Vilnius University Press

Contents lists available at Vilnius University Press

Problemos
2020, vol. 97, pp. 150–163

ISSN 1392-1126 eISSN 2424-6158
DOI: https://doi.org/10.15388/Problemos.97.12

Pažinimo ir sąmonės filosofija / Philosophy of Mind

The Interpretive-Sensory Access Theory of Self-Knowledge: Empirical Adequacy and Scientific Fruitfulness

Paulius Rimkevičius

Institute of Philosophy
Vilnius University
E-mail paulius.rimkevicius@fsf.vu.lt
ORCID https://orcid.org/0000-0002-6172-3275

Abstract. The interpretive-sensory access theory of self-knowledge claims that we come to know our own minds by turning our capacities for knowing other minds onto ourselves. Peter Carruthers argues that two of the theory's advantages are empirical adequacy and scientific fruitfulness: it leaves few of the old discoveries unexplained and makes new predictions that provide a framework for new discoveries. A decade has now passed since the theory's introduction. I review the most important developments during this time period regarding the two criteria: whether the theory's six main predictions were supported, and whether the theory's predictions contributed to new empirical studies. I argue that the interpretive-sensory access theory of self-knowledge remains empirically adequate and scientifically fruitful.
Keywords: self-knowledge, Peter Carruthers, interpretive-sensory access, empirical adequacy, scientific fruitfulness

Interpretacinės-sensorinės prieigos savižinos teorija: empirinis adekvatumas ir mokslinis vaisingumas

Santrauka. Interpretacinės-sensorinės prieigos savižinos teorija teigia, kad mes sužinome savo pačių mintis nukreipdami į save pačius savo gebėjimus, skirtus kitų žmonių mintims sužinoti. Peteris Carruthersas įrodinėja, kad du iš šios teorijos pranašumų yra jos empirinis adekvatumas ir mokslinis vaisingumas: ji paaiškina ligtolinius atradimus ir pateikia naujų numatymų, kurie suteikia pagrindą naujiems atradimams. Nuo šios teorijos atsiradimo jau praėjo dešimtmetis. Šiame straipsnyje apžvelgiami svarbiausi tuo laikotarpiu pasirodę atradimai, susiję su minėtais dviem kriterijais: ar empirinių tyrimų rezultatai parėmė šešis pagrindinius šios teorijos numatymus ir ar ši teorija pateikė naujų numatymų, kurie prisideda prie naujų empirinių tyrimų. Šiame straipsnyje įrodinėjama, kad interpretacinės-sensorinės prieigos savižinos teorija tebėra empiriškai adekvati ir moksliškai vaisinga.
Pagrindiniai žodžiai: savižina, Peteris Carruthersas, interpretacinė-sensorinė prieiga, empirinis adekvatumas, mokslinis vaisingumas

Acknowledgement. I am grateful to Tim Bayne, Peter Carruthers, Jérôme Dokic, Patrick Haggard, Sophie Keeling, Jay Olson, Joëlle Proust, Walter Sinnott-Armstrong, Tillman Vierkant for discussing with me some of the ideas presented in this paper. I also thank the two anonymous reviewers, whose comments helped to express my ideas more clearly. The mistakes are all mine.

Received: 09/08/2019. **Accepted:** 18/10/2019

Introduction

How we know our own minds is widely debated in philosophy of mind and cognitive science. Some philosophers, such as Peter Carruthers, argue that we come to know our own minds by turning our capacities for knowing other minds onto ourselves: we interpret our own sensory mental states (Carruthers 2009).[1] Others, such and Shaun Nichols and Stephen Stich, argue that we come to know our own minds using capacities that are dedicated for that purpose and that allow us to simply recognise our own mental states (Nichols & Stich 2003; see also Goldman 2006)[2].xxx

Theories of self-knowledge can be evaluated using the criteria of simplicity and coherence with surrounding theories in cognitive science (Rimkevičius 2019). They can also be evaluated using the criteria of empirical adequacy and scientific fruitfulness. A theory is empirically adequate if its predictions are supported by past empirical discoveries. It is scientifically fruitful if it makes new predictions that contribute to new empirical discoveries.

