Vilnius University Press

Contents lists available at Vilnius University Press

Problemos
2019, vol. 96, pp. 148–159

ISSN 1392-1126 eISSN 2424-6158
DOI: https://doi.org/10.15388/Problemos.96.12

The Interpretive-Sensory Access Theory of Self-Knowledge: Simplicity and Coherence with Surrounding Theories

Paulius Rimkevičius
Vilniaus universiteto Filosofijos institutas
E. paštas Paulius.rimkevicius@fsf.vu.lt

Abstract. The interpretive-sensory access (ISA) theory of self-knowledge claims that one knows one's own mind by turning one's capacity to know other minds onto oneself. Previously, researchers mostly debated whether the theory receives the most support from the results of empirical research. They have given much less attention to the question whether the theory is the simplest of the available alternatives. I argue that the question of simplicity should be considered in light of the well-established theories surrounding the ISA theory. I claim that the ISA theory then proves to be the simplest. I reply to objections to this claim related to recent developments in this area of research: the emergence of a unified transparency theory of self-knowledge and the relative establishment of the predictive processing theory.

Keywords: self-knowledge, simplicity, Peter Carruthers, transparency, predictive processing

Interpretacinės-sensorinės prieigos savižinos teorija: paprastumas ir derėjimas su supančiomis teorijomis

Santrauka. Interpretacinės-sensorinės prieigos (ISP) savižinos teorija teigia, kad mes sužinome savo pačių mintis nukreipdami į save pačius savo gebėjimą, skirtą kitų individų mintims sužinoti. Iki šiol daugiausia diskutuota, ar būtent šią teoriją daugiausia paremia empirinių tyrimų rezultatai. Daug mažiau dėmesio skirta klausimui, ar ši teorija yra paprasčiausia iš esamų alternatyvų. Šiame straipsnyje įrodinėjama, kad paprastumo klausimas turėtų būti nagrinėjamas atsižvelgiant į ISP teoriją supančias jau įsitvirtinusias teorijas. Teigiama, kad, atsižvelgus į supančias teorijas, ISP teorija pasirodo esanti paprasčiausia. Taip pat atsakoma į du prieštaravimus, susijusius su pastarojo meto pokyčiais šioje tyrimų srityje: su bendros skaidrumo savižinos teorijos atsiradimu ir su santykiniu numatančio informacijos apdorojimo teorijos įsitvirtinimu.

Pagrindiniai žodžiai: savižina, paprastumas, Peteris Carruthersas, skaidrumas, numatantis informacijos apdorojimas

Acknowledgement. I am grateful to Peter Carruthers, Keith Frankish, Joëlle Proust, and Lukas Schwengerer for their comments on some of my ideas put forward in this paper. The mistakes are all mine.

Received: 26/06/2019. **Accepted:** 03/09/2019

How one knows one's own mind is widely debated in contemporary philosophy of mind. A question that receives special attention is how one knows one's own propositional attitudes, such as beliefs, desires, and intentions, as well as judgements, and decisions. Philosophers who participate in this debate can roughly be divided into two groups: those who claim that one knows one's own propositional attitudes by turning one's capacity to know others' propositional attitudes onto oneself (Ryle 1949, Gopnik 1993, Carruthers 2011, Cassam 2014), and those who claim that one knows one's own propositional attitudes by some other means by which one can only know one's own propositional attitudes but not others' propositional attitudes (Moran 2001, Nichols & Stich 2003, Bar-On 2004, Frankish 2004, Bilgrami 2006, Goldman 2006, O'Brien 2007, Fernández 2013, Proust 2013, Coliva 2016, Byrne, A. 2018, Schwengerer 2018a). Since the philosophers who belong to the first group claim that the cases of self and other are essentially symmetrical, their theory can be called the symmetry theory, while their opponents' theory can be called the asymmetry theory (Schwitzgebel 2014).

