the end of this section.

First, note that there some important analogies and dis-analogies between the debate about self-knowledge and the debate about free will. The debate about self-knowledge is historically related to the one about the possibility of knowledge that, in turn, is related to the debate about the possible absence of knowledge, i.e., to the one about scepticism. In the debate about scepticism, there is an apparently widely endorsed hypothesis that what motivates the sceptic to hold on to their view, apart from rational argument, is the desire to deny moral responsibility.

For instance, Kant seems to hint in the direction of this hypothesis, when he chooses the phrase 'Dare to know' as the moto for the whole epoch of enlightenment (Kant 1784/1996). The phrase does not necessarily imply that the sceptic denies knowledge for the sake of denying responsibility. Perhaps it merely implies that it takes courage not to be a sceptic and not to suspend belief. A more recent example is the mildly menacing *Fear of Knowledge*, which the renowned anti-sceptic Paul Boghossian chose as the title for his book (Boghossian 2006). Again, on its own, the title does not say anything about the sceptic's reasons for fearing knowledge. However, saying that responsibility might be implied does not look to be far off from the target.

The most recent and the least equivocal example comes from last year's 'New Enlightenment' lecture given by Gloria Schönbaumsfeld at the University of Edinburgh (see also Schönbaumsfeld 2016). She started the lecture by arguing that the sceptic who doubts the existence of the external world has no sound arguments for their scepticism. She then raised the question what drives the sceptic to hold on to their scepticism in spite of the lack of sound argument. Drawing on Søren Kierkegaard, she suggested that the sceptic is driven by the desire to deny epistemic as well as moral responsibility. The reasoning behind this is quite simple: if you did not know, how could you be responsible? The suggestion here is that Schönbaumsfeld makes explicit what was implicit in Kant and Boghossian: the sceptic is supposed to deny knowledge in order to deny moral responsibility.

One could understand this as an empirical hypothesis about the relation between people's desires and their expressed beliefs. In particular, one could take it as a hypothesis about professional philosophers. The hypothesis would be that sceptics about the external world, or sceptics more generally, are more likely to have less robust beliefs about moral responsibility.

There is some evidence that indirectly supports this hypothesis. In the survey by Bourget and Chalmers, which probed the beliefs of thousands of professional philosophers, one of the findings was that empiricism is associated with moral anti-realism and, more generally, that less robust beliefs about morality tend to fall together with less robust beliefs about knowledge, such as: disbelief in a priori knowledge, disbelief in analytic truth, etc. (Bourget & Chalmers 2014; see also Figure 2, by Andrew Higgins, reproduced below). Less robust views about moral responsibility and knowledge were also associated with less robust views about free will, such as hard determinism ('free will scepticism') and compatibilism. For this reason, it is worthwhile to consider how the debate about self-knowledge compares with the debate about free will.

For a long time, the philosophical debate about free will had little input from empirical research. This changed, to a large extent, after experiments by the neurologist Benjamin Libet and his colleagues in the 1980s. Although these experiments did concern people's conscious experiences related to willing, they were not exactly experiments on people's intuitions about free will. At the same time, it was and still is quite common in the philosophical literature on free will to implicitly or explicitly appeal particularly to people's intuitions about free will. For example, philosophers appeal to how people would intuitively react to thought experiments (for two collections of examples of such appeals, see Nichols 2004 and Nahmias et al. 2005).

Many philosophers in the debate had assumed that laypeople are libertarians: that people think that they have free will in a sense that is incompatible with determinism. In this literature, it has also been suggested that if it turned out that there is no libertarian free will, as the Libet experiment was sometimes interpreted to have shown, then there would be important negative consequences: it would affect people's belief in morality and their actual moral behaviour. Some philosophers even suggested that if it turned out to be the case that there is no libertarian free will, then it would be better for the layperson not to know (Smilansky 2002).

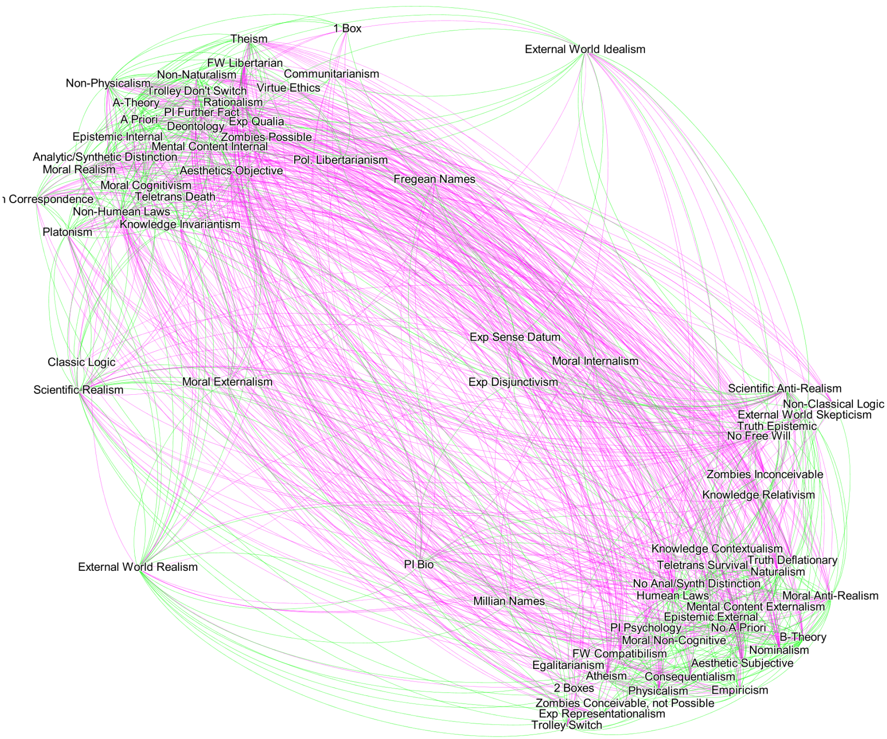

Figure 2: PhilSurvey Correlations: a visualisation by Andrew Higgins (please find a larger picture at: https://sites.google.com/site/aahiggi/home/pictures-of-philosophy).

However, since empirical research into the belief in free will and closely related beliefs, such as belief in determinism, started in earnest, it became apparent that people might well be compatibilists. At the very least the assumption that they are intuitive libertarians became highly questionable (see Nahmias 2018 for a recent review). It also turned out that belief in free will is largely driven by the desire to hold people morally responsible and to punish them for their misdeeds and to justify retributive punishment (Shariff et al. 2014, Clark et al. 2014, 2015, 2017, 2018). Very roughly, the history of this other philosophical debate now looks as follows: philosophers suggested that robust views about free will are intuitive, experimentalists found that this is questionable and, moreover, the people who do have such robust intuitions tend to be less lenient.

There is as yet no similar influx of empirical evidence in the debate on self-knowledge. While empirical research on belief in free will is now a burgeoning industry with literally hundreds of empirical studies already conducted, there are only six empirical studies on self-knowledge, discussed above, and none of them directly probes the relation between intuitions about self-knowledge and intuitions about free will, moral responsibility, or punishment. Research on free will belief is also far ahead methodologically. In addition to dozens of vignettes, there are more than ten questionnaires, and a wide variety of experimental manipulations. The measures evolved from asking for simple questions that did not really pit rival theories of free will against one another, to more complex ones that come much closer to it.

Here are some lessons one could learn from the free will debate. First, one could model the tools for measuring self-knowledge intuitions on the tools used for measuring free will intuitions. In particular, a questionnaire that could actually distinguish between rival theories of self-knowledge would be preferable to the simple questions that were asked. As argued above, it is crucial that the questions should probe intuitions on the way self-knowledge is acquired and not merely intuitions about the result, that is, whether we usually know our attitudes or not. This is not what the debate is about and it does not help distinguish between competing theories. As an example of what the questions could be about, one could ask whether certain factors are sufficient to mislead a person about their own attitudes. Following the example of Kozuch and Nichols, one could describe to the participants the studies that pit rival theories of self-knowledge against one another (e.g., the Olson et al. studies) and ask them to predict the result.

Second, one could compare the results obtained by measuring self-knowledge intuition with the results obtained by using other, more well-established measures. This would help us tell whether people actually have any intuitions about self-knowledge that possess any independent predictive power. It would be especially important to test the relation between self-knowledge intuitions and intuitions about free will and moral responsibility since we already have reasons to think that they are related. It might turn out that laypeople are so indecisive about these questions that talk about a distinct intuition is not warranted here. Likewise, it might turn out that in professional philosophers' intuitions, the variance is mostly explained by some more fundamental intuition.

In particular, judging from what we known from the debate about free will, one could already make some particular predictions. For example, one could predict that more robust beliefs about self-knowledge—a stronger belief in transparency—will be positively related to more robust beliefs about moral responsibility and more robust beliefs about punishment, and especially retributive punishment.

There is already some indirect support for this hypothesis in the studies conducted by Kozuch and Nichols. As noted above, they suggest that people think that reasons-responsive attitudes, like decisions, are more transparent than attitudes that are not reasons-responsive, like urges. Reasons-responsive attitudes are precisely the ones for which one will be held directly morally responsible and which are required for one's action to be considered an exercise of free will, at least as a minimal condition on most accounts of free will and moral responsibility. This is why, by analogy, one should predict that the desire to hold others morally responsible and punish transgressors will also be associated with a stronger belief in transparency.

Here is a way how one could test this hypothesis in an empirical study. First, one could give one group of participants a blameworthiness prime. For example, one could ask them to read a vignette about morally blameworthy behaviour, such as murder, while asking the other group to read a vignette about otherwise similar but morally irrelevant behaviour. Then, one could give both groups the same vignettes that Kozuch and Nichols gave to their participants and which measured the participants' belief in transparency (although developing a self-knowledge questionnaire would be preferable). After this, one could give both groups a manipulation check, for example, in the form of a questionnaire like the ones given in free will belief studies which elicit beliefs about free will, moral responsibility, and retributive punishment (we could use the questionnaires developed by Nadelhoffer et al. 2014 or Clark et al. 2015, or a combination of the two).

Of course, there are some legitimate doubts about the strength of priming effects in such contexts, so one might end up leaving the causal connection study and merely looking at correlations. In any case, the prediction is that, on the one hand, the group that would be given a blameworthiness prime will tend to think that people are more aware of their attitudes and will score higher on measures of belief in free will (especially libertarian free will), belief in moral responsibility, and belief in punishment (especially retributive punishment). On the other hand, those who are more sceptical about free will and moral responsibility should tend to be more sceptical about self-knowledge (should be more inclined to accept the ISA theory), and should also be more lenient when asked about punishment (especially retributive punishment).

4.5. Conclusion

The ISA theory receives support from empirical research on intuition. Other things being equal, one should choose a theory that is more intuitive, either for epistemic reasons, or for pragmatic reasons, or for dialectical reasons. The ISA theory is more intuitive than its rivals. It coheres with laypeople's intuition that one usually knows one's own attitudes but sometimes makes mistakes about them. It also coheres with many experts' intuition that one usually comes to know them 'silently', without noticing how it happens, but sometimes interprets oneself consciously. This is what one should conclude from the available empirical research. The available evidence is relatively scarce, however, and one should further investigate self-knowledge beliefs, perhaps taking the lead from empirical research on belief in free will.

5. MISLED MEDITATORS

> ‘But after having spent several years studying the book of the world and trying to acquire some experience of life, I took the decision one day to look into myself and to use all my mental powers to choose the paths I should follow.’
>
> Descartes 1637/2006: 11

In *Discourse on Method* and later in *Meditations on First Philosophy*, Descartes contrasts two broad paths to knowledge: that of the actor and that of the meditator. As he describes it, the path of the actor is the path of someone who travels the world, visits courts and armies, mixes with people of different character and rank, accumulates different experiences, puts himself to test in situations in which he finds himself by chance, and at all times gives reflection to things as they present themselves to him so as to derive some benefit from them. Whereas, as he describes it, the path of the meditator is the path of someone who leads as solitary and retiring a life as he would in the most remote of deserts, even if not lacking any of the comforts of the most populous cities. Descartes was of the opinion that the meditator has an advantage over the actor insofar as the ability to see one’s own mind clearly and distinctly is concerned and consequently insofar as one concerns oneself with the establishment of safe foundations for all knowledge.

The ISA theorist comes to derail this line of thought. For they insist that access to attitudes as well as their control remain as indirect and unsafe for the meditator as they are for the actor. This theorist can support their line of reasoning by noting the results of recent empirical research on meditation. They can draw support for their view from this recent literature, because it indicates that meditators are often misled about their own attitudes, just like other people are misled about them: they are not secluded from delusion in solitude. Moreover, the theorist is sooner emboldened than discouraged by the finding that practicing meditators are almost invariably struck to find out how daunting a task it is to control their own mind, even when one is left alone with one’s thoughts: some compare it with being left alone with a wild animal that one now has to make stand still. Although the theorist has no stake in denying that these experiences are instructive, they will be in a position to insist that, insofar as they are instructive, they primarily teach the lesson of the opacity of mind.

5.1. Minimal Interference

Meditators come in many varieties, and not all of them are equally relevant for the present discussion. For the purposes of this discussion, a meditator is someone who has only a minimal amount of behavioural or contextual evidence that could guide them in their interpretation of their own attitudes. Meditators are neither themselves engaged in any kind of overt behaviour that could suggest to them or anyone else that they possess a certain attitude, nor do they perceive events in the outside world or features of their own environment that could suggest either to themselves or to an outside observer that a certain attitude presently occupies their mind.

An example would be someone who sits comfortably in a quiet room, eyes closed, with no particular agenda other than to observe the goings-on of their own conscious mind. Descartes himself fits the description, sitting comfortably in his armchair in front of the fireplace in his quiet German refuge, in a room alone and shut out from the outside world, far away from any of his closer acquaintances, engaged in no other task than that of trying to observe as carefully as he can his own meandering thoughts. That is a meditator par excellence. But there are other, more mundane meditators, such as the one in bed before she falls asleep, or the one engaged in a daydream on a park bench on a Sunday afternoon. These too are meditators in the intended sense.

For the purposes of this discussion, the importance of meditative cases lies in their ability to serve as a test case for theories of self-knowledge. By definition, these are the cases where thinkers experience the least external interference regarding their own internally led train of thought. If one were to assume that people have a dedicated faculty for understanding their own minds—a clear window into one's soul, as it were—and especially if one aimed to explain away cases where people nevertheless misattribute mental states to themselves by invoking the always-present external interference—which taints the window, so to speak—then meditative cases are a good test for one's theory: there should be no such trouble for meditators. If, in a meditative state of mind, there were really nothing to mislead the person about their own attitudes, and if meditators were to find their own thoughts much easier to control, then this would furnish some support for the belief in transparency (Rey 2013a). If, on the contrary, one assumed that the mind is opaque by its nature, and that therefore it remains opaque to the same extent in meditative states as everywhere else, then one should predict that meditators too will be misled about their own mental states at least sometimes and will find their own thoughts difficult to control (Carruthers 2011).