A decade has now passed since the introduction of Carruthers's interpretive-sensory access (ISA) theory of self-knowledge. In this paper, I review the most significant developments during that time period, noting the relevance to the ISA theory of new empirical findings where the relevance has not yet been noted and providing responses from the perspective of the ISA theory to interpretations of those findings where such a response has not yet been given.

1. Empirical Adequacy

The ISA theory makes six main predictions that differentiate it from its rivals (Carruthers 2011: 3–7).[3] First, one will not attribute an attitude to oneself in absence of a sensory cue on which the attribution can be based. Second, concerning development, if a child is able to attribute a mental state to others, it will be able to attribute it to itself, and if it is unable to attribute it to others, it will be unable to attribute it to itself. Third, if one of the capacities is impaired, then the other will be impaired; relatedly, both capacities will be associated with the same brain regions.

Fourth, in absence of effortful training, monitoring of one's own attitudes will not be very reliable; relatedly, control of one's own attitudes will always be broadly behavioural in character. Fifth, one will be misled about one's own attitudes when presented with sensory cues that are analogous to those that would mislead one about others' attitudes. Sixth, concerning non-human animals, if an animal is able to attribute a mental state to

[1] To some extent, the theory can remain neutral on how we understand other minds (Carruthers 2011: 2). It can allow that this involves elements of theorising, simulation, innate knowledge, as well as perception (for a suggestion of how these elements might fit together, see Carruthers 2015a).

[2] There aare many more theories of self-knowledge on the market (for an overview, see Gertler 2015), but the ones cited above provide the clearest contrast in terms of empirical predictions.

[3] Arguably, in all six cases, the exact opposite from what the ISA theory predicts should be predicted by either Nichols and Stich, or Goldman (see the table in Carruthers 2011: 202).

others, then it will be able to attribute it to itself, and if it is unable to attribute it to others, then it will be unable to attribute it to itself.

1.1. No Non-Sensory Awareness

Carruthers argues that there is no non-sensory awareness of attitudes (Carruthers 2011: 214–221). The main challenge he responds to concerns reports of unsymbolised thinking. They come from descriptive experience sampling studies by Russell Hurlburt and colleagues (Hurlburt 2011: 291–308). In these studies, participants wear a beeper signalling at unpredictable intervals that they should report their immediately past inner experience. Some participants report episodes of unsymbolised thinking.

In brief, Carruthers' response is that these people fail to report the sensory cues that are present and, in any case, that 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 precise 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). Some philosophers claim that what it is like to entertain a thought is irreducible to other kinds of phenomenology, such as the quasi-perceptual experience of one's own inner speech. To the above discussion this debate adds historical perspective and expert reports. It harks back to a debate roughly a hundred years ago, when many of the same claims have been advanced, and 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. Despite the differences, 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 probably 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 thoughts 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 appears 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: ones where the subject has minimal external cues for interpreting their own state of mind. Georges Rey argues that if there would be no non-sensory awareness of attitudes, then our reports

of them would be unreliable in meditative cases: here, one just has to rely on non-sensory awareness (Rey 2013: 273–274). Following this line of thought, one should predict that non-sensory awareness will be reported more frequently in meditative cases.

It turns out that 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 not trained in meditation, suggests that people in such states have no less sensory cues than usual (Hurlburt et al. 2015). The crucial contribution to inner experience here is probably made by quasi-perceptual cues, such as mental imagery. 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 both of these studies were very small, one should be very cautious when making conclusions at this point.

1.2. Development

Carruthers argues that self-knowledge and other-knowledge develop together (Carruthers 2011: 203–209, 240–248). More precisely, he argues 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 responds to concerns reports of knowledge of one's own knowledge, pretence, and perspective in children who are as yet unable to attribute those mental states to others (Nichols & Stich 2003: 174–176). In brief, Carruthers' response to this challenge is 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, he notes that 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 responds to concerns an alternative explanation, based on behavioural rules, of reports of knowledge of other people's false beliefs, in children who are as yet unable to attribute false beliefs to themselves (Nichols & Stich 2003: 170–174). In brief, his response is 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.

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 such knowledge in children from six to thirty-six months of age (Scott & Baillargeon 2017). However, there have been some replications that were only partly successful as well as some failed replications (Baillargeon et al. 2018).