Perhaps the most well-developed version of the symmetry theory is the interpretive-sensory access (ISA) theory of self-knowledge. The ISA theory claims that: (1) there is a single mental faculty that underlies our attributions of propositional attitudes, whether to ourselves or to others; (2) this faculty has only sensory access to its domain; (3) its access to most kinds of our own propositional attitudes is interpretive; and (4) the faculty in question evolved to sustain and facilitate outward looking or other-directed forms of social cognition (Carruthers 2011: 1–2). These four claims constitute the core of the ISA theory and they imply six main empirical predictions.

The ISA theory predicts that: (1) people should be incapable of attributing attitudes to themselves in absence of relevant sensory cues; (2) there should be no fundamental differences between self-understanding and other-understanding in childhood development; (3) there should be no dissociations between people's competence for knowledge of self and other; nor should different brain areas be involved; (4) people should lack any deep or well-developed metacognitive competence; (5) people should confabulate plausible-seeming attitudes for themselves whenever presented with misleading sensory data; and (6) creatures capable of attributing mental states of a given kind to others should be capable of attributing states of that sort to oneself; and there should be no creatures capable of attributing states to themselves of a kind that they cannot attribute to others (Carruthers 2011: 370). Together, these six main predictions differentiate the ISA theory from its rivals.

Proponents of the ISA theory claim that it has the following four theoretical virtues: it is the simplest of the available alternatives, coheres with surrounding theories in cognitive science, receives the most support from the results of empirical research, and is scientifically fruitful in the sense of making more new predictions that contribute to further empirical research (Carruthers 2011: 368). The debate surrounding the ISA theory mostly concentrates on the question whether the theory receives the most support from the results of empirical research (for a recent overview see Rimkevičius, under review). A question that has received considerably less attention is whether the ISA theory is the most simple of the available alternatives. I suggest that there are two recent developments in this are of research that make the question of simplicity more pressing now.

One recent development that makes the question of simplicity more pressing now is the emergence of unified versions of the transparency theory of self-knowledge. The transparency theory claims that one knows one's own mind by attending to the relevant tracts of the outside world, as opposed to the mind itself, so the mind is in this sense transparent (Evans, G. 1982, Moran 2001, Fernández 2013, Byrne, A. 2018, Schwengerer 2018a). A disadvantage of early transparency theories is that they seem to apply to such kinds of mental states as beliefs, but not to such kinds of mental states as sensations. However, versions of the transparency theory have now emerged that explain how one knows all kinds of one's own mental states in the same way (Byrne, A. 2018, Schwengerer 2018a). These theories promise a relatively simple general explanation of self-knowledge.

Another recent development that makes the question of simplicity more pressing now is the relative establishment of the predictive processing theory of the mind. The predictive processing theory claims that the mind's function is to reduce error in predictions about the outside world (Clark 2013). The predictive processing theory offers a general framework for theories of cognitive processes and is said to have already become a well-established alternative to more traditional alternatives (Hohwy 2014, Clark 2016). If the question of simplicity should be considered in light of surrounding theories in cognitive science, as I shall argue, then the question whether the ISA theory is the simplest of the available alternatives should also be considered in light of the predictive processing theory of the mind.

In fact, some researchers explicitly suggest that the transparency theory of self-knowledge is both the simplest of the available alternatives (Byrne, A. 2018), and the one that best fits the predictive processing framework (Schwengerer 2018b). I want to question both of these suggestions.

In the following, I first provide a brief clarification of the theoretical virtues of simplicity and of coherence with surrounding theories, argue that they should be considered in light of one another, and suggest that the importance of this might have been overlooked in the debate about self-knowledge. In the second part, I suggest that when coherence with surrounding theories is taken into account, the ISA theory proves to be the simplest of the available alternatives, including the unified version of the transparency theory. Finally, I consider and respond to the possible objection that since the ISA theory does not fit the predictive processing framework as well as the transparency theory does, the ISA theory would constitute a part of a less simple overall picture of the mind.

