Beliefs in an Opaque Brain

Juan Andrés Abugattas Escalante

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Arts
In
Philosophy

Benjamin C. Jantzen
Kelly Trogdon
James C. Klagge

April 28, 2016
Blacksburg, VA

Keywords: self-knowledge, Peter Carruthers, introspection, cognitive science, belief

Beliefs in an Opaque Brain

Juan Andrés Abugattas Escalante

ABSTRACT

Peter Carruther's Interpretive Sensory-Access (ISA) theory of self-knowledge is an interesting account of the opaqueness of our own minds that draws upon a wide range of theories from cognitive science and philosophy. In the present paper, I argue that the theory's assumptions support the conclusion that the available perceptual evidence massively underdetermines all of an agent's second-order beliefs about her own beliefs. Such a result is far more negative than the ISA's well-known pessimism regarding self-knowledge. Furthermore, I also argue that, from the same assumptions, it is possible to build an argument to the effect that cognitive scientists trying to determine an agents' true behavior-causing attitude face similar underdetermination problems. Toward the end of the paper, I suggest that the theory's problems arise from a conflation of two different ways in which terms denoting propositional attitudes, such as "belief", are used in its formulation. Distinguishing between the two usages of these terms, in turn, leads to a further distinction between two different senses in which we can talk about the "opaqueness" of our own minds.

Acknowledgements

I would like to thank the members of the Committee for this thesis, professor Benjamin C. Jantzen, the Committee Chair, and professors Kelly Trogdon and James C. Klagge for their enormous help throughout the processes of developing, writing and defending the arguments expressed here. Their guidance both as teachers and advisors will continue to influence my philosophical endeavors.

I also thank my fellow students and friends in the Department of Philosophy at Virginia Tech, specially Caitlin Parker and Jonathan Lindsey, for their intelligent and considerate commentaries and conversation on the drafts of the present paper.

None of my work in philosophy and elsewhere would be possible without the constant support and comprehension from my family, in particular my mother Frida, Joanna and my beloved partner Zoila.

Table of Contents

1. Introduction

The aim of the present paper is to develop an argument about a conceptual aspect of Peter Carruthers' Interpretive Sensory-Access (ISA) theory of self-knowledge (as he calls it in his 2011 book). While I find the theory generally appealing, I believe that it yields an implausible result: following its premises, both regular agents and cognitive scientists lack, *at all times*, any means to ascertain whether the agents' self-ascriptions of attitudes are successful. Even though the theory's original, somewhat pessimistic, view of self-knowledge is explicit and well known – after all, it is intended to be an account of the opaqueness of our own minds - the consequences that I contend follow from it are more extreme and conflict with some of its own assumptions.

More specifically, I will argue that, given the ISA's assumption that attitudes are inferred from perceptual information only, plus its assumption that attitudes in general are only realized as amodal representations in a language of thought, ascribed attitudes are always massively underdetermined. Nonetheless, I do not intend to present a knockdown argument against the theory. One reason for this is that the argument relies on the fact that we currently lack empirical information regarding many of the ISA's assumed components of human cognitive architecture. Hence, future empirical investigation might enhance the theory's plausibility, *contra* my *a priori* epistemological arguments. Another reason is that, as I will argue in the final section of the paper, it is possible to distinguish between two senses in which we can speak about attitudes, such as "beliefs," that the ISA seems to conflate. Making such a distinction will force the theory's proponents to, at the very least, drop the claim that, for instance, beliefs (in both senses) are realized in a language of thought (a position that I will refer to as "belief realism"), which, in turn, could render the theory plausible *a priori*.