In response, following René Baillargeon and colleagues, one might note that there are some procedural differences between some of the original and replication studies, and that there are still many original paradigms that remain unaffected. Moreover, even if all of the non-verbal studies eventually failed to replicate, this would not yet show that knowledge of one's own false beliefs develops earlier. Verbal studies suggest that knowledge of one's own and others' false beliefs develops at the same time, at four years of age (Wellman et al. 2001), and 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 itself is (Goupil et al. 2016, Goupil & Kouider 2016). The authors interpret these results as suggesting that core metacognitive capacities are already in place in infancy.

One could respond to this in three different ways. The first is to insist, despite the replication crisis, that there is evidence of knowledge of others' false beliefs at the same age or even earlier. The second is to argue that in the experiments by Goupil and others the infants merely had to be influenced by their uncertainty, not to monitor it as such. In fact, the authors make no explicit suggestion that the exhibited abilities are meta-representational. The third is to note that there is some evidence (to be discussed below) suggesting that the implicit metacognition tasks that they used measure something different from what is measured by explicit metacognition tasks, which clearly measure meta-representational abilities.

1.3. Dissociations

Carruthers argues that there are no dissociations between self-knowledge and other-knowledge (Carruthers 2011: 293–324). More precisely, he argues that both capacities are systematically impaired in schizophrenia and autism, but not in alexithymia. Also, he argues that the brain areas that are activated when using either capacity are these: the medial prefrontal cortex, the posterior cingulate cortex, the temporal pole, the temporo-parietal junction, and the superior temporal sulcus.

One of the two main challenges that he responds to concerns 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 is 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, he claims that 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 about actual or mentally simulated movement. It does not represent the mental states as such, and therefore does not engage the person's meta-representational capacities. As for alexithymia, Carruthers claims that it probably involves a fault in the mechanism that makes the valence of one's affect directly accessible to the person. He also claims that the valence of one's own affect is simply recognised, so the fault would not lie with one's meta-representational abilities.

The second of the two main challenges that he responds to concerns purported dissociations between the level of activation of certain brain areas when exercising the two capacities (Lombardo et al. 2009). In brief, his response is that the results are not consistent across studies and that the tasks for 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, he argues, since information about oneself is generally more familiar, more emotionally charged, and more deeply processed, it should be expected that activation levels will be somewhat different in the two 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 when compared to 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). They interpret these results as suggesting that mindreading is significantly impaired in autism.

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 has recently put forward a book-length argument that working memory and consciousness are sensory-based (Carruthers 2015b). If this were so, then it would further support the ISA theory. He supports his claim in part by referring to new 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 causal and computational involvement of these 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 notes that there is a case where both of these impairments are present in the same person and argues that the ISA theory cannot explain how this is possible (Peters 2014a). In particular, the patient 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 reports concerning 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 do.

Conceding that this is a difficult case, one could try to respond by noting that the rival theory 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 reports concerning 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.

1.4. Metacognition

Carruthers argues that people have no strong innate metacognitive capacities (Carruthers 2011: 263–278). More precisely, he argues that one controls one's own learning by initiating behaviour or intervening on it (e.g., rehearsing what one wants to memorise). Likewise, he argues that one relies on heuristics (e.g., retrieval fluency) to judge whether one has learned something. He also claims that one rarely exercises control over one's own attitudes as such, and when one does so, one operates on symbols rather than thoughts.

An important challenge to these claims concerns the link between direct knowledge and direct control. In principle, one could have direct (non-behavioural) control without direct (non-interpretive) 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 is 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 attitudes indirectly, it is unclear what evolutionary function 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'. Situational strategies are indirect ways of controlling one's own attitudes. Therefore, the suggestion is in line with the prediction 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 prediction that people will find it difficult to control their own minds in absence of effortful training. Of course, one might speculate that the results might have been different if the participants were extensively trained.

Yet another important development in this area concerns the distinction between evaluative and voluntary processes. Tillmann Vierkant argues that if conscious processes are voluntary (as most cognitive scientists agree), and if judgements are not voluntary (as most philosophers agree), then conscious judgements do not exist (Vierkant, in draft). This line of reasoning could probably be extended to cover 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 processes directly.