1. Two Theoretical Virtues

Theoretical virtues are such theoretical features by which it is rational to guide one's choice of theory. They are relied upon as steady guides in theory choice because through the long history of science they have emerged as reliable indications that a research programme is progressing, as opposed to degenerating (Lakatos 1970: 116; see also Newton-Smith 1981: 225). Theoretical virtues are now widely agreed in the scientific community to include the following four theoretical features: simplicity, coherence with surrounding theories that are already relatively well-established, support from the results of empirical research, and scientific fertility or fruitfulness in the sense of the theory predictions making a contribution to further empirical research (Newton-Smith 1981: 223–232). Therefore, showing that simplicity is possessed by a given theory to a higher degree than by other, gives the proponent of that theory a considerable dialectical advantage.

The philosophical debate about self-knowledge is no exception in this respect and therefore the participants in this debate are trying to show that their theory has this virtue, while their opponents' theories lack this virtue. However, I suggest that the participants of this debate might have overlooked how important it is to consider simplicity in light of other theoretical virtues. In particular, they might have overlooked how important it is for simplicity to be considered in light of coherence with surrounding theories. These virtues should be considered in light of one another, because a theory is simpler only if it postulates fewer new entities, but whether a postulated entity is new can only be seen in light of those entities that are already postulated by relatively well-established surrounding theories. In the case of self-knowledge, such surrounding theories clearly include theories of how one knows other minds, as well as more general theories of how the mind works.

Furthermore, when considering whether a given theory of self-knowledge coheres with one of the given surrounding theories, it is important bear in mind that the link that makes them cohere with one another might be provided by a third theory. If that third theory is itself already well-established, then relying on it to provide the link between the first two theories will not make the overall picture of the mind any more complicated. In the following, I suggest that the ISA theory benefits in this respect from the presence of relatively well-established theories of how one knows other minds and how our minds have evolved.

2. Simplicity

Before considering the new challenges posed by recent developments, it is worthwhile to review how the proponents of the ISA theory originally argued that it the ISA theory simplest of the available alternatives (Carruthers 2011: 6, 369). They gave two main reasons: one has to do with a comparison between one's knowledge of one's own mind and one's knowledge of other minds, while the other has to do with a comparison between one's knowledge of one's own mind and one's ignorance of one's own mind.

The first of the originally given reasons why the ISA theory is simpler than its rivals is that the ISA theory gives a unified explanation of one's knowledge of one's own and other minds. In fact, all versions of the symmetry theory claim that one knows one's own and other minds by using the same mental capacity. Whereas all versions of the asymmetry theory claim that one knows one's own and other minds in different ways. The relative complexity of the asymmetry theory is perhaps the most evident in the case of the inner-sense theory of self-knowledge. The inner-sense theory claims that one knows one own mind using a mental faculty that is dedicated specifically for that purpose and that functions similarly to those mental faculties that underly perception (Nichols & Stich 2003, Goldman 2006). Evidently, the symmetry theory provides a simpler overall explanation of one's knowledge of one's own and other minds than the asymmetry theory.

The second of the originally given reasons why the ISA theory is simpler than its rivals is that the ISA theory gives a unified account of self-knowledge and self-ignorance. All contemporary theories of self-knowledge agree that one sometimes misinterprets one's own propositional attitudes. For instance, they agree that one might be led to interpret and sometimes to misinterpret one's own desires in such circumstances as a psychotherapy session. Since the symmetry theory claims that one's access to one's own propositional attitudes is always interpretive, they do not need to postulate an additional means of access to one's own propositional attitudes in order to explain the occasional misinterpretations. In contrast, since the asymmetry theory claims that normally one's access to one's own propositional attitudes is not interpretive, they need to postulate an additional means of access to one's own propositional attitudes in order to explain the occasional misinterpretations. Evidently, the symmetry theory also provides a simpler overall explanation of one's knowledge and one's ignorance of one's own mind.