Moreover, if there is such a thing as an unsymbolised thought, then meditators should report more of them. This is because meditative cases are precisely the ones where the need for symbolic vehicles to carry one's thought is the most meagre: if your thoughts are ever for yourself only, then it is in meditative states that they should be so, and therefore you should find no need to cloak them in robes of language when you meditate. If there is no need to communicate your thought, and if you are capable of presenting your thoughts to yourself without any sensory vehicle to bear them, why use any sensory means to do so? If, on the contrary, there is no such thing as an unsymbolised thought, then one should expect people to report plenty of sensory events to serve as vehicles for their thoughts even in meditation.

5.2. Delusions in Solitude

Although all of the experiments discussed here are still imperfect meditative tests, some of them approach the ideal closely. One such experiment is where the experimenter leaves the room and bids the participant to attend to their inner experience in solitude in order to report it afterwards. Another is where the participant is left alone in a room and is gently advised to make notes for themselves that, the participant is told, no one will ever read. Yet another is where the participant is in a resting state in a brain scanner, and it is communicated to them at unpredictable intervals by an auditory signal (from a beeper they carry) that they should report their inner experience of that moment. Finally, there is the experiment where the method just described is used in an everyday context with the crucial difference that the participant who is describing their inner experience is an expert meditator (someone who has practiced formal meditation for more than 10 000 hours). The question to be discussed presently is whether in such cases people are also misled about their attitudes.

The first two experiments to consider were conducted by Jay Olson and his collaborators (Olson et al. 2016). In the first phase of both of these experiments, each participant was asked to lie in an fMRI brain scanner (which was actually only a mock scanner, with the same sights and sounds, but no real functionality beyond deceiving the participants) and told to think of a number from zero to one hundred after the experimenter leaves the room and gives them the signal through a speaker to do so, that is, after the scanner is 'turned on'. Each participant performed two variations of this task in counterbalanced order: in the first variant of the task, they were told that the scanner reads their thoughts, while in the second variant of the task, they were told that the scanner influences their thoughts.

After the first phase of both of these experiments, participants were asked to report their inner experience while in the scanner. In particular, they filled out a questionnaire about their sense of agency, among other things, and in the second experiment, were interviewed generally about their experience when choosing the number. Crucially, they were questioned in a way that is called an ‘elicitation interview’, which emphasises ‘how’ and not ‘why’ they experienced things, and in which the interviewer’s questions are framed in such a way that they are almost entirely devoid of content, serving merely as prompts for the participant to continue talking.

The main findings were that the participants reported a significantly lower sense of agency in the Influence task compared to the Reading task and, in the interviews that followed, they also tended to relate various sorts of outlandish experiences specifically from the Influencing task and not from the Reading task. Examples of these outlandish experiences include the experience that the number was inserted into their minds, that they were unable to change the number when they tried, that they felt hot in the face, pulsations in the brain, or someone else leading them to their decision. In contrast, participants were not inclined to relate outlandish experiences from their time at the Reading task. The mock scanner was, of course, equally inert in both cases, only producing the humming, lights, etc., as if it were a working fMRI scanner.

How should one interpret these results? From the present perspective, these experiments have the limitation that, although the experimenter did not interfere with the participant while the latter was alone in the scanner, they did interfere with the participant earlier: they did try to convince the participant that the scanner will read or influence the mind. Moreover, the mock scanner remained there throughout the two experiments in order to encourage the illusion that it actually works. Consequently, this is not an ideal meditative test. However, if one were to say that cases like this, where traces of earlier interference are still present, are unsuitable conditions for attending to one’s own mind, then one would have to admit that the number of cases suitable for this purpose is rather small. For these sources of interference are already somewhere in-between internal and external: they have their origin outside but remain inside afterwards. If so, are there ever favourable circumstances to observe one’s own mind?

Clearly, some misleading suggestion or other is bound to be present in most cases. For instance, if one were to follow the aforementioned line of reasoning, then one should exclude everyone who reads substandard psychology books, or generally has misguided ideas in the domain of psychology, from ever being able to use the supposed faculty that grants transparent access to one’s own mind. For this reason, the experiment by Olson and colleagues already poses a challenge to the asymmetrical theory.

The next experiment to consider has been conducted by Timothy Wilson and colleagues (reported in Wilson et al. 1989). In this experiment, each participant had to choose a picture to take home. The first group were encouraged to think of reasons for their choice, while the second group did not have to think of reasons for their choice and were, in fact, partly prevented from doing so because they had to engage in a different task in the meantime. All participants were told that the experimenters will not know which picture an individual participant chose to take home. When some time had passed after the initial phase of the experiment, the participants were contacted by the experimenters and asked to evaluate how much they still liked the picture that they took home. It turned out that the Reasons group liked their pictures significantly less, presumably because they were misled about their long-term preferences by their own reasoning about them.

The important point for the purposes of this discussion is that when thinking about the pictures the participants had genuine false beliefs about their own minds and had them in solitude. For as much as any of the participants knew, no one else was ever going to know what their choice was or whether it corresponded to the reasons they thought they had to choose as they did. Under these circumstances, they still believed that they desired the most what, in fact, they did not desire the most, at least not in the long run.

Wilson and colleagues stress that the following is a common feature of experiments eliciting people's mistaken beliefs about their own minds. That is, participants are often left to think in solitude, they are often told that they do not need to write at all, that the things they will write will never be read, that the text will be immediately aggregated by the computer, that it will be thrown in the dustbin, that the things they write are just for them to organise their thoughts, that the things they write are unrelated to the main purpose of the study, etc. (Wilson et al. 1989: 325–326; see also Carruthers 2011: 337–338). Special care is taken to ensure that the participants do not change their reasoning because of the presence of (possible) observers.

Unlike in the experiment by Olson and others, in this experiment by Wilson and colleagues, the participants misattributed attitudes to themselves without there being an expectation that there will be anyone to listen to their judgements or explanations. Of course, the worry remains that suggestions made by the experimenter in the recent past might have influenced the participants and led them to make rationalisations. Another set of more recent experiments allows to address this worry more directly.

The next experiment to be considered was conducted by Hurlburt and colleagues (Hurlburt et al. 2015). Hurlburt is the researcher responsible for developing the descriptive experience sampling method. The essence of this method is that the participant carries a beeper which gives an auditory signal at unpredictable intervals to jot down notes about inner experience at that moment when the signal sounded. The method also involves interviewing participants at length during multiple sessions. These studies often take place over an extended period of time, such as a month. Hurlburt writes that there is a discrepancy between what the participants note immediately after the signal and what they note when they are interviewed several minutes, hours, or days later. This leads to the conclusion that, at one point or another, people's beliefs about their own inner experience are false.

Most importantly, for the purposes of the present discussion, such a discrepancy was found in a recent experiment that was conducted under circumstances that come close to the ideal meditative test (unfortunately, the number of participants was small, so conclusions should still be cautious). Hurlburt and colleagues found such a discrepancy when they compared the reports that participants gave while in a resting state in an fMRI scanner (a real one this time) and the reports that these same participants gave immediately after getting out of the scanner (Hurlburt et al. 2015).

In this experiment, the participants were told: 'please relax, without falling asleep, and do keep your eyes open'. The instructions therefore come close to what one would want to have in an ideal meditative test. The participants did not have any specific agenda and were assailed by almost no external disturbances apart from the rather unusual experience of being in a scanner. The only significant incongruity with the ideal meditative test is that the eyes were to be kept open (but then again, the scenery was not very distracting). From the point of view of the present discussion, however, there is another problem with this experiment, namely, that it does not allow one to compare the participants' present experience itself with their reports about their present experience. The brain imaging data that the experiment generated are but a very imperfect independent measure of what the inner experience of the participants was like at that moment.

However, the significant fact remains that immediately after coming out of the scanner the participants modified their reports. One might suggest that this is due to some sort of memory malfunction. This suggestion is supported by results from recent experiments suggesting that meditation aggravates false-memory recall (Wilson et al. 2015). Be that as it may, appealing to memory malfunction would be problematic for the asymmetrical theory, since it further narrows down the circumstances in which one might expect the supposed capacity to understand one's attitudes without interpretation to work unhindered. If one made this move, then one should also accept that, even if there is no one to mislead you, a few minutes will, by themselves, suffice to taint the clear window opening into your own mind.

Some other cases were already discounted because of immediately prior suggestions. If one were likewise to discount the cases that are now in question because of probable lapses of memory that could occur almost immediately after the experience, then one would be left with a very thin slice of time during which introspection is possible. If one added to this that no external influences should be present during that thin time slice itself and that one should be unsullied by prior misguided beliefs about the mind, then the result would be that, even if there is a special way to know one's own attitudes, one is hardly ever able to use it. This is especially problematic for the inner sense theory. Is it not too great a luxury to evolve a faculty that you hardly ever use?

But one does not need to appeal to memory malfunction, since there is an alternative explanation of the discrepancy. It might be that the reports differ not because people forget, but because people are interpreting the cues that they have at the moment: they give different reports at different times because they have different cues at different times. Moreover, the cues that they have at any one moment underdetermine interpretation: even if the same cues were present at two different moments and were directly available to the participant, they would still allow for a variety of interpretations and consequently the participants' reports could differ. In other words, the ISA theory has no trouble explaining these results.

Finally, there is a more general consideration against the idea that one comes to see one's mind in a special way in meditation. This is that there are so much disagreements about inner experience between expert meditators. These sorts of disagreements contributed greatly to the eventual downfall of introspectionist psychology in the early twentieth century (see Lyons 1986).

It is important to note the precise nature of those disagreements. Here, the people who disagree are experts in their own field, which is that of studying one's own mind. They have had thousands of hours of practice. To them, meditative states are all too familiar. They have plenty of strong incentives and opportunities to set themselves in the perfect meditative state to allow them to report on their own mental life with utmost clarity. In spite of this, there are immense disagreements between introspectionist psychologists.

The disagreement that is perhaps the most relevant here concerns the existence of a sui generis inner experience of decisions, judgements, and similar amodal mental phenomena (that is, phenomena not proper to either of the sensory modalities). The debate raged a hundred years ago and it has been recently revived with no less vigour and probably no better chances of closure (Bayne & Montague 2011; Breyer & Gutland 2015).

Crucially, one might safely assume that these theorists are not different in their relevant mental makeup. No theorist presently working in the field suggests otherwise. One reason why this is so is that the presence of the supposed sui generis cognitive experiences should be a very coarse-grained feature of one's mental make-up. Not being able to experience them would be like not being able to have visual experiences. One might also safely suppose that this is not merely a verbal dispute. Assuming otherwise would be very discouraging indeed: it would mean that all these intelligent and well-intending people who dedicated themselves to the study of this issue were unable to settle a merely verbal dispute during the course of more than a hundred years. If these two suppositions are true, then it seems inevitable that some of these expert meditators are misattributing mental states to themselves and doing so in perfect meditative conditions.

5.3. Monkey Mind

As noted above, the ISA theory also predicts that even meditators will have to resort to indirect behavioural means to control their own attitudes and that even meditators will find their minds difficult to control. The opposite prediction does not follow directly from the asymmetrical theory as such, because there could in principle be direct access without direct control. Yet it would be a blow to the theorist if the mind turned out to be difficult to master even in meditation. This is because the theory needs to explain why the special way of understanding one's own attitudes evolved and for this end the theorist usually appeals to the idea that it helps to control them, which in turn gives an evolutionary advantage. So the theorist should predict that, in cases where the postulated special way works at its best—and meditative cases are supposed to be such cases—one will be able to control one's own attitudes better. But evidence on inner experiences of meditators regarding control of their mental states turns out to support the ISA theory instead: meditation only reveals the difficulty of controlling one's own mind.

The formal meditation practices that are most thoroughly studied today in cognitive science all hark back to the Buddhist tradition (Creswell 2017). Although one should be careful about making inferences from genuinely Buddhist meditation because of the religious factors that come into play and are entirely irrelevant for the present discussion, there is still a lesson to be learned from the general tenor of Buddhist thought about meditation.

One of the symbols for the mind that the Buddhist tradition chooses is the monkey. The animal is chosen for its sprightly nature and to indicate that the mind as the meditator finds it is always restless: like a wild beast, it would be exasperatingly hard to make stand still. In particular, the phrase 'monkey mind' implies that the mind is unsettled, restless, capricious, whimsical, fanciful, inconstant, confused, indecisive, and uncontrollable (Carr 1993). From this one might draw the tentative conclusion that Buddhist meditators are on the ISA theorist's side on this issue.

But Carruthers himself argued otherwise. Based on textual evidence, he claimed that the Buddhist tradition understands the mind as transparent. This contributed to his argument that the transparency assumption is a human universal. However, as noted above, some experts on the relevance of Buddhism to contemporary philosophy, such as Jay Garfield, claim that it is a mistake to think that all Buddhist traditions regard the mind as transparent to itself. If the Buddhists to whom Garfield refers disbelieve in transparency, then they should also be inclined to disbelieve in direct control, even in meditation: one can only directly control what one is directly aware of, at least in the conscious and intentional sense of 'control' that is relevant here.

Admittedly, this argument from Buddhist conceptions of meditation is rather impressionistic, but it does put some pressure on anyone who wants to hold that only external interference prevents one from directly accessing and controlling one's own mind, and that meditation is the solution. Why would one find the opposite suggestion in the oldest and best-established tradition of formal meditation in the world?

There is another source of evidence on what meditators experience when they try to control their own thoughts, which is, in certain respects, more precise. The evidence comes from self-reports delivered by participants who are involved in meditation-based psychotherapy interventions. Today, there is already an industry of empirical research on these programmes and a wide variety of programmes on offer. The interventions that are most relevant for the present discussion all have at their core a formal meditative exercise that looks roughly as follows. The participant is instructed to keep an upright posture, close their eyes (unless they prefer otherwise), and direct their attention to sensory aspects of their own experience, for example, to the physical sensations of their own breath, wherever they feel it most clearly; when the mind wanders, the participant is to note the thought, feeling, or sensation that drew their attention away, stay with it for a while, and then redirect their attention back to their breath; the exercise takes from three to forty-five minutes (Segal et al. 2013: 383).