1.5. Misattribution

Carruthers argues that misleading sensory cues lead to mistakes about one's own attitudes (Carruthers 2011: 147–154, 325–367). More precisely, he argues that such cues lead to mistakes about the causes of one's own attitudes as well as about the occurrence of particular judgements and decisions. He also argues that they mislead about the more fine-grained properties of the objects of one's own affects (as opposed to the valence of those affects or the identity of the objects of those affects).

The main challenge that he responds to concerns an alternative explanation of misattribution, which appeals to pragmatic pressures (Rey 2008). In brief, his response is that pragmatic pressures are unlikely to account for all or even most of 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). Note that one could concede this point and insist on some other explanation for cases where pragmatic pressures are minimal, but then the overall explanation would swiftly become more complex.

An important development in this area concerns one such proposal of an alternative way to explain misattribution. Sophie Keeling argues 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 attitudes knowledgeably and with reference to motivating reasons (Keeling 2018). One could note two issues with the explanation Keeling offers.

First, although it seems to be presented as an alternative account, the existence of such a 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 that people do not only misattribute reasons for their attitudes or actions. They also 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 explanation of their own attitude often closely 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 relies 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 2014b, Walter 2014). However, now there are 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 the 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 the decision. However, there are studies that support a very similar conclusion: people sometimes confirm having made one decision, when they in fact have made another decision (Johansson et al. 2005, Hall et al. 2010, 2012).

1.6. Comparative Evidence

Carruthers argues that self-knowledge and other-knowledge evolved together (Carruthers 2011: 254–259, 278–287). More precisely, he argues that some non-human animals are

able to attribute percepts and goals to others and themselves, but all non-human animals are unable to attribute false beliefs, whether to themselves, or to others.

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

The second of the three main challenges that he responds to concerns the purported capacity of some non-human animals to monitor their own uncertainty (Couchman et al. 2009). The suggestion is that the animal thinks about its uncertainty as such. In brief, his response is 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 uncertainty as such.

The third of the three main challenges that he responds to concerns the capacity of some non-human animals to seek information (Kornell et al. 2007). Again, the suggestion is that the animal thinks about its knowledge as such. In brief, his response is that it 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?'; it does 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 action, in a similar way that fear directly motivates action, 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 in this section 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 are able to pass the anticipatory looking task (Krupenye et al. 2016). Buttelmann and colleagues report that great apes are able to pass the interactive helping task (Buttelmann et al. 2017). This suggests that the animals can attribute false beliefs to others. There is no evidence that great apes are able to attribute false-beliefs to themselves, so this presents a problem.

In response, 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 and colleagues 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 in case of the interactive helping task. Here, two other responses are possible.

One is to note that the core of the prediction about other animals is that metacognition should not evolve earlier than mindreading. The other part, 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 would 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, in any case, the evolutionary claim and prediction should be given relatively less weight (Carruthers 2011: 7). The ISA theory's claims and predictions about humans might be true and supported even if its claim and prediction about other animal species are not. This would deprive the theory of an evolutionary argument in its favour, but would not yet yield a very strong argument against it.

2. Scientific Fruitfulness

It is now generally agreed that scientific fruitfulness is a theoretical virtue. This is because the fact that a theory continues to make new predictions that provide a framework for new empirical research, in addition to explaining the evidence that is already there, tends to be a reliable sign that the theory is progressing rather than degenerating (Newton-Smith 1981: 223–232). Carruthers argues that the ISA theory does make new predictions, and that it makes more of them than its rivals do (Carruthers 2011: 370).

First, he argues that in the past, the framework for most of the empirical research on self-knowledge was provided by predictions that were drawn from theories that are at least similar to the ISA theory (Bem 1967; Nisbett & Wilson 1977; Gazzaniga 1998; Wegner 2002; Wilson 2002). They are similar in that, like the ISA theory, but unlike its main rival, these theories predict misattributions, and not merely accommodate them.

Second, he argues that the ISA theory itself makes new predictions that could potentially contribute to new empirical research. First of all, it gives a set of six main predictions that separate it from others (Carruthers 2011: 202). In contrast, it is not always clear what many of the rival theories should predict, since this is not clearly spelled out by their authors (the clearest exceptions are Nichols and Stich (2003), and Goldman (2006)). Moreover, in addition to the six main predictions, the ISA theory explicitly makes many more predictions that are more specific.[4]

The most significant development in this area is that in some cases the ISA theory's new predictions have already contributed to new empirical studies. In particular, it contributed to new empirical studies on intuitions about self-knowledge and to new empirical studies on autism.