These two originally given reasons why the ISA theory is simpler than its rivals are now widely acknowledged by the opponents of the ISA theory. For instance, 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 the symmetry theory, which claims one's access to one's own propositional attitudes is interpretive (Byrne, A. 2018: 177). However, the opponents of the ISA theory question whether it is the simplest of the available alternatives on other grounds.

In particular, Byrne suggests that the ISA theory gives a less unified account of one's knowledge of one's own propositional attitudes and one's knowledge of one's own sensory mental states (Byrne, A. 2012, 2018: 16). According to him, most 'neo-Ryleans', and perhaps even Ryle himself, claim that one's knowledge of one's own sensory mental states is not always interpretive. Carruthers is fairly explicit about this when he writes that one's access to one's own sensory mental states might be more like recognition than interpretation, or closer to how the transparency theory describes one's knowledge of one's own propositional attitudes (Carruthers 2011: 81). For this reason, Byrne concludes that the ISA theory is in a sense a complex theory of self-knowledge.

It is true that earlier versions of the transparency theory also were complex in this sense. For they only seemed to apply to one's knowledge of one's own propositional attitudes such as beliefs, but not other mental states. However, unified versions of the transparency theory have now emerged that account for one's knowledge of all kinds of one's own mental states in the same way (Byrne, A. 2018, Schwengerer 2018a).

For instance, Byrne's new version of transparency theory claims that one knows all kinds of one's own mental states by inferring them from corresponding tracts of the outside world. According to this theory, one normally knows that one believes that *p* by applying the inference rule 'If *p*, believe that you believe that *p*' (Byrne, A. 2018: 102). Likewise, one knows 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, A. 2018: 149). Crucially, the theory claims that applying these rules of inferences only requires one to possess an ordinary reasoning capacity, not a mental faculty dedicated for self-knowledge. From this Byrne concludes that a unified 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 knowledge and ignorance of one's own mind. That is to say that the originally given reasons to think that the ISA theory is simpler in those respects would remain standing even if Byrne's suggestion were also left to stand. One would then have to concede that the ISA theory is simpler in some respects, while the transparency theory is simpler in another. However, it is unclear whether Byrne's suggestion stands. Here I want to suggest two responses to it that a proponent of the ISA theory could make.

The first response is to say that the ISA theory is compatible with the claim that all self-knowledge is interpretive. If one were to add to the four core claims of the ISA theory described above a fifth claim that says one's access to one's own sensory mental states is interpretive, then one would get a more unified interpretive theory of self-knowledge. At some points, it seems that Quassim Cassam suggests that the ISA theory should make this fifth claim (Cassam 2014: Ch. 12). If one were to commit to this fifth claim, then the ISA theory would certainly give a unified account of one's knowledge of one's own propositional attitudes and one's knowledge of one's own sensory mental states. Yet the four core claims of the ISA theory themselves are silent about the kind of access one has to one's own sensory mental states. 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 ISA theorist.

At some points, it seems that Carruthers suggests that the ISA theory should remain neutral on this fifth claim (Carruthers 2011: xi), or that it should reject it (Carruthers 2011: 81). One could reject the fifth claim and concede Byrne's point about simplicity, but then argue that the ISA theory is as simple as any theory of self-knowledge should be, while the transparency oversimplifies things. Generally, the simplicity or complexity of a theory should reflect the simplicity or complexity of reality. One might argue that one's knowledge of one's own propositional attitudes and one's knowledge of one's own sensory mental states are relevantly different in reality. For instance, one might suggest that self-attributions of propositional attitudes and self-attributions of sensory mental states 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 to the effect that one often misinterprets one's own propositional attitudes (Carruthers

2011: 325–367), but a similar case has not yet been made to the effect that one often similarly misinterprets one's own sensory mental states. In other words, one might argue that the transparency theory buys simplicity at the cost of empirical support. However, I suggest that proponents of the ISA theory, even they reject the fifth claim, do not need not concede Byrne's point about simplicity, since there is another way to respond to his suggestion.