Of paramount importance for present purposes is that people engaged in these exercises almost invariably find them difficult to perform and express surprise at the unruliness of their own mind (Creswell 2017: 494). So much so, that this form of psychotherapy is considered dangerous for certain groups of people (Creswell: 507). In particular, present symptoms of severe depression are considered a contraindication, since the feeling of failure, which is only to be expected, might adversely affect the patient. It is a contraindication even in spite of the fact that patients who have relapsed to major depression more than twice are precisely the ones for whom the therapy is known to be the most effective. So it is not that people who find the practice difficult are very different from the ones that benefit from it. Consequently, authors of these therapy programmes suggest special caution regarding participants' feelings of failure and administer advice on coping with it from the perspectives of both the therapist and the patient.

In order to see just how difficult it is for an ordinary individual to 'just think', one might also consider participants who have no psychopathological symptoms and are not even engaged in therapy (Wilson et al. 2014). Wilson and colleagues did a series of eleven experiments with such participants. The typical setup was that they asked participants to put away their belongings, including their mobile phones, and to stay alone in an unadorned room for 6 to 15 minutes. The participants had no instructions other than to entertain themselves with their thoughts and stay awake. Afterwards, the participants answered such questions about their experience as how enjoyable it was and how hard it was to concentrate. In some variations of the experiment, participants were in the laboratory, in other variations they were at home. Sometimes they could choose a mundane activity, such as reading a book or texting, as an alternative to attending to their own thoughts. Sometimes they were encouraged to decide beforehand on a pleasant topic to think about. In one final version, participants were first given a mild electric shock and asked how much they would pay to avoid experiencing it again. Afterwards, they could either just think, or administer the shocks to themselves if they wanted to.

From the perspective of the asymmetrical theory, the results are rather surprising. Most participants found 'just thinking' unpleasant, said it was difficult for them to concentrate, and preferred any other social or non-social activity to entertainment by pure thought. Even more surprisingly, many of them freely chose to administer themselves electric shocks, presumably just to avoid being alone with their thoughts. From the group of people who said they would pay money not to receive the electric shock again two-thirds of men and a quarter of women administered further shocks to themselves. The interviews revealed that the participants found 'just thinking' difficult irrespective of whether the contents of their thoughts were self-centered or not and whether the thoughts were positive or negative.

These findings seem much less unnatural from the perspective of the ISA theory. They are also less surprising when one considers what is known about controlling one's own mind in situations that are not obviously meditative in character. It is well known that keeping a thought out of one's mind is hard ('Don't think of the pink elephant', earworms, etc.), and also that, as far as anybody knows, the best method for keeping them out is concentrating on something else (Wegner 1994, Smallwood & Schooler 2013), which is a rather indirect method.

Again, this does not mean that meditation is not instructive or that, when one practices it, it does not help to improve one's grip on one's own mind. However, it does suggest that staying alone with one's own thoughts reveals a harsh reality, rather than the serene refuge that some might have hoped for.

5.4. Coming to One's Senses

The previously discussed evidence supports rather than challenges the claim to adequacy by the ISA theorist in relation to the results of the available empirical research on meditation. But in addition to explaining the already available evidence, the theorist should also try to make new predictions in this domain. In particular, the theorist could make certain new predictions on the intuitions of expert meditators about self-knowledge, on the reports by meditators of unsymbolised thought, and about the attitudes meditators take towards themselves, explicitly or implicitly. Viewed from the ISA theorist's perspective, the crucial point of the discussed formal meditative practices is, quite literally, to bring the meditator to their senses: to bring their attention to their own sensory experience. The theory has the resources to explain why this should have the effects that it does.

Here is the first prediction that the ISA theorist should make: the theorist should predict that expert meditators will intuitively think that their relation to their own minds is more similar to their relation to other minds. Here is the rationale: if meditation reveals the nature of the mind—and Buddhist meditators would certainly stress the point—and if the nature of the mind is as the ISA theory describes it, then the theorist should predict that expert meditators will be more inclined to think that the mind is opaque than people who are not expert meditators.

There is an important caveat, however. As argued earlier, acquiring self-knowledge is mostly an unconscious process. If the mindreading faculty does most of its work below consciousness, then the theorist should not predict that even expert meditators will be able to report on the process accurately in all of its stages. But the theorist should predict that expert meditators will be less averse to the idea that the contents of one's conscious mind are mostly sensory in character. They should also predict that expert meditators will tend to agree that answers to questions about one’s own or others’ attitudes spring to mind without one being aware of the entire process by which such judgements and answers come about. Finally, the theorist should predict that the interpretive nature of self-understanding will become apparent at least to some extent to those who spend time meditating, because these people become aware of the way in which judgements about attitudes tend to follow relevant sensory cues.

Here are two ways how one could test the first prediction. One could take a group of expert meditators and a well-matched group of non-meditators. For example, one could take a group of Buddhist monks, who regularly meditate, and a group of their compatriot Buddhists who do not regularly meditate (sampling Buddhists in both cases might be important for the reason that Buddhism as such might have an effect on their views about self-knowledge). Then one could measure their beliefs about self-knowledge (lacking a better instrument, one could use the vignettes from Kozuch & Nichols 2011). The prediction is that monks will tend to agree that the mind is opaque more than laypeople. Alternatively, one could take a group of participants enrolled in a meditation-based psychotherapy programme and administer the same measures to them before and after they have completed the standard eight-week course (such as the one described in Segal et al. 2013). The prediction is that the participants should tend to agree that the mind is opaque more after they have taken the course.

There already is some indirect evidence suggesting that the prediction would be confirmed. This is because it is commonplace in both the Buddhist and the psychotherapeutic tradition to stress that one discovers the nature of one’s own mind by taking a step back: by recognising your thoughts as ‘just thoughts’ and learning not to immediately identify with them, just like one would not unhesitatingly attribute a thought to someone else merely on the basis that they are saying something, if one reflected on it.

Here is the second prediction: the ISA theorist should predict that meditators will report more sensory experiences and less unsymbolised thought. If one were to assume that meditation makes one more aware of the nature of one's inner experience, and that the nature of one's inner experience is as the ISA theory describes it, then the theorist should predict that meditators will report more experiences that are sensory in character compared to non-meditators.

Again, there is an important caveat. As noted above, the introspectionist psychologists, who are expert meditators in a sense, were unable to settle the issue whether unsymbolised thought exists. If one were to address exactly the same question again, then there seems to be no reason why one should expect to succeed where others failed. The hypothesis under consideration, however, does not broach the very same question again. The aim is not to settle whether sui generis cognitive phenomenology exists. It is merely to find out whether the tendency to report unsymbolised thought grows or withers with meditation practice. No such experiment has been conducted, although people in the cognitive phenomenology debate do sometimes make suggestions one way or the other.

There is some evidence suggesting that the second prediction would be confirmed. In the experiment by Hurlburt and colleagues, which was discussed earlier, one of the findings was that people in a resting state in a scanner reported more sensory experiences than is typically reported by similar populations in more everyday circumstances (Hurlburt et al. 2015). Likewise, the sole case of an actual expert meditator who participated in a long-term study by Hurlburt and colleagues showed that he mainly diverged from the norm in that his reports of inner experiences referred to sensory events more than is usual (Hurlburt & Heavey 2006: 246) (of course, one should be cautious not to make strong inferences from a sample of one).

Here is the third prediction: the ISA theorist should predict that attention to one's own sensory experience plays the central role in meditation-based psychotherapy interventions. The intervention that is widely agreed to weather criticism directed at such interventions the best (Van Dam et al. 2018) is Mindfulness-Based Cognitive Therapy (MBCT), which is tailored to formerly depressed patients that have relapsed more than twice (Segal et al. 2013). It is clear that the programme relies on the kind of exercise of redirecting attention to one's sensory experience that was briefly described above. We know that this is what the therapy consists in, at least to a large extent, and we know that the results are diminished depressive symptoms mediated by diminished rumination.

Based on this knowledge and the ISA theory, one can hypothesise that it works through something like the following mental mechanism. From the perspective of the ISA theorist, conscious experience consist of sensory representations, and conceptual and affective representations that are bound into them. In performing a formal meditation exercise, one is focusing on the sensory aspects of experience and not the conceptual or affective ones. During the course of the programme, one slowly forms the habit to attend first and foremost to the sensory aspects. As a result, the content of one's conscious experience tends to become more sensory overall and to leave out the conceptual and affective aspects because one forms the habit not to attend to them and without attention they do not reach consciousness. The faculty responsible for attributing mental states to oneself, according to the ISA theory, feeds on information that becomes conscious through being bound into sensory states. This means that, in the meditator, one has the mindreading faculty fed almost entirely by sensory information, so it should attribute attitudes accordingly.

In the case of the chronically depressed patient, this means that their experience is deprived of such conceptual content that tends to perpetuate rumination and of affective content that is mostly negative. Both of these are usually bound into their sensory experience in the conscious mind. If so, then one should predict that the habit of distilling the sensory aspects and leaving out all the others should result in less rumination. Here is a more concrete example of how this would work. The inner speech utterance 'I am a failure' is a sensory-affective-conceptual bundle: it includes the quasi-sensory experience of uttering the sentence in inner speech, the negative affective content attached to it, and the conceptual content that the speaker is a failure. During the exercise, the participant habituates themselves to attend primarily to the sensory part of that bundle and to disregard others. If this hypothesis is right, then the intervention should work even if one eliminated all the other elements, such as extensive group discussions, if they do not encourage selective attention to sensory experience. No such simplified intervention has yet been tested.

Finally, here are two more particular hypotheses. The previous discussion leaves open two possibilities regarding the mental mechanism in meditation: meditation could work by revealing the nature of the mind, or it could work by fostering a positive illusion. On the first interpretation, one would say that attending to sensory experience allows one to understand how the mind interprets itself and to take the process into one's own hands, insofar as this is possible. On the second interpretation, one would say that attending to sensory experience fosters the positive illusion that one no longer has the negative thoughts that one used to have, which might be false, but in some cases it might be better not to know what one's attitudes really are, such as in the case of the chronically depressed. To be sure, the second interpretation jars loudly with how the Buddhists and the psychotherapeutic tradition just discussed describe things.

One could pit these two competing hypotheses against one another by measuring both explicit and implicit attitudes of meditators towards themselves. If the first interpretation is right, then one should find that the intervention increases the correspondence between implicit and explicit attitudes. If the second interpretation is right, then one should find that the intervention decreases the correspondence between implicit and explicit attitudes. Again, one could measure this before and after a standard meditation-based intervention. No similar test has yet been carried out.

5.5. Conclusion

The ISA theory receives support from empirical research on meditation. It predicts that meditators will misattribute attitudes and find them hard to control, while its rivals predict that meditative cases will be safe and easy. Experimental evidence and disagreements between introspectionists suggest that meditators misattribute attitudes to themselves. Likewise, experimental evidence and traditional Buddhist and psychotherapeutic conceptions of meditation suggest that meditators find their attitudes difficult to control. If the theory is right, then experienced meditators should find opacity more intuitive and report more sensory inner experiences. Finally, it offers an explanation of the mechanism behind meditation that could be tested by measuring meditators' implicit and explicit attitudes towards themselves.

6. UNWITTING ACTORS

> '…I have often said that man's unhappiness springs from one thing alone, his incapacity to stay quietly in one room.'
>
> Pascal 1670/1995: 44

If the mind is opaque, then there is no conscious will. Consciousness is the surface of the mind, while decisions and intentions, the expressions of one's will, are always under that surface. One always interprets one's own sensory states to know one's will. One relies on two sources of evidence. Sometimes one interprets overt behaviour, which is something that can only be done after an overt bodily movement has occurred. The interpretation is mostly done unawares, which makes one an unconscious behaviourist, so to speak. Sometimes, there is no overt behaviour to be interpreted. In such cases, one interprets imagined behaviour, which is something that can be done before any overt bodily movement occurs. This is like staging a play in your mind's eye where you yourself play the lead role and the critic. The performance is almost always produced without any awareness of the purpose, which makes one an inadvertent performer. So the our predicament makes unwitting actors in a double sense.

6.1. New Willusionism

One might think that the ISA theory is an illusionist approach to conscious will. The latter view has sometimes been referred to as 'willusionism'. But if one were to say that conscious will is an illusion, one would say two things: first, that people believe conscious will exists, and second, that conscious will does not exist. The following discussion certainly goes in the footsteps of researchers who accept both of these claims, in particular, Daniel Wegner (Wegner 2002/2017; see also Carruthers 2007). However, it will only advance the second claim: that there is no conscious will; not the first claim: that people believe otherwise. As noted above, laypeople are probably quite indecisive when it comes to the finer points of theories of self-knowledge, those that would help one to tell between current competitors. As it happens, there is a reason to think that their intuition will be largely silent in the case at hand, that is, the existence of conscious will.

The reason is that at the center of the current debate about our knowledge of our own decisions lies a rather fine distinction between two possible causal roles for certain events in the conscious mind. All sides in the debate agree that there are events in the conscious mind that one controls directly. For instance, I can say to myself, whether on my own whim or when asked by someone else to do it, that: 'I will be going out now'. All sides in the debate agree that such events have an effect on one's own behaviour, such as influencing one to eventually go out, in the example given above. What the opposing sides disagree on is whether such events affect behaviour directly or through their effect on further unconscious reasoning that then leads to action (see Vierkant 2015; analogous discussion concerns the causal role of epistemic emotions, see Dokic 2012). Since this is a rather fine theoretical point, perhaps one should not expect laypeople to have strong opinions here. In any case, the currently available evidence on self-knowledge intuitions, discussed above, will certainly not settle the issue.

However, perhaps one could test this in a future experiment. For example, one could ask people whether it is possible that a certain candidate mental event occurred in their conscious mind, such as saying to oneself 'I will get out of bed now', but the corresponding behaviour failed to follow, even though there were no external hindrances to following it through. As might be suspected from the choice of example, the hypothesis suggested here is that people will be quite willing to accept that all sorts of events in the conscious minds fail to directly cause corresponding behaviour. If so, then one should conclude that, in laypeople's conception, these events do not play the causal role of decisions, as philosophers define them. It is another issue, whether laypeople would still want to call them decisions (see Frankish 2016). One could predict that they probably would want to call them decisions, but then again, one might also predict that they would retract their words after a philosopher gave them their reasons for not calling those things decisions.

The final point to note about the question of belief before moving on to the question of existence is this: to refrain from saying that people believe that conscious will exists is not to say that they have no illusions about their will. Illusions about the will play a crucial role in the following discussion, but those are illusions about other features of the will than its supposedly conscious character. They will include, for example, the timing of the act of will and the very existence of a particular act of the will. In this sense, the view proposed in the following does qualify as a (new) kind of willusionism. It is a kind of willusionism at least in this sense: since the ISA theorist says that one only knows one's own acts of will through interpretation, they should predict that one will be prone to illusions about one's own will. In particular, the theorist should predict that one will fall for illusions about one's own will when misled by sensory cues that are analogous to those that would mislead an outside observer.