Before I begin, an important clarification is pertinent. Carruthers argues that it would be mistaken to claim that the philosophical accounts of self-knowledge that he rejects (among other types, the so-called "inner sense" views; aprioristic accounts, such as Shoemaker's (1994); and expressivist and "constitutive" accounts, which Carruthers considers as the closest to his own ISA) belong to an "explanatory space" different than the one corresponding to his approach (2011, pp. 21-24). I think he is right. As he asserts, all these philosophical accounts do entail commitments regarding the sort of subpersonal processes whose study grounds the ISA model (*ibid*., p. 22). These philosophical accounts have in common that they try to justify the ascription of a higher degree of reliability to our self-attributions of attitudes than the ISA. Whether it is due to the workings of a specialized introspection mechanism, or because it is in the very nature of our own mental states to be introspectible, or perhaps due to some expressive traits and/or practical consequences of the speech acts through which we refer to them, a common premise in said accounts, as Carruthers notes, is that self-knowledge is non-interpretive. To put it differently, the aforementioned philosophical positions are not consistent with the sort of non-conscious interpretive processes that, according to the ISA, determine self-knowledge.

As I stated above, I think that Carruthers' characterization of the ISA's disagreements with some of the most discussed philosophical theories of self-knowledge is correct. Nevertheless, if at least some of the arguments that I develop in the present paper are sound, then, *eo ipso*, there are reasons for changing certain aspects of the way in which the contenders in the debate, including Carruthers himself, have been thinking about self-knowledge.

2. Carruthers' Perception-Based Account of Self-Knowledge

2.1 Perception and Consciousness

The strongest argument for endorsing some theory of consciousness according to which it (mostly) depends on perception is pretty straightforward: so far, our empirical evidence about the mind/brain points in that direction (Carruthers, 2006/2011; Prinz 2007/2012). Accordingly, Carruthers' ISA theory makes use of a vast array of theories that explain different aspects of the way in which access consciousness might arise from what is currently known about the human brain's processing of perceptual data.[1] Without going into much detail, and with the intention of providing a helpful background for the discussion that follows, it is worth mentioning two of the principal theories upon which Carruthers bases his view of the "stream of consciousness" and, hence, his ISA theory.

The first of said theories is the idea that (access) consciousness is coextensive with "globally broadcasted" conceptualized perceptual information (Carruthers 2006; for the direct relation between this notion and the ISA model, see Carruthers 2011, in particular pp. 47-55).[2/3] The core idea is that the sensory representation of whatever stimulus (that is, pertaining to any modality) that gains attentional focus enters a neural network to which many sub-systems of the brain are connected. The upshot of this model is that the many specialized sub-systems in charge of our distinct cognitive capabilities can potentially access the attended string of perceptual data; in a way, it is through the shared network of perceptual representations that several of our mind's sub-systems interact, since it is conceivable that the outputs of each sub-system may enter the shared perceptual data network (as long as such outputs produce some form of perceptual

[1] In the sense of Block's "access consciousness" (1995).
[2] The idea that consciousness is coextensive with globally broadcasted perceptual representations is also, though briefly, discussed in Carruthers 2009/2014).
[3] The notion of a global broadcast network within the brain was first developed by Bernard J. Baars (1988).

representation that may, in turn, receive attentional focus).

The second theory on which I want to briefly comment here, and which is tightly related to the one on which I commented above, is the idea that the brain's "working memory" mostly employs perceptual data (Carruthers 2011, pp. 56-64; 2014). Basically, what our perception-based working memory does is to retain and manipulate a limited amount of the perceptual representations shared through the global broadcast network. Such retained representations then re-enter the shared network's flow so that other sub-systems of the mind/brain can employ them in their tasks.

In Carruthers' view, hence, conscious thought results, for the most part, from the workings of both the global broadcast architecture and the working memory. It is also important to note that this particular image of the "stream of consciousness" is his response to the notion of a "central work space" where "attitudes of all types can become active, engaging with one another and with systems of inference and decision making" (2014, p. 144), something similar to Fodor's hypothetical "central processing unit" (or "CPU") (Fodor, 1983), where an allegedly distinctively human holistic kind of reasoning takes place.[4] According to Fodor, the amodal medium of the properly contentful representations is a so-called "language of thought" (Fodor, 1975). In the language of thought, representations are produced compositionally, hence concepts can function as parts of more complex thoughts.