The second response to Byrne's suggestion is to say that none of the entities postulated by the ISA theory are new. This is because every entity that the ISA theory postulates is already postulated by surrounding theories that are already relatively well-established. In particular, these entities are postulated by theories of how one knows other minds. These theories lead us to believe that there 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.

More precisely, the input in this process is sensory in both cases, although there are kinds of sensory input that are related primarily to the self. These kinds of sensory input include those that come from interoception and proprioception, as well as inner speech and other kinds of mental imagery. Similarly, the processing rules in this process are rules of inference of one's own intuitive theory of mind, although different rules may be applied to processing information about different individuals. For instance, a given kind of sensory input that is related to the self may be processed more deeply than that kind of sensory input that is related to other people. Lastly, the outputs in this process are beliefs about mental states, although these beliefs may be stored somewhat differently. Note, however, that different mental files or different 'person models' (Newen 2015) for different people is something that one should already assume in order to explain knowledge of other minds. I conclude that the entities that the ISA theory postulates to explain self-knowledge are not new.

In contrast, some of the entities that are postulated by the transparency theory are new. The transparency theory falls short of 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 the process resemble the workings of a separate mental faculty at least to some degree. In particular, the theory claims that the rules are applied unconsciously, because if the process were made conscious the rules would strike their user as irrational. The reason they would strike one as irrational is that, 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 were now snowing at the North Pole, it would not be a good reason to believe that someone believes it. I conclude that the entities that the transparency theory postulates are new, even if they are less weighty than some of those postulated by other versions of the asymmetry theory.

3. Coherence with Surrounding Theories

To the preceding suggestion the ISA theory proves to be simpler than the transparency theory when one considers surrounding theories, such as the theory of how one knows other minds, one might propose the following objection. One might argue that the transparency theory fits the predictive processing framework better than the ISA theory and that therefore the transparency theory promises to be a part of a simpler overall account of the mind. This objection concedes the point that one should take account of surrounding theories and uses to argue against the ISA theory. I want to suggest that the objection fails to take note of an important link that makes the ISA theory perfectly coherent with the predictive processing theory.

But before moving to this new challenge posed by a recent development, again it might be worthwhile to first review how proponents of the ISA theory have originally argued that the ISA theory coheres with surrounding theories. The suggestion was that the ISA theory receives indirect support from three surrounding theories that at the time were already relatively well-established (Carruthers 2011: 47–68). These were the theories of global workspace, working memory, and Machiavellian intelligence. Here is why the ISA theory was said to receive support from them.

First, the global workspace theory claims that our mind consists of many specialised systems that communicate by means of sensory information through the one central system that is consciousness (Baars 1988). Since the ISA theory claims that attribution of mental states is subserved by one such specialised system that feeds on sensory information, it seems to cohere with the global workspace theory. Second, 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 consciously operate on them (Baddeley & Hitch 1974). Since the ISA theory claims that the mental faculty underlying mental state attribution is largely dependant on such manipulations 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, R. W. & 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 specialised system responsible for attributing mental states would have non-sensory access to its domain. Likewise, there is no suggestion in the Machiavellian intelligence theory that there were comparable evolutionary pressures for a specialised cognitive system for understanding one's own mind to evolve. These theories provide indirect support for the ISA theory, because they make it seem natural that one should have evolved a specialised cognitive system for understanding other minds that feeds on sensory input and is repurposed for understanding one's own mind. The ISA theory would receive indirect support from these theories even if it did not fit the predictive processing framework.

In addition to this original argument for the ISA theory, there have also been early suggestions that it does not cohere with some of the more general surrounding theories. In particular, it has been suggested that the ISA theory does not fully 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 J. St. B. T. & Stanovich 2013). Keith Frankish and Joëlle Proust have both expressed worries about the ISA theory that were 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). Whereas 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 the ISA theory only explains intuitive self-knowledge.