Most of the experiments that will be discussed here belong to a tradition that stems from empirical researcher into the neural antecedents of one's intentions and urges. These experiments have cast a new shadow of doubt on the idea that one has direct access to one's own acts of will. Probably the first and certainly the most influential of these experiments was conducted by the neuroscientist Benjamin Libet and his colleagues (Libet et al. 1983).

Libet started his experiment by seating his participants in front of a clock with a quickly rotating light beam. He then connected them to machines that measure electric activity: an electroencephalograph (EEG), which measures it in the brain, and an electromyograph (EMG), which measures it in the muscles, in this particular case, muscles of the right hand. He told his participants to look at the center of the clock and to skip the first full rotation of the light beam. He also asked them to then spontaneously flex their right finger or wrist at any time of their choosing. He entreated them not to plan when they will move beforehand. Finally, he instructed them to memorise and report the location of the light beam at the moment when they first became aware of their urge or intention to move. The most important finding was that certain electric activity reliably started in the brain 500 ms before the start of electric activity in the muscles, which indicated the start of the movement, and 300 ms before the participants first became aware of their urge or intention to move.

Different interpreters differ widely on the right understanding of Libet's results (see Sinnott-Armstrong & Nadel 2010). Some of them have argued that these results suggest the following: if the event that decided the time of movement occurred at -500 ms and not at -200 ms, then the real decision to move occurred at -500 ms, and not at -200 ms. They argued that conscious 'decisions' inertly follow unconscious decisions, and therefore the will itself is unconscious, and so there is no conscious will. However, few researchers working in the field today would rely in their argument on Libet's original findings, even if their argument is largely inspired by them. This is because most would agree that there is now a question mark on almost every point of Libet's picture: what the brain signals are, what the time of their onset is, what the participants are reporting, what the time of their reports is—all of these are now moot.

Luckily for the purposes of this discussion, one can bypass much of this debate and concentrate on a particular strand of experimental work that was inspired by Libet's experiment. This particular strand of empirical research often follows his original paradigm quite closely. However, it puts most of the emphasis on behavioural measurements. This research provides one with two sorts of evidence. On the one hand, it suggests that participants rely on external cues: on perceptual feedback on their own overt movement, such as visual feedback on the movements of their own hand. Because of what is known about the role of external cues, the research poses a challenge to the asymmetrical theory: if you have non-interpretive access to your decisions, why rely on external cues?

On the other hand, results from this line of research suggest that participants also rely on internal cues: on something that exists before the participant performs an overt movement, such as when they are interrupted just before they start to move. Because of what is known about the role of internal cues, the research poses a challenge to the symmetrical theory: if they have not yet performed an overt movement, what are they interpreting? Both the asymmetrical theory and the traditional willusionist—who holds that one always relies on both perceived movement and prior thoughts—face problems when trying to account for the full range of evidence. However, the new sort of willusionism proposed here smoothly accounts for it all.

6.2. Unconscious Behaviourism

The ISA theorist's argument, as it relates to knowledge of one's decisions, was originally based on a study that, although it might have been inspired by the tradition stemming from Libet's work, did not at all closely follow his original paradigm. However, over the course of the years that have passed since the ISA theory entered the scene, it became apparent that this particular study has severe methodological limitations. On its own, it provides only very unsteady support for the willusionist's argument. Luckily for the willusionist, however, there are now methodologically stronger studies that serve the same purpose. A feature of these studies is that they also follow Libet's paradigm more closely. One such study suggests that people's reports change with changing perceptual (doctored) feedback. Another suggests that they sometimes misattribute their own decisions to move to something else.

But first, perhaps it is worthwhile to consider the experiment upon which Carruthers originally relied, to a large extent, in arguing that decisions are not conscious (Carruthers 2011: 339–342). More precisely, one of the few experiments that he cited to directly support his claim about decisions in particular: one might well be sceptical about the existence of conscious decisions based on general considerations concerning the mind's opacity.

The study in question is the famous 'I Spy 'experiment by Daniel Wegner and Thalia Wheatley (1999). In this experiment, participants were seated together with a confederate of the experimenter in front of a computer screen showing roughly fifty small objects from a children's game called 'I Spy' (e.g., a swan). The participant and the confederate wore headphones, through which, the participant was told, they will hear music and words. Some of the words were names of the objects on the screen. Participants were told that the other 'participant' will hear different words, and that the words serve as a minor distraction. For the first 30 s, they were to move a cursor on the screen in small circles together with the other 'participant'. After this, they were to hear 10 s of music and sometime into it to stop the cursor on an individually selected picture on the screen. Finally, they had to report whether it was their decision to stop the cursor there or that of the other 'participant' (0 = 'I allowed the stop to happen', 100 = 'I intended to make the stop').

Here are the main findings. On the trials where they had full control of the cursor—where the confederate was told not to interfere—participants tended to say that it was their decision to stop, although this tendency was only slight (*M* = 56.09, *SD* = 11.76). On the trials where they had no control of the cursor, participants tended to say that it was their decision to stop if they heard the name of the object on which the cursor stopped, although again the tendency was rather slight (*M* = 52% of trials, *SD* = 23.95). For the Forced trials, reported intentionality varied depending on the time when the participant heard the name of the object: the stop was reported as not intentional if the word was heard 30 s before the stop (*M* = 43), as intentional if it was heard 1 s (*M* = 60) or 5 s (*M* = 62) before the stop, and again, as unintentional if it was heard 1 s after the stop (*M* = 47).

The researchers who conducted this experiment concur with Carruthers in reading these results as suggesting that the participants interpreted the word they heard before the stop and the stop itself that followed as signs that they themselves had made the decision to stop. If they did not have to interpret—if they had transparent access to their own decisions—why would they say that they made the stop when they did not, and that they did not make the stop when they did?

Unfortunately for willusionists, this experiment has garnered intense methodological criticism over the years that followed. A good example are the worries raised by Sven Walter (Walter 2014). First, 60 or 62 out of 100, in the Forced condition, does not quite amount to 'I intended to make the stop', which would be 100. Second, 56 out of 100, in the Free condition, is roughly in the middle and not that different from 60 or 62. Third, the intentionality rating in the Free condition was averaged over 1275 trials, while in the Forced condition it was averaged over as little as 37 trials (see the calculations by Walter). Fourth, Wegner's own studies on facilitated communication suggest that sometimes when people are supposed to merely recognise the decisions of others they nevertheless actively interfere without themselves noticing that they do (Wegner 2002/2017). If that is true, then the confederate was likely to actively interfere when they were supposed to merely let the participant make the stop. Fifth, on some of the trials in the Forced condition, participants probably decided to stop on the same object as the confederate. This is quite likely for several reasons: they had to make the stop during a 10 s interval, while the music played; people tend to follow this instruction by stopping midway through the music; there were not too many objects that could be reached by the confederate in time without making the pattern of movement suspicious (hastening it, etc.); participants reported that they sometimes searched for the object that was named, and that is the object where the confederate had to stop on many of these trials. All of this leaves the willusionist with the feeling that another experiment to support their cause would be welcome.

Luckily, there is a more recent attempt in this domain that follows Libet not only in spirit but in a more literal way: the two experiments conducted by William Banks and Eve Isham (Banks & Isham 2009). In the first experiment, they used basically the same setup as Libet. One difference was that they only connected their participants to a machine measuring electric activity in the muscles and did not measure electric activity in the brain. Another difference was that the participants had their finger on a button which they had to press down during the second rotation of the light beam. Closure took place when the button was depressed 2.5 mm. The button gave no tactile feedback when closed. The last and most crucial difference was that the participants heard a signal at 5, 20, 40, or 60 s after closure. The main finding was that participants' reported time of their first becoming aware of the decision to press the button moved forward in time together with the delay of auditory feedback on the button press.

The second experiment differed from the first mainly in that instead of hearing a delayed signal participants saw a video of their own hand pressing the button which was sometimes in real time and sometimes delayed by 120 ms. The main finding was that the participants' reported time of their first becoming aware of their decision to press the button shifted forward in time by 44 ms when the visual feedback was delayed.

The authors interpret these results as suggesting that one infers rather than perceives the moment when one decided to act. Why rely on auditory or visual feedback, when reporting the time of your decision, if you have transparent access to your decisions? One limitation of these two experiments, when considered with the goals of the present discussion in mind, is that they concern illusions about properties of decisions and, in particular, their timing, rather than the existence of those decisions. The following experiment addresses this further issue.

This is a study by Alexander Schlegel and colleagues, which they named ’Hypnotising Libet’ (Schlegel et al. 2015). Overall, the setup was again similar to Libet’s. One difference was that the participants were seated in front of a computer screen with their hands on their lap and under an occluder, palms up, each hand loosely holding a stress ball. During the first phase of the experiment, the participants watched nature videos of 20 s duration, with a Phillip Glass soundtrack. They had to press a stress ball at some point during each of the videos with either their left, or their right hand, depending on which way a red arrow pointed on the side of the screen. After this, hypnotic induction followed, during which the participants were instructed to press the stress ball with either their left, or their right hand, depending on the cue on the screen. Then, participants were woken up from their hypnotic state and told a cover story which said that the experimenters will now calibrate the EMG machine and that this might make the participant’s forearm muscles contract.

The second phase of the experiment followed, which was like the first phase in all other respects except that blue semicircles were used instead of red arrows and that the task for the participant was now to just watch the videos while the EMG machine was being ‘calibrated’ once for every video of 20 s duration. After this, another hypnotic induction followed, which removed the suggestion from the first hypnotic session. Finally, during the third phase of the experiment, the participants had to respond to the blue semicircles like they responded to the red arrows. After this phase, the participants were thoroughly interviewed for any suspicions about the hypothesis of the study.

The following findings only concern the participants who did not guess the hypothesis. The main finding was that participants thought it was the EMG machine that made them press the stress ball during the ‘calibrating’ phase, not the participants themselves. This was in spite of the fact that the machine did nothing else besides measuring their physiological reactions: it did not influence their movement in any way. Another finding was that there were no significant differences in the brain signals between the second and the third phases of the experiment, so the disowned movements bore all the neural marks of voluntary action.

The authors interpret these findings as suggesting that conscious willing is not necessary for voluntary action. They define voluntary action as an action that is caused endogenously, as opposed to being merely a reaction to an external cue, and that is not merely a reflex.

Here is a summary of what the experimental literature suggests regarding the influence of external cues on people's reports about their decisions. The experiment by Wegner and Wheatley provides (weak) evidence that people sometimes fail to attribute to themselves decisions they did make and sometimes attribute to themselves decisions they did not make. The experiments by Banks and Isham provide strong evidence that people rely on perceptual feedback in their reports of the timing of their decisions. The experiment by Schlegel and colleagues provides strong evidence that people sometimes fail to attribute to themselves decisions they did make.

All of these cases constitute anomalies for the symmetrical theory. If people have special access to their own decisions, then it is not clear why they do not use that access in these cases and choose to attribute decisions to themselves based on their perceived behaviour. Perhaps this penchant for unconscious behaviourism can be accommodated by the symmetrical theory. However, there seems to be no obvious way of doing so that would not require postulating an additional mechanism, which would be responsible for these interpretations of behaviour, and therefore no way of doing so without making the symmetrical theory more complicated.

6.3. Inadvertent Performances

In his recent review of the literature on the neurobiological foundations of voluntary action, Patrick Haggard stresses that traditional willusionism also finds some of the recent results difficult to explain (Haggard 2019). The kind of willusionism he has in mind is the theory that the subjective experience of volition is always based on the interpretation of prior thoughts and perceived behaviour, a view that he attributes to Wegner. He notes that we now know that the subjective experience of volition can be dissociated from perceived behaviour and prior thoughts.

The dissociation goes both ways. On the one hand, sometimes subjective experience of volition and prior thoughts about action are present when corresponding overt behaviour is absent. On the other hand, sometimes perceived behaviour and prior thoughts about it are present when the subjective experience of volition is absent. The suggestion pursued in the following is that an updated form of willusionism, which based on the ISA theory, has the resources to deal with these challenges. The crucial role here is played by inadvertent performances of action in the mind's eye.

It is worthwhile to first examine Haggard's point in more detail. One of the studies he cites is an experiment by Itzhak Fried and colleagues (Fried et al. 1991). The express purpose of the experiment was to map the functional organisation of the supplementary motor area of the human cortex. The participants were patients suffering from intractable seizures, undergoing evaluation for surgery. These uncommon circumstances allowed Fried and colleagues to place electrodes directly onto various areas of the cortex of the patients. After doing so, they were able to observe the motor reactions that were elicited by electric stimulation. They also collected verbal reports from the patients on their experience during the stimulation of each of those areas.

They found that participants sometimes report things like an ‘urge to move right arm’ or a feeling as if a right hand movement ‘was about to occur’ when certain of these brain areas were stimulated. They also found that stronger stimulation of the same areas often resulted in actual movement, not unlike the one that was implied in the participant’s felt urge or premonition. Crucially, the urges were reported even when there was no actual movement, no perceived behaviour.

More recently, Michel Desmurget and colleagues followed up on this experiment, extending it in many respects (Desmurget et al. 2009). Again, they used the opportunity for direct electrical stimulation of the cortex in patients undergoing awake brain surgery. They stimulated premotor and parietal areas. Like in the experiment by Fried and colleagues, they also collected reports from their participants about their experiences when each of the areas was stimulated.

What they found was similar, insofar as the interests of the present discussion are concerned, but it went further in many respects. They found that stimulating the right inferior parietal region triggered a strong intention and desire to move the contralateral hand, arm, or foot. When stimulation to this area was increased, participants thought that they had actually executed these movements, even though there was no discernible change in the electrical activity of the corresponding muscles. Stimulating the premotor region triggered overt mouth and contralateral limb movements, but patients firmly denied that they had moved intentionally.

The authors interpret these results as suggesting that conscious intentions arise from increased parietal activity prior to movement execution. This means that the subjective experience of intending to move is dissociable from perceived movement and even from the first discernible brain signals showing that the movement’s execution has been initiated.

Masao Matsuhashi and Mark Hallett conducted an experiment that follows Libet’s paradigm more closely, and found a similar dissociation (Matsuhashi & Hallett 2008). In their experiment, participants had to extend the index finger every 5–10 s as briskly as possible, without thinking when they will move beforehand, counting, or keeping time. Participants were also told that throughout the experiment they will occasionally hear a certain tone. If they heard the tone when they were already thinking about the next movement, then they had to refrain from moving and wait for another 5–10 s. If they heard the tone when they had already started the movement, then they had to refrain from finishing it, if they were still able to do it. They were to keep the hand muscles relaxed throughout.

The main finding that is relevant here was that the participants were perfectly able to refrain from moving after, they said, they had already formed the intention to move. This poses a problem for the traditional willusionist, because here we have an intention to move without any perceived movement which would follow and which could serve as the basis for behavioural interpretation.