First, as noted above, Carruthers and colleagues conducted three studies to test the predictions that mindreading capacities will be impaired in autism, that this impairment will be matched by an impairment in explicit metacognition, and that these impairments

[4] 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.

will not be matched by an impairment in implicit metacognition (Williams et al. 2018; Nicholson et al. 2019).

Second, partly in order to show that the burden of proof does not lie on the side of the ISA theory, Carruthers argues that we should expect people to have evolved an intuitive belief in 'transparency' (or non-interpretive access) even if this belief is false (Carruthers 2011: 11–45). This is because the belief is adaptive: in most cases, it simplifies mental state attribution without making it less accurate. He predicts that this belief will turn out to be a human universal. The idea has led to a pilot study with Clark Barrett (reported in Carruthers 2008). It also led to five follow-up studies by others (Kozuch & Nichols 2011).

Conclusion

The ISA theory is still empirically adequate and scientifically fruitful. It remains empirically adequate, because during the first decade after its introduction its predictions have received significant further support and none of the challenges have yet been such that the theory could not provide a reasonable answer to them. Perhaps the most significant new source of support comes from studies suggesting that, in autism, impairments in explicit metacognition are related to impairments in explicit mindreading but not implicit metacognition. Perhaps the most significant new challenge comes from studies suggesting that great apes attribute false beliefs only to others. The ISA theory also remains scientifically fruitful, since it offers new empirical predictions, some of which are yet to be tested, and since it is contributing to new empirical studies on intuitions and autism.

References

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

Bayne, T., Montague, M., eds., 2011. *Cognitive Phenomenology*. Oxford: Oxford University Press.

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

Brasil-Neto, J. P., Pascual-Leone, A., Valls-Solé, J., Cohen, L. G., Hallett, M., 1992. Focal Transcranial Magnetic Stimulation and Response Bias in Forced-Choice Task. *Journal of Neurology, Neurosurgery, and Psychiatry* 55: 964–966. http://dx.doi.org/10.1136/jnnp.55.10.964

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

Carruthers, P., 2018. Basic Questions. *Mind and Language* 32(2): 140–147. https://doi.org/10.1111/mila.12167

Carruthers, P., 2015b. *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought*. Oxford: Oxford University Press.

Carruthers, P. 2015a. Perceiving Mental States. *Consciousness and Cognition* 36: 498–507. https://doi.org/10.1016/j.concog.2015.04.009

Carruthers, P., 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge.* Oxford: Oxford University Press.

Carruthers, P., 2009. How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition. *Behavioral 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.

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

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

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

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

Goldman, A. I., 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

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

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., 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., 2011. *Investigating Pristine Inner Experience.* Cambridge: Cambridge University Press.

Hurlburt, R. T., Heavey, C. L., 2006. *Describing Inner Experience*. Amsterdam: John Benjamins.

Johansson, P., Hall, L., Sikström, S., Olsson, A., 2005. Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science* 310(5745): 116–119. https://doi.org/10.1126/science.1111709

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

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.

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

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

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

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

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., 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

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

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

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

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

Rey, G., 2013. 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

Rimkevičius, P. 2019. The Interpretive-Sensory Access Theory of Self-Knowledge: Simplicity and Coherence with Surrounding Theories. *Problemos* 96: 148–159. https://www.journals.vu.lt/problemos/article/view/14617

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

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

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

Vierkant, T., in draft. System 2 Judgements Do Not Exist!

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. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wegner, D. M., Wheatley, T., 1999. Apparent Mental Causation: Sources of the Experience of Will. *American Psychologist* 54(7): 480–492. http://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

Williams, D. M., Nicholson, T., Grainger, C., Lind, S. L., Carruthers, P., 2019. 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, T. D., 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.

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. http://dx.doi.org/10.1126/science.1250830

Wu, W., 2014. Being in the Workspace, from a Neural Point of View, Comments on Peter Carruthers, 'On Central Cognition'. *Philosophical Studies* 170: 163–174. https://doi.org/10.1007/s11098-013-0169-8