However, it seems that at least some of the disagreement here is 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 to the person but not meta-representational. They disagree whether these feelings should then be called 'meta-cognitive' (Carruthers 2017). Also in a recent paper, Carruthers agrees 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 to the person but do not constitute an attitude such as a belief on their own: they only do so in conjunction with things that are not directly accessible to the person, such as a commitment to what one says. They disagree whether the directly accessible event and those conjoined with it should then together be called a kind of attitude, an 'explicit belief' (Carruthers 2018). There might be deeper disagreements lurking beneath these terminological ones, but on the face of it, the ISA theory seems to cohere with most of what Proust and Frankish say about intuitive and reflective processes involved in self-knowledge.

I suggest that clarification might similarly show that the ISA theory is compatible with the predictive processing theory. Since the predictive processing framework is a very general theory, what it explains inevitably overlaps with what the ISA theory explains. If it turned out that the ISA theory is not readily compatible with the predictive processing theory, then either one would have to do more work and complicate the overall picture in order to graft the ISA theory onto the predictive processing theory, or one would have to reject one of the theories. I suggest that neither needs to be done, because a third theory provides the link that makes the ISA theory perfectly compatible with the predictive processing theory.

But first, here are the reasons for thinking that it is the transparency theory that best fits the predictive processing framework (Schwengerer 2018b). In a more traditional framework, one would roughly understand a piece of self-knowledge as a reliably formed true belief about one's own mental states. Since the predictive processing theory substitutes talk of propositional attitudes such as beliefs with talk of sub-personal predictions and error-correction, Schwengerer suggests that in this new framework one should understand self-knowledge as a pattern of higher-level predictions that accurately predict a pattern of lower-level predictions. He also suggests that since all of these predictions are ultimately about the outside world, the way of looking at self-knowledge that this new framework suggests coheres with the transparency theorist's notion that one acquires self-knowledge by attending to the relevant tracts of the outside world. Schwengerer also suggests a prediction that might differentiate a theory of self-knowledge based in the predictive processing framework from other theories of self-knowledge. Namely, he suggests that the theory should predict occasional surprise at the workings of one's own mind, which it would explain as an error being registered in the higher level of prediction. I suggest that proponents of the ISA theory can respond to Schwengerer's suggestion in two different ways.

The first response to Schwengerer's suggestion is to say that the ISA theory should predict occasional surprise at the workings of one's own mind. In fact, any of version of the symmetry theory should predict this, since the theory claims that one knows one's propositional attitudes by self-interpretation, which might lead first to error and then to the realisation that one has made that error and one's surprise at discovering it. Therefore, if this is the only prediction that is specific to theories of self-knowledge that are embedded in the predictive processing framework, then the ISA theory fits this framework perfectly.

The second response to Schwengerer's suggestion is to say that the predictive processing theory becomes 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 one simply seeks to minimise prediction error, why does one not to stay forever in such especially predictable environments as an empty darkened room? To solve this problem, the predictive processing theorist assumes that one evolved to have certain rigid prediction patters, such as that one will get food: one does not correct the prediction and predict that one will never get food, even though this would be a simple way to reduce prediction error. Crucially, the predictive processing theory itself insufficient to explain what set of rigid prediction patterns humans evolved to have.

Therefore, for all we know, the rigid predictions patterns that humans evolved to have might turn out to correspond to the mental architecture that is postulated by any of the theories of self-knowledge that are currently on offer. They might correspond, for instance, to the mental architecture that is postulated by the ISA theory. In fact, the ISA theory already receives indirect support from a relatively well-established theory of the driving forces behind the evolution of intelligence. Therefore, the Machiavellian intelligence theory might well provide the link that makes the ISA theory fit the predictive processing framework perfectly.

Conclusion

I have suggested that the importance of looking at the two theoretical virtues of simplicity and coherence with surrounding theories in light of one another might have been overlooked in the debate about self-knowledge. For a theory is simpler than the available alternatives only if fewer of the entities that it postulates are new, but whether the postulated entities are new can only be seen in light of what entities are already postulated by surrounding theories that are already relatively well-established.