Christos Ganos and colleagues discuss a dissociation from the other side (Ganos et al. 2018). It concerns patients with tics. These repetitive movements range from very simple ones such as twitches, to more complex ones such as jumping, cursing, or even inappropriate remarks at other people's expense that are sensitive to the fine points of their particular context of utterance. The patients often think about those behaviours before they happen. The behaviours have also been shown to depend on the same channels as ordinary voluntary action. For instance, if a patient with a tic of shaking their left arm starts rhythmically moving their right arm, then their left arm stops shaking and follows the rhythm of the right arm. In this respect, the patients are just like a healthy individual who would also find it hard to randomly shake their left arm and at the same time move their right arm in slow rhythm.

If we agree that people with tics have thoughts about the movement before it happens, that they perceive the movement, and that the movement is voluntary, then we have a case where the subjective experience of volition is absent when perceived movement and preceding thoughts about it are present. Again, this is an anomaly for the traditional willusionist, who holds that subjective experience of volition is based on interpretations of perceived movements and corresponding prior thoughts. Yet it is easily explained by the asymmetrical theory, which holds that one has transparent access to one's own intuitions, so there is no need to wait for bodily movement to occur or to resort to interpretation. However, it has been argued above that, while the traditional willusionist has trouble explaining internal factors influencing the subjective experience of volition, the asymmetrical theory has trouble explaining external factors influencing the subjective experience of volition. Neither of these theories accounts for the full range of available evidence.

Whereas, the ISA theorist—the new willusionist—accommodates both internal and external influences on the subjective experience of volition, without making their theory any more complicated. The key to solving the problem is mental imagery. Obviously, the theorist has no problem with external influences on the subjective experience of volition: these are only to be expected if one attributes acts of will to oneself by using the same mental faculty which one uses to attribute acts of will to others. More to the point, the theorist has no problem explaining the cases where the subjective experience of volition dissociates from perceived behaviour. This is because, according to the theorist, the evidential basis from which one infers one's own attitudes includes sensory states broadly construed: not only perceptual states, but quasi-perceptual states too. This means that the evidential basis includes mental imagery of movements that is generated before one moves.

There are good reasons to think that such simulations of movement take place prior to movement. In fact, some researchers have suggested that it is the main reason why one has quasi-perceptual states at all, and Carruthers himself is sympathetic to this view (see Carruthers 2015). According to this theory, one generates motor intentions and then refrains from executing them at the last moment in order to merely simulate the execution offline. One does this in order to see what it would be like if one actually executed them without the dangers of actual consequences. This generates quasi-perceptual feedback which is similar to the one that the mind predicts that one would receive if one actually engaged in the simulated behaviour. Evaluating the feedback helps choose the course of action that, from the available options, is likely to bring about the most desirable results.

Here is how this account applies to cases that were just discussed. In the experiments just described, even if participants really manage to refrain from settling on an answer, as they are asked to do by the experimenters, they still have a question on their mind throughout: when will it be best to move and which movement will it be best to perform? The presence of the question is likely to generate mental imagery of performing different movements at different times. Likewise, patients with tic disorders probably have ample imagery of movements that are likely to come, on the basis of which they might ascribe to themselves the intention to act. This is not to say that they need not ascribe intentions to themselves automatically whenever they have such mental images. After all, they are supposed to help choose between alternatives. So these images should be of different possible movements, not all of them is going to get executed.

The suggestion that mental imagery is at play here, is also corroborated by some of the neurobiological findings. The areas that are involved in 'intention generation' (Desmurget et al. 2009) are also known to be involved when people generate mental imagery. Moreover, brain areas responsible for somatosensation are also known to show signs of increased activation prior to the commencement of movement, in experiments very similar as those discussed above (Schurger, personal communication 2019).

It is important that this practice—of performing mental simulations of the possible courses of action before committing oneself to any of them—is performed without pre-planning and often experienced as having no purpose. People often think of mental imagery as resulting from the idleness of their own mind and as serving no specific goal. However, there are reasons to believe that this is not the case: the difference between that part of the stream of consciousness which seems purposeful and that part of it which does not seem purposeful probably lies not in the goals they do or do not serve but rather in one's knowledge or ignorance of those goals (Carruthers 2015). Not knowing what goal the imagery might serve, one concludes that it is idle. But it would be rather strange, from an evolutionary perspective, if people exerted those precious resources that are required for generating mental imagery for no purpose. More likely, the goals that mental imagery is there to serve are unconscious. One inadvertently acts out in the mind's eye what one might or might not do in the future, in order to decide on the best course to take. This is done quite spontaneously, without pre-planning or dedicating special attention it. The suggestion is that, likewise, one often spontaneously interprets these imaginings as signs that one has decided to act.

Here are a few ways one might try to tests this new suggestion. Nobody seems to have interviewed the participants in Libet-style experiments about the presence or absence of mental imagery at or before the time they decided to move. So one could ask just that. If the above account is on the right track, one might also predict that people with a stronger tendency to generate mental imagery should have a stronger presentiment that they will act before the movement happens. A particularly interesting population to study in this respect would be those who say that their mind's eye is blind, people who say that they never experience mental imagery at all (Zeman et al. 2015). More generally, to test this account, one could look more closely into the relation between feelings and judgements of agency on the one hand and mindreading abilities on the other hand. Perhaps, people who have impaired mindreading abilities also have an impaired sense of agency, and people who have an impaired sense of agency also have impaired mindreading abilities.

Some indirect support for this hypothesis can already be gleaned from recent empirical research. This research finds a positive correlation between people's score on mindreading tests—how likely they are to see an action as intentional, for instance—and their belief in free will, whether their own or other people's (Genschow et al. 2019). This suggests people's mindreading abilities influence their judgements about the intentionality of other people's movements as well as their own and other people's possession of free will. Perhaps their mindreading abilities also influence their judgements about the intentionality of their own movements, and maybe even their feeling that they have moved intentionally.

6.4. Mixed Feelings

Suppose one agreed that all knowledge of one's own will is mediated by interpretation. From this point onwards, one could develop the theory in two different ways. On the one hand, one could say that the mindreading faculty is charged with interpreting sensory cues which indicate that one has made a decision. Since, on this first reading, the process of attributing decisions to oneself is interpretive, one should predict that it will be susceptible to the usual errors of interpretation: to the influence of misleading sensory cues. This is what was found in the experiments discussed above. However, on this first reading, the sensory cues themselves are not affected by the mindreading mechanism, and so one should predict that participants will be able to report the sensory cues very reliably: their reports on the sensory cues themselves will be free from similar interpretive errors.

On the other hand, one could say that the mindreading mechanism is not only charged with interpreting sensory cues which indicate that one has made a decision, but that it is also involved in generating the sensory or sensory-like evidential basis itself. Since, on this reading, the process of generating sensory cues is itself influenced by interpretation, one should predict that it will be susceptible to the usual errors of interpretation. Then it should be possible to generate not only false judgements about the will, but also illusory feelings about the will. This would be to suggest that a higher-order process penetrates a lower-order process: mindreading penetrates the generation of sensory-like states. It is common in the literature to distinguish 'feelings of agency' and 'judgements of agency' (Saito et al. 2015). In these terms, the question is: does interpretation penetrate to the level of feelings of agency or merely to the level of judgements of agency?

For the ISA theory, this is a question of how one should go beyond its basic tenets. As noted above, Cassam seems to suggest that the theory should endorse cognitive penetration of this sort generally, not only for decisions (Cassam 2014). He seems to suggest that knowledge of one's own sensory states as well as one's own attitudinal states is permeated with interpretation, or in his own terms, that self-knowledge is 'inferential' through and through.

It may be that many other researchers would be ready to claim that a lot of perception is also inferential in some sense (Bayne, Haggard, Sinnott-Armstrong, Vierkant, personal communication, 2019). However, it is one thing to say that sensory processes are inferential, and another thing to say that they are interpretive, in the sense that our knowledge of them depends on mindreading mindreading, or that the latter is even involved in generating sensations. The latter is a much stronger theses. It is not completely clear, which of these two claims the extended ISA theory should adopt. Does the available empirical evidence suggest that one should develop it in one of these directions?

There is some evidence that suggests that feelings of agency, as opposed to judgements of agency, are affected by such external factors as whether the agent succeeded to achieve their goal (Moore et al. 2009). James Moore and colleagues found that priming participants with the end-state of their action increased the sense of control over the movement for both voluntary and involuntary movements. Crucially, Moore and colleagues used an indirect measure of the sense of control: they did not ask the subject to directly report on it. This makes it more likely that what one is getting at here is the feeling rather than the judgement.

There is also some evidence suggesting that the sense of control depends on whether other agents succeed in achieving their goal if the primary agent considers them to be part of their own group and striving for the same goal (Dewey et al. 2014). John Dewey and colleagues report that when a group of participants was able to reach their common goal, individual participants attributed more control to themselves in particular. Crucially, this was not because the participants were attributing some of their partner's contribution to themselves. In this respect, they were as generous as before. These results suggest that reaching 'our' goal independently contributes to the feeling that 'I' am in control. In many ways, this experiment is similar to the study by Moore and colleagues. However, Dewey and colleagues asked for explicit judgements about agency. For the purposes of testing the present hypothesis, it would be more interesting to see whether similar results would be found if one changed the experiment so that participants would be asked about their contribution indirectly.

To sum up, the idea that interpretation penetrates to the level of feelings of agency has already garnered some support. However, more research is needed to convincingly answer the question whether the feeling of agency comes to us as a mixture of feeling and judgement.

6.5. Conclusion

The ISA theory receives support from empirical research on free will. These studies have probed people's attributions of decisions to themselves. They suggest that these attributions depend on both internal and external evidence, such as evidence on one's overt behaviour. The evidence challenges its rivals, since they claim these attributions do not depend on external behavioural cues. The evidence supports the ISA theory, since it claims that these attributions depend on perceptual feedback on behaviour as well as mental imagery. The suggestion that mental imagery plays a role in attributions of decisions to oneself is yet to be directly tested in future studies. One avenue would be to look at people with aphantasia.

CONCLUSION

1. The ISA theory of self-knowledge focuses on self-attribution of attitudes, claims it is done by turning the mindreading faculty onto oneself, predicts it will result in misattribution when sensations are misleading, and implies that there are no conscious decisions.

1.1. The theory has a broad target, but its focus is quite narrow. The target is self-knowledge. The focus is not the body, but the mind, not the complex features, but the simple features, not the standing states, but the occurrent states, and not the sensations, but the attitudes. Likewise, the focus is not knowledge, much less privileged, peculiar, and authoritative knowledge, but attribution. The point of focus is self-attribution of judgement and decision. This point of focus is more basic than the others, except for self-attribution of sensations, and the latter is given a relatively simple and uncontroversial explanation by the theory.

1.2. The theory makes claims that make it rival all other theories of self-knowledge, but it has one most important rival. It is not the other versions of the symmetrical theory, but the asymmetrical theory, and not all versions of the asymmetrical but the inner sense theory. The inner sense theory claims that self-attribution of attitudes is done by the introspection faculty. This theory contrasts with all the main claims of the ISA theory and it is the strongest competitor for empirical support.

1.3. The ISA theory makes predictions that make it possible to pit it against all of its main rivals in empirical research, and here one prediction is central. It predicts that people will misattribute attitudes to themselves when given cues analogous to the ones that mislead them about others. It pits the theory against all those claiming that mindreading is not the only way of finding out about one's own attitudes and that therefore such cues should not be sufficient to mislead people about themselves.

1.4. The theory has many wide implications, but one of them gives rise to more debate than others. It claims that there is no non-interpretive access to one's decisions. If non-interpretive access to one's decisions is necessary for conscious decisions, then there are no conscious decisions. If conscious decisions are necessary for free will, then there is no free will. Whether interpretive-sensory access to one's decisions is sufficient for free will is the question that probably receives the most attention from those who think that the ISA theory might well be right.

2. The empirical evidence that emerged during the first decade since the ISA theory entered the scene generally continues to support the ISA theory

and challenges its rivals.

2.1. The evidence on non-sensory awareness continues to suggest that those who report non-sensory awareness fail to report the available sensory cues. One new source of support is evidence that sensory cues are plentiful even in meditative cases. Another is evidence that even experts' reports of non-sensory awareness are not very consistent.

2.2. The evidence on childhood development continues to suggest that mindreading comes first. One new challenge is evidence that studies of early implicit attribution of false belief to others are not replicating. One response is to say that this affects only some of the paradigms. Another is to say that studies of explicit attribution of false belief would put attribution to oneself and others at the same time in development. Another new challenge is evidence of early implicit attribution of uncertainty to oneself. One response is to say that it would put attribution of uncertainty to oneself at the same time as attribution of uncertainty to others if some of the studies of early implicit attribution of attitudes to others would remain unaffected. Another is that the implicit attribution tests that were used here do not measure meta-representational abilities, which are the only ones that concern our theory.

2.3. The evidence on dissociations continues to suggest that mindreading and metacognition go together. One new source of support is evidence that, in autism, impairment in explicit metacognition is related to impairment in explicit mindreading, not in implicit metacognition. One new challenge is to explain what the impairment is in the case of the patient with both anarchic hand syndrome and utilisation behaviour. One response is to say that the case is equally puzzling for the main rival theories. Another new challenge is to show that sensory brain areas activated during metacognition are involved not merely causally, but computationally.

2.4. The evidence on metacognition continues to suggest that monitoring is not very reliable and control is broadly behavioural. One new source of support is evidence that situational strategies are central to self-control. Another is evidence that even solitary metacognition is reported as difficult. Yet another is the point that something cannot be both evaluative and under direct voluntary control, while decisions and judgements are evaluative by definition.

2.5. The evidence on misattribution continues to suggest that people are misled by cues that are analogous to those that would mislead others. One new challenge is the suggestion that it might be explained by the desire to explain one's own attitudes knowledgeably and with reference to reasons. One response is to say that the desire's influence is compatible with the theory, but the desire alone cannot explain misattributions of brute causes and of reasons that parallel those misattributed to others. Another new challenge is evidence that studies of misattribution of decisions that were originally appealed to have methodological problems. One response is to say that there are now stronger studies that make the same point.

2.6. The comparative evidence continues to suggest that mindreading comes first. One new source of support is the theory of questioning attitudes as sui generis first-order attitudes, which helps explain away some of the problematic evidence. One new challenge is evidence that great apes attribute false beliefs to others but not to themselves. One response is to say that the mindreading faculty is not repurposed immediately. Another is to concede the point but note that this only challenges one the less central claims of the theory.

3. Explanatory considerations continue to suggest that one should choose the ISA theory over its rivals, because it is scientifically fruitful, relatively simple, and externally coherent, even if not necessarily intuitive.