I have also suggested that when surrounding theories are taken into account, it becomes evident that the entities postulated by the ISA theory are not new. This is because theories of how one knows other minds already postulate a process in which sensory input is fed into a mental mechanism that processes it according to the inference rules of an intuitive theory of mind and produces beliefs about mental states as output. In contrast, even the more simple of the alternatives postulate entities that are new: the unified transparency theory postulates processing rules that are introduced specifically for the purpose of explaining self-knowledge.

Finally, I have considered and replied to the objection that the ISA theory might yet prove to complicate the overall picture of the mind if it turned out that the ISA theory is not readily compatible with some newly established general theory of the mind, such as the predictive processing theory. I have suggested that the ISA theory fits this framework perfectly. This is because the ISA theory should make the prediction that was said to be specific to theories of self-knowledge embedded in this new framework (that one will sometimes be surprised at the workings of one's own mind), and because the framework relies upon assumptions about rigid prediction patters that humans evolved to have, which might well correspond to the one's postulated by the ISA theory.

References

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

Baddeley, A. D., Hitch, G., 1974. Working Memory. In: *The Psychology of Learning and Motivation: Advances in Research and Theory* 8, ed. G. H. Bower. New York: Academic Press, 47–89. https://doi.org/10.1016/s0079-7421(08)60452-1.

Bar-On, D., 2004. *Speaking My Mind: Expression and Self-Knowledge*. Oxford: Clarendon Press.

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

Byrne, A., 2018. *Transparency and Self-Knowledge.* Oxford: Oxford University Press.

Byrne, A., 2012. Peter Carruthers, *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. *Notre Dame Philosophical Reviews.* https://ndpr.nd.edu/news/the-opacity-of-mind-an-integrative-theory-of-self-knowledge. https://doi.org/10.1093/mind/fzt025

Byrne, R. W., Whiten, A., eds., 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Oxford: Oxoford University Press. https://doi.org/10.2307/2804121

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

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/9780198738824.003.0004.

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. *Behavioural and Brain Sciences* 32(2): 1–18. https://doi.org/10.1017/s0140525x09000545

Cassam, Q., 2014. *Self-Knowledge for Humans.* Oxford: Oxford University Press.

Clark, A., 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.

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

Coliva, A., 2016. *The Varieties of Self-Knowledge*. London: Palgrave Macmillan.

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

Evans, J. St. B. T., Stanovich, K. E., 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspective 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.

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

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

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

Hohwy, J., 2013. *The Predictive Mind*. Oxford: Oxford University Press.

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/10.1017/cbo9781139171434.009

Moran, R., 2001. *Authority and Estrangement: An Essay on Self-Knowledge.* Princeton: Princeton University Press. https://doi.org/10.1086/374016

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/mind/fzi181

Newen, A., 2015. Understanding Others: The Person Model Theory. *Open MIND* 26: 1–28. https://doi.org/10.15502/9783958570320

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

O'Brien, L., 2007. *Self-Knowing Agents*. Oxford: Oxford University Press.

Proust, J., 2013. *The Philosophy of Metacognition: Mental Agency and Self-Awareness.* Oxford: Oxford University Press.

Rimkevičius, P., under review. *The Interpretive-Sensory Access Theory of Self-Knowledge: Empirical Adequacy and Scientific Fruitfulness*.

Ryle, G., 1949. *The Concept of Mind*. New York: Barnes and Noble.

Schwengerer, L., 2018b. Self-Knowledge in a Predictive Processing Framework. *Review of Philosophy and Psychology*, first online. https://doi.org/10.1007/s1316

Schwengerer, L., 2018a. *A Unified Transparency Account of Self-Knowledge.* Doctoral dissertation, University of Edinburgh [unpublished].

Schwitzgebel, E., 2014. Introspection. In *The Stanford Encyclopaedia of Philosophy*, ed. E.N. Zalta. https://plato.stanford.edu/archives/win2016/entries/introspection/>