3.1. The theory is scientifically fruitful. It was already predictable upon its introduction, partly because more sceptical theories like that have driven empirical research on self-knowledge before, and partly because it provided a large number of explicit predictions while most of its rivals did not. It is even more evident now since some of those predictions contributed to new empirical research on self-knowledge intuitions and mentalisation in autism.

3.2. The theory is relatively simple. It explains the cases of self and other, and of typical and interpretive attribution in the same way. One challenge is to show that it also gives a relatively simple explanation of self-attribution of sensations and attitudes. One response is to say that it should explain them differently, and it can do it without postulating any new entities that are not already postulated by theories of mindreading. Its rivals, like the unified transparency theory, ad minimum postulate new processing rules.

3.3. The theory is externally coherent. It receives indirect support from global workspace, working memory, and Machiavellian intelligence theories. It also enters larger frameworks of how the mind works, like dual-processing theory, without complicating the overall picture. One challenge is to show that this is so in the case of the predictive processing framework. One response is to say that humans evolved to make rigid predictions in the pattern that the ISA theory describes.

3.4. The theory needed to better explain why its counter-intuitiveness would not be an argument against it. On the original explanation, it was said that the (incorrect) assumption of transparency simplifies attribution without loss of accuracy. The challenge is to explain why the (correct) assumption of knowledge would not suffice for that end, or to show that the ISA theory is not counter-intuitive.

4. The ISA theory receives support from empirical research on intuition since they both suggest that the acquisition of self-knowledge is either unconscious, or conscious and interpretive.

4.1. Other things being equal, one should choose the more intuitive theory. Intuition is the pre-reflective attitude that can be elicited by asking for quick answers, first impressions, or by asking non-specialists. One might think that a more intuitive theory has an epistemic or a pragmatic advantage. It is clearly in a better position dialectically. For these reasons, it is rational (ceteris paribus) to choose the more intuitive theory over its rival.

4.2. Laypeople's intuition is in favour of the theory. The evidence on laypeople's intuition suggests that they intuitively think that people usually know their own attitudes but also make mistakes about them. The theory is equally supported by both of these findings while its rivals have a harder time explaining the intuitiveness of mistakes.

4.3. Some experts' intuition is also in favour of the theory. Many experts note, more or less explicitly, that the acquisition of self-knowledge is mostly 'silent' or unconscious and that at other times it is experienced as conscious interpreting. Since the theory claims that the acquisition of self-knowledge is a mostly unconscious interpretive process, it fits the description. Since its rivals claim that there is something more to the experience, they do not fit the description.

4.4. Self-knowledge intuitions could be further tested drawing on new predictions drawn from the theory and from the lessons from empirical research on free will intuitions. One such lesson is that it is crucial to fine-tune one's tools so that the test pits one rival theory against another. In the case of self-knowledge, rivals disagree on how attitudes are known, not whether they are usually known, and not whether mistakes about them are possible. Another is that one might expect more robust beliefs about self-knowledge to be related to more robust beliefs about free will, moral responsibility, and punishment.

5. The ISA theory receives support from empirical research on meditation since they both suggest that meditators misattribute attitudes to themselves and find them difficult to control.

5.1. Meditation is a test case for theories of self-knowledge. It is the case where one has only a minimal amount of behavioural cues. Here, the ISA theory predicts that monitoring will still be prone to mistakes, since it will remain interpretive, and control will still be hard, since it will remain broadly behavioural. Its rivals predict that monitoring will not be prone to mistakes and control will now be easy, since behavioural cues were distractions.

5.2. The evidence on monitoring in meditators suggests that they make mistakes about their own attitudes. They make these mistakes in solitude, in a resting state, and thinking their responses will not be known to others. This puts the pressure on anyone claiming that people only mistake their attitudes in unfavourable circumstances.

5.3. The evidence on control in meditators suggests that they find it hard to control their attitudes. One source of support is evidence that experts in formal Buddhist and psychotherapeutic meditation suggest that meditators actually find their mind difficult to control. Another is evidence that people generally find the task of entertaining themselves with their own thoughts to be difficult and unpleasant.

5.4. Self-knowledge in meditators could be further tested by drawing on new predictions of the ISA theory. They include the predictions that more experienced meditators will find the symmetrical theory more intuitive, that they will report more sensory inner experience, and perhaps also that they will have more coherent explicit and implicit attitudes towards themselves.

6. The ISA theory receives support from empirical research on free will since they both suggest that self-attribution of decisions depends on both internal and external evidence.

6.1. Free will experiments are a test for theories of self-knowledge. The ISA theory predicts that people will self-attribute decisions based on both internal evidence (mental imagery) and external evidence (perceived behaviour). Its rivals predict that people will not self-attribute attitudes based on perceived behaviour.

6.2. The evidence on external factors influencing decision self-attribution suggests that it is sometimes based on perceived behaviour. One source of support is evidence that reported time when one first became aware of one's decision shifts with delayed perceptual feedback. Another is evidence that perceptual feedback and a background story can mislead one into thinking that one did not make a decision when one did.

6.3. The evidence on internal factors influencing decision self-attribution suggests that it is sometimes based on something that is available in absence of perceived behaviour. One source of support for this claim is evidence that people self-attribute decisions that they do not implement. Another is that people (with tics) fail to self-attribute decisions in presence of prior thoughts about moving and the perceived movement.

6.4. Self-knowledge of decisions could be further tested drawing on new predictions of the ISA theory. One new prediction is that mental imagery will play an important role in decision self-attribution, so one might expect to find certain peculiarities in people with aphantasia. Another is that perhaps one could expect cases of illusory feelings of agency, not only of mistaken judgements of agency.

BIBLIOGRAPHY

Allen, T. & May, J., 2014. Does Opacity Undermine Privileged Access? *Philosophical Studies* 22(4): 617–629. https://doi.org/10.1080/09672559.2014.948714

Antony, L., Rey, G., 2016. Philosophy and Psychology. In: *The Oxford Handbook of Philosophical Methodology*. Eds. H. Cappelen, T. S. Gendler, J. Hawthorne. Oxford: Oxford University Press, 554–585. https://doi.org/10.1093/oxfordhb/9780199668779.013.36

Armstrong, D. M., 1968/2002. *A Materialist Theory of Mind*. 2nd edition. London: Routledge & Kegan Paul. https://doi.org/10.4324/9780203003237

Baars, B. J., 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baddeley, A. D., Hitch, G., 1974. Working Memory. *The Psychology of Learning and Motivation: Advances in Research and Theory* 8: 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Baillargeon, R., Buttelmann, S., Southgate, V., 2018. Interpreting Failed Replications of Early False-Belief Findings: Methodological and Theoretical Considerations. *Cognitive Development* 46: 112–124. https://doi.org/10.1016/j.cogdev.2018.06.001

Banks, W. P., Isham, E., A., 2009. We Infer Rather Than Perceive the Moment We Decided to Act. *Psychological Science* 20(1): 17–21. https://doi.org/10.1111/j.1467-9280.2008.02254.x

Bar-On, D., 2015. Transparency, Expression, and Self-Knowledge. *Philosophical Explorations* 18(2): 134–152. https://doi.org/10.1080/13869795.2015.1032334

Bar-On, D., 2004. *Speaking My Mind: Expression and Self-Knowledge.* Oxford*:* Clarendon Press. https://doi.org/10.1093/0199276285.001.0001

Bayne, T., Montague, M., eds., 2011. *Cognitive Phenomenology.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199579938.001.0001

Bem, D. J., 1967. Self-Perception: An Alternative Explanation of Cognitive Dissonance Phenomena. *Psychological Science* 74(3): 183–200. http://dx.doi.org/10.1037/h0024835

Bermúdez, J. L., 2013. *The Opacity of Mind: An Integrative Theory of Self Knowledge*, by Peter Carruthers. *Mind* 122: 263–266. https://doi.org/10.1093/mind/fzt025

Bilgrami, A., 2006. *Self-Knowledge and Resentment*. Cambridge, Massachusetts: Harvard University Press.

Boghossian, P. A., 2006. *Fear of Knowledge: Against Relativism and Constructivism.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199287185.001.0001

Bourget, D., Chalmers, D. J., 2014. What Do Philosophers Believe? *Philosophical Studies* 170: 465–500. https://doi.org/10.1007/s11098-013-0259-7

Boyle, M., 2009. Two Kinds of Self-Knowledge. *Philosophy and Phenomenological Research* 78(1): 133-164. https://doi.org/10.1111/j.1933-1592.2008.00235.x

Brasil-Neto, J. P., Pascual-Leone, A., Valls-Solé, J., Cohen, L. G., Hallett, M., 1992. Focal Transcranial Magnetic Stimulation and Response Bias in a Forced-Choice Task. *Journal of Neurology, Neurosurgery, and Psychiatry* 55(10) : 964–6. https://doi.org/10.1136/jnnp.55.10.964

Breyer, T., Gutland, C., eds., 2015. *The Phenomenology of Thought: Philosophical Investigations into the Character of Cognitive Experiences.* London: Routledge.

Briñol, P., Petty, R. E., 2003. Overt Head Movements and Persuasion: A Self-Validation Analysis. *Journal of Personality and Social Psychology* 84(6): 1123–1139. https://psycnet.apa.org/record/2003-00779-004

Burge, T., 2013. *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection*. Philosophical Essays, Volume 3. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199672028.001.0001

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J. & Tomasello, M., 2017. Great Apes Distinguish True from False Beliefs in an Interactive Helping Task. *PLoS ONE* 12(4): e0173793. https://doi.org/10.1371/journal.pone.0173793

Byrne, A., 2018. *Transparency and Self-Knowledge.* Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198821618.001.0001

Byrne, A., 2012. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, by Peter Carruthers. *Notre Dame Philosophical Reviews.* https://ndpr.nd.edu/news/the-opacity-of-mind-an-integrative-theory-of-self-knowledge/

Byrne, R. W., Whiten, A., eds., 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Oxford: Oxofrd University Press.

Carr, M., 1993. 'Mind-Monkey' Metaphors in Chinese and Japanese Dictionaries. *International Journal of Lexicography* 6(3): 149–180. https://doi.org/10.1093/ijl/6.3.149

Carruthers, P., 2018. Implicit Versus Explicit Attitudes: Differing Manifestations of the Same Representational Structures? *Review of Philosophy and Psychology* 9(1): 51–72. https://doi.org/10.1007/s13164-017-0354-3

Carruthers, P., 2017b. Are Epistemic Emotions Metacognitive? *Philosophical Psychology* 30(1–2): 58–78. https://doi.org/10.1080/09515089.2016.1262536

Carruthers, P., 2017a. The Illusion of Conscious Thought. *Journal of Consciousness Studies* 24(9–10): 228–252. https://www.ingentaconnect.com/contentone/imp/jcs/2017/00000024/F0020009/art00013?crawler=true

Carruthers, P., 2015. *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198738824.001.0001

Carruthers, P., 2013b. Mindreading the Self. In *Understanding Other Minds* (third edition), eds. S. Baron-Cohen, H. Tager-Flusberg, M.V. Lombardo. Oxford: Oxford University Press, 467–486. https://doi.org/10.1093/acprof:oso/9780199692972.003.0026

Carruthers, P., 2013a. On Knowing Your Own Beliefs: A Representationalist Account. In: *New Essays on Belief: Constitution, Content, and Structure*, ed. N. Nottelmann, Palgrave Macmillan, 145–165. https://doi.org/10.1057/9781137026521_8

Carruthers, P., 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199596195.001.0001

Carruthers, P., 2010. Introspection: Divided and Partly Eliminated. *Philosophy and Phenomenological Research* 80(1): 76–111. https://doi.org/10.1111/j.1933-1592.2009.00311.x

Carruthers, P., 2009. How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition. *Behavioural and Brain Sciences* 32(2): 1–18. https://doi.org/10.1017/S0140525X09000545

Carruthers, P., 2008. Cartesian Epistemology: Is the Theory of the Self-Transparent Mind Innate? *Journal of Consciousness Studies* 15(4): 28–53. https://www.ingentaconnect.com/content/imp/jcs/2008/00000015/00000004/art00002

Carruthers, P., 2007. The Illusion of Conscious Will. *Synthese* 159: 197–213. https://doi.org/10.1007/s11229-007-9204-7

Carruthers, P., 2006. Conscious Experience Versus Conscious Thought. In:, *Self-Referential Approaches to Consciousness*, eds. U. Kriegel & K. Williford. Cambridge, Massachusetts: The Massachusetts Institute of Technology Press, 299–320.

Carruthers, P., Ritchie, J. B., 2012. The Emergence of Metacognition: Affect and Uncertainty in Animals. In: *Foundations of Metacognition*, eds. M. J. Beran, J. L. Brandl, J. Perner, J. Proust. Oxford: Oxford University Press, 191–234.
https://doi.org/10.1093/acprof:oso/9780199646739.003.0006

Cassam, Q., 2017. What Asymmetry? Knowledge of Self, Knowledge of Others, and the Inferentialist Challenge. *Synthese* 194: 723–741.
https://doi.org/10.1007/s11229-015-0772-7

Cassam, Q., 2014. *Self-Knowledge for Humans.* Oxford: Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780199657575.001.0001

Clark, A., 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780190217013.001.0001

Clark, A., 2013. Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioural and Brain Sciences* 36(3): 181–204.
https://doi.org/10.1017/S0140525X12000477

Clark, C. J., Ditto, P. H., Shariff, A. F., Luguri, J. B., Knobe, J., Baumeister, R. F., 2014. Free To Punish? A Motivated Account of Free Will Belief. *Journal of Personality and Social Psychology* 106(4): 501–513.
http://dx.doi.org/10.1037/a0035880

Clark, C. J., Baumeister, R. F., Ditto, P. H., 2017. Making Punishment Palatable: Belief in Free Will Alleviates Punitive Distress. *Consciousness and Cognition* 51: 193–211.
https://doi.org/10.1016/j.concog.2017.03.010

Clark, C. J., Chen, E. E., Ditto, P. H. 2015. Moral Coherence Processes: Constructing Culpability and Consequences. *Current Opinion in Psychology* 6: 123–128.

https://doi.org/10.1016/j.copsyc.2015.07.016

Clark, C. J., Shniderman, A., Luguri, J. B., Baumeister, R. F., Ditto, P. H., 2018. Are Morally Good Actions Ever Free? *Consciousness and Cognition* 63: 161–182.
https://doi.org/10.1016/j.concog.2018.05.006

Coliva, A., 2016. *The Varieties of Self-Knowledge.* London: Palgrave Macmillan.
https://doi.org/10.1057/978-1-137-32613-3

Couchman, J., Coutinho, M., Beran, M., Smith, D., 2009. Metacognition is prior. *Behavioral and Brain Sciences* 32(2): 142.
https://doi.org/10.1017/S0140525X09000594

Creswell, J. D., 2017. Mindfulness Interventions. *Annual Review of Psychology* 68: 491–516.
https://doi.org/10.1146/annurev-psych-042716-051139

Descartes, R., 1637/2006. *A Discourse on the Method.* Trans. I. McLean. Oxford: Oxford University Press.
https://doi.org/10.1093/acref/9780199399680.013.0320

Descartes, R., 1641/2008. *Meditations on First Philosophy with Selections from Objections and Replies.* Trans. M. Moriarty. Oxford: Oxford University Press.

Desmurget, M., Reilly, K. T., Richard, N., Szathmari, A., Mottolese, C., Sirigu, A., 2009. Movement Intention after Parietal Cortex Stimulation in Humans. *Science* 324(5928): 811–813.
https://doi.org/10.1126/science.1169896

Dewey, J. A., Pacherie, E., Knoblich, G., 2014. The Phenomenology of Controlling a Moving Object with Another Person. *Cognition* 132: 383–397.
https://doi.org/10.1016/j.cognition.2014.05.002

Dokic, J., 2012. Seeds of Self-Knowledge: Noetic Feelings and Metacognition. In: *Foundations of Metacognition*, eds. M. J. Beran, J. L. Brandl, J. Perner, J. Proust. Oxford: Oxford University Press, 716–761.
https://doi.org/10.1093/acprof:oso/9780199646739.003.0020

Doris, J. M., 2015. *Talking to Ourselves: Reflection, Ignorance, and Agency.* Oxford: Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780199570393.001.0001

Duckworth, A. L., Gendler, T. S., Gross, J. J., 2016. Situational Strategies for Self-Control. *Perspectives on Psychological Science* 11(1): 35–55.
https://doi.org/10.1177/1745691615623247

Dutton, D. G., Aron, A. P., 1974. Some Evidence for Heightened Sexual Attraction Under Conditions of High Anxiety. *Journal of Personality and Social Psychology* 30(4) : 510–507. https://doi.org/10.1037/h0037031

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

Evans, J. St. B. T., Stanovich, K., 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science* 8(3) 223–241. https://doi.org/10.1177/1745691612460685

Fernández, J., 2013. *Transparent Minds: A Study of Self-Knowledge.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199664023.001.0001

Frankish, K. 2016. Playing Double: Implicit Bias, Dual Levels, and Self-Control. In: *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*, eds. M. Braunstein, J. Saul. Oxford: Oxford University Press, 23–46. https://doi.org/10.1093/acprof:oso/9780198713241.003.0002

Frankish, K., 2012. Dual Systems and Dual Attitudes. *Mind and Society* 11(1): 41–51. https://doi.org/10.1007/s11299-011-0094-5

Frankish, K., 2009. How We Know Our Own Minds: Introspective Access to Conscious Thoughts. *Behavioural and Brain Sciences* 32(2): 25–26. https://doi.org/10.1017/S0140525X09000636

Frankish, K., 2004. *Mind and Supermind.* Cambridge: Cambridge University Press.

Fricke, M. F., 2014. Transparency or Opacity of Mind? *Contributions to the Austrian Ludwig Wittgenstein Society* 22: 97–99.

Fried, I., Katz, A., McCarthy, G., Sass, K. J., Williamson, P., Spencer, S. S., Spencer, D. D., 1991. Functional Organization of Human Supplementary Motor Cortex Studied by Electrical Stimulation. *The Journal of Neuroscience* 11(11): 3656–3666. https://doi.org/10.1523/JNEUROSCI.11-11-03656.1991

Garfield, J., 2015. *Engaging Buddhism: Why It Matters to Philosophy*. Oxford: Oxford University Press.

Ganos, C., Rothwell, J. & Haggard, P., 2018. Voluntary Inhibitory Motor Control Over Involuntary Tic Movements. *Movement Disorders* 33(6): 937–946. https://doi.org/10.1002/mds.27346

Gazzaniga, M. S., 1998. *The Mind's Past*. Berkeley: University of California Press.

Genschow, O., Rigoni, D., Brass, M., 2019. The Hand of God or the Hand of Maradona? Believing in Free Will Increases Perceived Intentionality of Others' Behavior. *Consciousness and Cognition* 70: 80–87. https://doi.org/10.1016/j.concog.2019.02.004

Gertler, B., 2015. Self-Knowledge. In: *Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. https://plato.stanford.edu/archives/win2019/entries/self-knowledge

Gertler, B., 2011. *Self-Knowledge.* London: Routledge.

Goldman, A. I., 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading.* Oxford: Oxford University Press. https://doi.org/10.1093/0195138929.001.0001

Goldman, A. I., Jordan, L. C. 2013. Mindreading by Simulation: The Roles of Imagination and Mirroring. In: *Understanding Other Minds* (third edition), eds. S. Baron-Cohen, H. Tager-Flusberg, M. V. Lombardo. Oxford: Oxford University Press, pp. 448–466. https://doi.org/10.1093/acprof:oso/9780199692972.003.0025

Gopnik, A., 1993. How We Know Our Own Minds: The Illusion of First Person Knowledge of Intentionality. *Behavioural and Brain Sciences* 16: 1–14. https://doi.org/10.1017/S0140525X00028636

Goupil, L. & Kouider, S., 2019. Developing a Reflective Mind: From Core Metacognition to Explicit Self-Reflection. *Current Directions in Psychological Science* 28(4): 403–408. https://doi.org/10.1177/0963721419848672

Goupil, L. & Kouider, S., 2016. Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal Infants. *Current Biology*: 26(22): 3038–3045. https://doi.org/10.1016/j.cub.2016.09.004

Goupil, L., Romand-Monnier, M. & Kouider, S., 2016. Infants Ask for Help When They Know They Don't Know. *PNAS* 113(13): 3492–3496. https://doi.org/10.1073/pnas.1515129113

Guerini, R., Marraffa, M., Paloscia, C., 2015. Mentalisation, Attachment, and Subjective Identity. *Frontiers in Psychology* 6: art. 1022. https://doi.org/10.3389/fpsyg.2015.01022

Haggard, P., 2019. The Neurocognitive Bases of Human Volition. *Annual Review of Psychology* 70: 9–28. https://doi.org/10.1146/annurev-psych-010418-103348

Haley, K. J, Fessler, D. M. T., 2005. Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game. *Evolution and Human Behavior* 26: 245–256. https://doi.org/10.1016/j.evolhumbehav.2005.01.002

Hall, L., Johansson, P., Strandberg, T., 2012. Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PloS ONE* 7(9): e45457. https://doi.org/10.1371/journal.pone.0045457

Hall, L., Johansson, P., Tärning, B., Sikström, S., Deutgen, T., 2010. Magic at the Marketplace: Choice Blindness for the Taste of Jam and the Smell of Tea. *Cognition* 117(1): 54–61. https://doi.org/10.1016/j.cognition.2010.06.010

Hurlburt, R. T., 2011. *Investigating Pristine Inner Experience.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511842627

Hurlburt, R. T., Heavey, C. L., 2006. *Describing Inner Experience.* Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/aicr.64

Hurlburt, R. T., Alderson-Day, B., Fernyhough, C., Kühn, S., 2015. What Goes on in a Resting State? A Qualitative Glimpse into Resting State Experience in the Scanner. *Frontiers in Psychology* 6: 1535. https://doi.org/10.3389/fpsyg.2015.01535

Hurlburt, R. T., Akhter, S. A., 2008. Unsymbolized thinking. *Consciousness and Cognition* 17: 1364–1374. https://doi.org/10.1016/j.concog.2008.03.021

Johansson, P., Hall, L., Sikström, S., Olsson, A., 2005. Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science* 310: 116–119. https://doi.org/10.1126/science.1111709

Kant, I., 1781/1997. *The Critique of Pure Reason*. Eds. trans. Paul Guyer & Allen W. Wood. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804649

Kant, I., 1784/1996. Answer to the Question: What is Enlightenment? In: *Practical Philosophy,* trans. ed. M. J. Gregor. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511813306

Keeling, S., 2018. Confabulation and Rational Obligations for Self-Knowledge. *Philosophical Psychology* 31(8): 1215–1238. https://doi.org/10.1080/09515089.2018.1484086

King, M., Carruthers, P., forthcoming. Responsibility and Consciousness. In: *Handbook of Moral Responsibility,* eds. D. Nelkin, D. Pereboom. https://faculty.philosophy.umd.edu/pcarruthers/#

King, M., Carruthers, P. 2012. Moral Responsibility and Consciousness. *Journal of Moral Philosophy* 9: 200–228. https://doi.org/10.1163/174552412X625682

Knappik, F., 2015. Self-Knowledge about Attitudes: Rationalism Meets Interpretation. *Philosophical Explorations* 18(2): 183–198. https://doi.org/10.1080/13869795.2015.1032328

Kornell, N., Son, L. K., Terrace, H. S., 2007. Transfer of Metacognitive Skills and Hint Seeking in Monkeys. *Psychological Science* 18(1): 64–71. https://doi.org/10.1111/j.1467-9280.2007.01850.x

Kozuch, B. & Nichols, S., 2011. Awareness of Unawareness: Folk Psychology and Introspective Transparency. *Journal of Consciousness Studies* 18(11–12): 135–160. https://www.ingentaconnect.com/contentone/imp/jcs/2011/00000018/F0020011/art00006

Krachun, C., Call, J., Tomasello, M., 2009. Can Chimpanzees (*Pan troglodytes*) Discriminate Appearance from Reality? *Cognition* 112(3): 435–50. https://doi.org/10.1016/j.cognition.2009.06.012

Krupenye, C., Kano, F., Hirata, S., Call, J. & Tomasello, M., 2016. Great Apes Anticipate that Other Individuals Will Act According to False Beliefs. *Science* 354(6308): 110–114. https://doi.org/10.1126/science.aaf8110

Lakatos, I., 1970. Falsification and the Methodology of Scientific Research Programmes. In: *Criticism and the Growth of Knowledge*, eds. I. Lakatos, A. Musgrave. Cambridge: Cambridge University Press, 91–196. https://doi.org/10.1007/978-94-010-1863-0_14

Levy, N., 2014. Consciousness, Implicit Attitudes, and Moral Responsibility. *Noûs* 48(1): 21–40. https://doi.org/10.1111/j.1468-0068.2011.00853.x

Levy, N., 2012. A Role for Consciousness After All. *Journal of Moral Philosophy* 9(2): 255–264. https://doi.org/10.1163/174552411X612092

Libet, B., Gleason, C. A., Wright, E. W., Pearl, D. K., 1983. Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential): The Unconscious Initiation of a Freely Voluntary Act. *Brain: A Journal of Neurology* 106: 623–642.

https://doi.org/10.1093/brain/106.3.623

Locke, J., 1689/1975. *An Essay Concerning Human Understanding*. Ed. P. H. Nidditch. Oxford: Oxford University Press.

Lombardo, M. V., Chakrabarti, B., Baron-Cohen, S., 2009. What Neuroimaging and Perceptions of Self-Other Similarity Can Tell Us About the Mechanism Underlying Mentalizing. *Behavioral and Brain Sciences* 32(2): 152–153. https://doi.org/10.1017/S0140525X09000715

Lycan, W. G., 1996. *Consciousness and Experience*. Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

Lyons, W., 1986. *The Disappearance of Introspection*. Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

Machery, É., 2017. *Philosophy Within Its Proper Bounds*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198807520.001.0001

Marraffa, M., 2014. The Unconscious, Self-Consciousness, and Responsibility. *Rivista Internazionale di Filosofia e Psicologia* 5(2): 207–220. https://doi.org/10.4453/rifp.2014.0016

Matsugasaki, K., Tsukamoto, W., Ohtsubo, Y., 2015. Two Failed Replications of the Watching Eyes Effect. *Letters on Evolutionary Behavioral Science* 6(2): 17–20. https://doi.org/10.5178/lebs.2015.36

Matsuhashi, M. & Hallett, M., 2008. The Timing of the Conscious Intention to Move. *European Journal of Neuroscience* 28(11): 2344–2351. https://doi.org/10.1111/j.1460-9568.2008.06525.x

McGlynn, A., 2012. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, by Peter Carruthers. *The Philosophical Quarterly* 62(248): 635–637. https://doi.org/10.1111/j.1467-9213.2012.00051.x

Mercier, H., Sperber, D., 2017. *The Enigma of Reason.* Cambridge, Massachusetts: Harvard University Press.

Moore, J. W., Wegner, D. M., Haggard, P., 2009. Modulating the Sense of Agency with External Cues. *Consciousness and Cognition* 18(4): 1056–1064. https://doi.org/10.1016/j.concog.2009.05.004

Moran, R., 2017b. Frankfurt on Identification: Ambiguities of Activity in Mental Life. In: *The Philosophical Imagination: Selected Essays.* Oxford: Oxford University Press, 136–157.

https://doi.org/10.1093/acprof:oso/9780190633776.003.0008

Moran, R., 2017a. Self-Knowledge, "Transparency", and the Forms of Agency. In: *The Philosophical Imagination: Selected Essays.* Oxford: Oxford University Press, 275–296. https://doi.org/10.1093/acprof:oso/9780190633776.003.0015

Moran, R., 2001. *Authority and Estrangement: An Essay on Self-Knowledge.* Princeton: Princeton University Press.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., Ross, L. T., 2014. The Free Will Inventory: Measuring Beliefs about Agency and Responsibility. *Consciousness and Cognition* 25: 27–41. https://doi.org/10.1016/j.concog.2014.01.006

Nagel, J., 2014. Intuition, Reflection, and the Command of Knowledge. *Proceedings of the Aristotelian Society Supplementary Volume* 88(1): 219–241. https://doi-org.eres.qnl.qa/10.1111/j.1467-8349.2014.00240.x

Nahmias, E., 2018. Your Brain as the Source of Free Will Worth Wanting: Understanding Free Will in the Age of Neuroscience. In: *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*, eds. G.G. Caruso, O. Flanagan. Oxford: Oxford University Press, 251–268. https://doi.org/10.1093/oso/9780190460723.003.0014

Nahmias, E., Morris, S. G., Nadelhoffer, T., Turner, J., 2005. Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility. *Philosophical Psychology* 18(5): 561–584. https://doi.org/10.1080/09515080500264180

Newen, A. 2015. Understanding Others: The Person Model Theory. *Open MIND* 26: 1–28. https://open-mind.net/papers/understanding-others-the-person-model-theory

Newton-Smith, W. H., 1981. *The Rationality of Science.* London: Routledge.

Nichols, S., forthcoming, Mindreading and the Philosophy of Mind. In: *The Oxford Handbook of Philosophy of Psychology*, ed. J. Prinz. Oxford: Oxford University Press.

Nichols, S., 2004. The Folk Psychology of Free Will: Fits and Starts. *Mind and Language* 19(5): 473–502. https://doi.org/10.1111/j.0268-1064.2004.00269.x

Nichols, S., Stich, S., 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds.* Oxford: Clarendon Press.

https://doi.org/10.1093/0198236107.001.0001

Nicholson, T., Williams, D. M., Grainger, C., Lind, S. E., Carruthers, P. 2019. Relationships Between Implicit and Explicit Uncertainty Monitoring and Mindreading: Evidence from Autism Spectrum Disorder. *Consciousness and Cognition* 70: 11–24. https://doi.org/10.1016/j.concog.2019.01.013

Nisbett, R. E., Schachter, S., 1966. The Cognitive Manipulation of Pain. *Journal of Experimental Social Psychology* 2(3): 227–236. https://dx.doi.org/10.1016/0022-1031(66)90081-3.

Nisbett, R. E., Wilson, T. D., 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review* 84(3): 231–259. http://dx.doi.org/10.1037/0033-295X.84.3.231

O'Brien, L., 2007. *Self-Knowing Agents*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199261482.001.0001

O'Shaughnessy, B., 2000. *Consciousness and the World.* Oxford: Clarendon Press. https://doi.org/10.1093/0199256721.001.0001

Olson, J. A., Landry, M., Appourchaux, K., Raz, A., 2016. Simulated Thought Insertion: Influencing the Sense of Agency Using Deception and Magic. *Consciousness and Cognition* 43: 11–26. https://doi.org/10.1016/j.concog.2016.04.010

Pascal, B., 1670/1995. Pensées. In: *Pensées and Other Writings,* trans. H. Levi, ed. A. Levi. Oxford: Oxford University Press.

Peters, U., 2018. Introspection, Mindreading, and Transparency of Belief. *European Journal of Philosophy* 26: 1–17. https://doi.org/10.1111/ejop.12318

Peters, U., 2014c. Self-Knowledge and Conscious Attitudes. J*ournal of Consciousness Studies* 21(1-2): 139–155. https://www.ingentaconnect.com/content/imp/jcs/2014/00000021/F0020001/art00008

Peters, U., 2014a. Conscious Propositional Attitudes and Moral Responsibility. *Journal of Moral Philosophy:* 11(5): 585–597. https://doi.org/10.1163/17455243-4681017

Peters, U., 2014b. Interpretive Sensory-Access Theory and Conscious Intentions. *Philosophical Psychology* 27(4): 583–595. https://doi.org/10.1080/09515089.2012.749560

Proust, J., 2016. The Evolution of Primate Communication and Metacommunication. *Mind and Language* 31(2): 177–203. https://doi.org/10.1111/mila.12100

Proust, J., 2013. *The Philosophy of Metacognition: Mental Agency and Self-Awareness.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199602162.001.0001

Proust, J., 2012. Metacognition and Mindreading: One or Two Functions? In: *Foundations of Metacognition*, eds. M. J. Beran, J. L. Brandl, J. Perner, J. Proust. Oxford: Oxford University Press: 234–251. https://doi.org/10.1093/acprof:oso/9780199646739.003.0015

Renz, U., ed., 2017. *Self-Knowledge: A History*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190226411.001.0001

Rey, G., 2013b. The Possibility of a Naturalistic Cartesianism Regarding Intuitions and Introspection. In: *Philosophical Methodology: The Armchair or the Laboratory?*, ed. M.C. Haug. London: Routledge.

Rey, G., 2013a. We Are Not All 'Self-Blind': A Defense of a Modest Introspectionism. *Mind and Language* 28(3): 259–285. https://doi.org/10.1111/mila.12018

Rey, G., 2012. Postscript to 'We Are Not All "Self-Blind": A Defense of a Modest Introspectionism'. Homepage of George Rey. https://sites.google.com/site/georgesrey/>.

Rey, G., 2011. Introspection, Inatentional Blindness and an Insufficient Inferential Base. Website of National Humanities Center's project *On The Human*. https://nationalhumanitiescenter.org/on-the-human/2011/10/knowledge-of-our-own-thoughts/.

Rey, G., 2008. (Even Higher-Order) Intentionality Without Consciousness. *Revue Internationale de Philosophie* 62: 51-78. https://doi.org/10.3917/rip.243.0051

Roessler, J., 2013. The Silence of Self-Knowledge. *Philosophical Explorations* 16(1): 1–17. https://doi.org/10.1080/13869795.2013.744084

Ryle, G., 1949/2009. *The Concept of Mind.* 60th anniversary edition. London: Routledge. https://doi.org/10.4324/9780203875858

Saito, N., Takahata, K., Murai, T., Takahashi, H., 2015. Discrepancy between explicit judgement of agency and implicit feeling of agency: Implications for sense of agency and its disorders. *Consciousness and Cognition* 37: 1–7. https://doi.org/10.1016/j.concog.2015.07.011

Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., Baumeister, R. F., Vohs, K. D. 2014. Free Will and Punishment; A Mechanistic View of Human Nature Reduces Retribution. *Psychological Science* 25(8): 1–8. https://doi.org/10.1177/0956797614534693

Shoemaker, S., 1994. Self-knowledge and 'inner sense'. *Philosophy and Phenomenological Research* 54: 249–314.

Schlegel, A., Alexander, P., Sinnott-Armstrong, W., Roskies, A., Tse, P. U., Wheatley, T., 2015. Hypnotizing Libet: Readiness Potentials with Non-Conscious Volition. *Consciousness and Cognition* 33: 196–203. https://doi.org/10.1016/j.concog.2015.01.002

Schönbaumsfeld, G., 2016. *The Illusion of Doubt.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/97801987

Schwengerer, L., unpublished. *A Unified Transparency Account of Self-Knowledge*. Doctoral dissertation, The University of Edinburgh.

Schwengerer, L., 2019. Self-Knowledge in a Predictive Processing Framework. *Review of Philosophy and Psychology* 10(3): 563–585. https://doi.org/10.1007/s13164-018-0416-1

Schwitzgebel, E., 2019. Introspection. In: *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. https://plato.stanford.edu/archives/win2019/entries/introspection/

Schwitzgebel, E., 2013. A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box. In: New Essays on Belief, ed. N. Nottelman. Dordrecht: Springer, 75–99. https://doi.org/10.1057/9781137026521_5

Schwitzgebel, E., 2012. Introspection, What? In: Introspection and Consciousness, eds. D. Smithies, D. Stoljar. https://doi.org10.1093/acprof:oso/9780199744794.003.0001

Schwitzgebel, E., 2011. *Perplexities of Consciousness.* Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

Scott, R. M., Baillargeon, R., 2017. Early False-Belief Understanding. *Trends in Cognitive Sciences* 21(4): 237–249. https://doi.org/10.1016/j.tics.2017.01.012

Segal, Z. V., Williams, J. M. G., Teasdale, J. D., 2013. *Mindfulness-Based Cognitive Therapy for Depression.* New York: Guilford Press.

Sellars, W., 1956. Empiricism and the Philosophy of Mind. *Minnesota Studies in the Philosophy of Science* 1: 253-329.

Serban, M., 2014. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, by Peter Carruthers. *Philosophical Psychology:* 27(6): 934–938.
https://doi.org/10.1080/09515089.2013.791747

Shepherd, J., 2013. The Apparent Illusion of Conscious Deciding. *Philosophical Explorations* 16(1): 18–30.
https://doi.org/10.1080/13869795.2013.723035

Sinnot-Armstrong, W., Nadel, L., eds., 2011. *Conscious Will and Responsibility.* Oxford: Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780195381641.001.0001

Smallwood, J., Schooler, J. W., 2013. The Restless Mind. *Psychology of Consciousness: Theory, Research, and Practice* 1(S): 130–149.
http://dx.doi.org/10.1037/2326-5523.1.S.130

Smilansky, S., 2002. *Free Will and Illusion.* Oxford: Clarendon Press.

Sperber, D., Mercier, H., 2010. Reasoning as a Social Competence. In: *Collective Wisdom: Principles and Mechanisms*, eds. H. Landemore, J. Elster. Cambridge: Cambridge University Press, 368–392.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1862669

Tanney, J., 2009. Rethinking Ryle: A Critical Discussion of *The Concept of Mind*. In: *The Concept of Mind.* 60th anniversary edition. London: Routledge.
https://doi.org/10.4324/9780203875858

Van Dam, N. T., Van Vugt, M. K., Vago, D. R., Schmalzl, L., Saron, C. D., Olendzki, A., Meissner, T., Lazar, S. W., Kerr, C. E., Gorchov, J., Fox, K. C. R., Field, B. A., Britton, W. B., Brefczynski-Lewis, J. A. & Meyer, D. E., 2017. Mind the Hype: A Critical Evaluation and Prescriptive Agenda for Research on Mindfulness and Meditation. *Perspectives on Psychological Science* 13(1): 36–61.
https://doi.org/1o0i.1or1g7/170/.1177475/61794156197176107975088995

Vazire, S. & Carlson, E. N., 2010. Self-Knowledge of Personality: Do People Know Themselves? *Social and Personality Psychology Compass* 4(8): 605–620.
https://doi.org/10.1111/j.1751-9004.2010.00280.x

Vierkant, T., forthcoming. *The Tinkering Mind.* Oxford University Press.

Vierkant, T., 2015. How Do You Know that You Settled a Question? *Philosophical Explorations* 18(2): 199–211.
https://doi.org/10.1080/13869795.2015.1032330

Walter, S., 2014. Willusionism, Epiphenomenalism, and the Feeling of Conscious Will. *Synthese* 191: 2215–2238.

https://doi.org/10.1007/s11229-013-0393-y

Wegner, D. M., 2002/2017. *The Illusion of Conscious Will.* New Edition. Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

Wegner, D. M., 1994. Ironic Processes of Mental Control. *Psychological Review* 101(1): 34–52. https://dx.doi.org/10.1037/0033-295X.101.1.34

Wegner, D. M., Wheatley, T., 1999. Apparent Mental Causation: Sources of the Experience of Will. *American Psychologist* 54(7): 480–492. https://dx.doi.org/10.1037/0003-066X.54.7.480

Wellman, D., Cross, D., Watson, J., 2001. Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development* 72(3): 655–684. https://doi.org/10.1111/1467-8624.00304

Westra, E., Carruthers, P., 2017. Pragmatic Development Explains the Theory-of-Mind Scale. *Cognition* 158: 165–176. https://doi.org/10.1016/j.cognition.2016.10.021

Williams, D. M., Nicholson, T., Grainger, C., Lind, S. L., Carruthers, P., 2018. Can You Spot a Liar? Deception, Mindreading, and the Case of Autism Spectrum Disorder. *Autism Research* 11: 1129–1137. https://doi.org/10.1002/aur.1962

Wilson, B. M., Mickes, L., Stolarz-Fantino, S., Evrard, M., Fantino, E., 2015. Increased False-Memory Susceptibility after Mindfulness Meditation. *Psychological Science* 26(10): 1567–1573. https://doi.org/10.1177/0956797615593705

Wilson, T. D., 2009. Know Thyself. *Perspectives on Psychological Science* 4(4): 384–389. https://doi.org/10.1111/j.1745-6924.2009.01143.x

Wilson, T. D., 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious.* Cambridge, Massachusetts: Harvard University Press.

Wilson, T. D., Dunn, D. S., Kraft, D., Lisle, D. J., 1989. Introspection, Attitude Change, and Attitude-Behavior Consistency: The Disruptive Effects of Explaining Why We Feel the Way We Do. *Advances in Experimental Social Psychology* 22: 287–343. https://doi.org/10.1016/S0065-2601(08)60311-1

Wilson, T. D., Reinhard, D. A., Westage, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., Brown, C. L. & Shaked, A., 2014. Just Think: The Challenges of the Disengaged Mind. *Science* 345(6192): 75–77. https://doi.org/10.1126/science.1250830

Wittgenstein, L., 1953/2009. *Philosophical Investigations*. 4th edition. Trans. G. E. M. Anscombe, P. M. S. Hacker, J. Schulte. Eds. P. M. S. Hacker, J. Schulte. Oxford: Blackwell.

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.
https://doi.org/10.1007/s11098-013-0169-8

Zeman, A., Dewar, M., Della Salla, S., 2015. Lives Without Imagery – Congenital Aphantasia. *Cortex* 73: 378–80.
https://doi.org/10.1016/j.cortex.2015.05.019

PUBLICATION LIST

Rimkevičius, P., forthcoming. The Interpretive-Sensory Access Theory of Self-Knowledge: Empirical Support and Scientific Fruitfulness. *Problemos* 97.

Rimkevičius, P., 2019. The Interpretive-Sensory Access Theory of Self-Knowledge: Simplicity and Coherence with Surrounding Theories. *Problemos* 96: 148-159. https://doi.org/10.15388/Problemos.96.12

Rimkevičius, P., 2016. Psychologism, Relativism, and Self-Refutation in Husserl's *Prolegomena*. *Problemos* 89: 153–166. https://doi.org/10.15388/Problemos.2016.89.9895

Rimkevičius, P. & Gutauskas, M., 2014. The Question of the Compatibility of Phenomenology and Naturalism in Shaun Gallagher's Philosophy. *Problemos* 86: 120–126. https://doi.org/10.15388/Problemos.2014.0.3950

ACKNOWLEDGEMENTS

I would like to express my gratitude to the people who have contributed to making this dissertation better. My supervisor, Marius Povilas Šaulauskas, and my consultant, Renatas Berniūnas, guided me throughout the writing process. Jérôme Dokic and Tillmann Vierkant supervised me during my stays in Paris and Edinburgh, and provided comments on early drafts. Jonas Dagys and Albinas Plėšnys wrote reviews for a later draft. Matthew Sims and Urtė Survilaitė commented on the last draft and helped me with the proofreading. Many others discussed the ideas with me, including: Tim Bayne, Marcel Brass, Peter Carruthers, Vilius Dranseika, Keith Frankish, Mindaugas Gilaitis, Patrick Haggard, Mark Hallett, Eve Isham, Sophie Keeling, Amy Kind, Uriah Kriegel, Jennifer Nagel, Eddy Nahmias, Jay Olson, Elisabeth Pacherie, Joëlle Proust, Aaron Schurger, Lukas Schwengerer, and Walter Sinnot-Armstrong. I am grateful to all of them. I would also like to express my gratitude to my former supervisor, Mintautas Gutauskas, who was in that capacity before I changed my dissertation topic. Finally, I thank my family and friends for supporting me throughout my studies.

NOTES

Vilniaus universiteto leidykla
Saulėtekio al. 9, LT-10222 Vilnius
El. p. info@leidykla.vu.lt,
www.leidykla.vu.lt
Tiražas 15 egz.