Now, crucially, Carruthers' view on human cognition in general assumes the existence of an amodal and compositional medium for representations, such as a language of thought. This is by no means inconsistent with his rejection of a propositional working space (such as Fodor's CPU). In fact, Carruthers' own belief realism is a commitment to the idea that attitudes are

[4] Carruthers exhaustively explores his thesis of working memory as a perceptual only "work space" in his recent *The Centered Mind* (2015).

realized as discrete, causally effective units, precisely as the theory of a language of thought postulates. The core of his discrepancy with the more standard conception of reflective reasoning is his idea that there are no attitudes in the stream of consciousness, only perceptual data. This is what grounds his interpretive view of second order cognition, since referring to reflection and self-knowledge as "interpretive" means that they operate on the basis of inferences from perceptual cues. Nevertheless, a perception-based account of self-knowledge/metacognition doesn't really require the assumption that our attitudes are realized in some subsystems of the mind/brain, and this is exactly what I will argue in Section 3.

2.2 The ISA at Work

According to the ISA theory, metacognition, metaphorically speaking, "results from us turning our mindreading abilities upon ourselves" (2009, p. 123). Our mindreading abilities are, in this view, the doings of a metarepresentational faculty that has access to data from perception. Upon receiving perceptual inputs (and "quasi-perceptual" data), the mindreading/metarepresentational faculty interprets these data by forming judgments regarding the attitudes of the relevant agent, including, of course, oneself. The *quid* of this idea is that the mindreading faculty doesn't have access to the outputs of the thinker's own attitude forming and decision making systems, just as it doesn't have access to another thinker's attitudes (2009, p. 124; also, cf. 2011 for a thorough discussion of these notions).

Importantly, Carruthers' theory rules out the hypothesis that we possess some special mechanism for introspecting our own mental states, where "introspection" is understood as "any reliable method for forming beliefs about one's own mental states that is not self-interpretative and that differs in kind from the ways in which we form beliefs about the mental states of other people" (2009, p. 123). Nevertheless, this view does leave room for the introspection of

perceptual judgments (2009, pp. 124-125; 2015, pp. 64-68) and, in fact, the theory assumes that such is the case, i.e., that conscious access to data from perception *is significantly more reliable than conscious access to one's own attitudes*.

An obvious consequence of all this is that we are left without much protection from confabulation; not even speech (overt or inner) is an absolute guarantor of privileged access to our own minds. In Carruthers' model, the utterance of sentences such as "I believe that x" is the result of the language faculty attaching first order sentences expressing the content x (selected after a search in memory) to phrases such as "I believe that" (2009, p. 125; 2011, p. 86). Notice that, according to this picture, no metacognitive process takes place before the utterance; instead, it is the utterance itself that triggers our entertaining the metacognitive thought. Furthermore, given that speech (including inner speech) is perceived, it needs to be interpreted (ibid.) Nevertheless, Carruthers thinks there are two reasons why interpreting our own speech is a more reliable endeavor than interpreting that of others.[5] One reason is that (a) we usually have at our disposal more perceptual cues for the disambiguation of our utterances than cues for disambiguating utterances of others. The second reason is that (b) we seem to possess a certain "relative accessibility to the concepts involved [in the utterance], which is a pervasive feature of speech comprehension generally" (2009, p. 126).[6] He explicates this notion by adding that

> [b]ecause the goals that initiated the utterance, "I shall walk to the bank," would almost certainly have included an activation of one or other specific concept bank, this will ensure the increased accessibility of that concept to the comprehension system when the utterance is processed and interpreted (ibid).

In this picture of consciousness, then, the interpretive system or module receives the perceptual information that is attached to some of the stored first order amodal concepts that comprise the

[5] After all, it doesn't seem to be plausible that we can confabulate in every possible context where self-ascription occurs.
[6] This theory was originally formulated by Dan Sperber and Deirdre Wilson (1986).

actual attitudes. For example, the amodal concept BANK is attached to certain visual and phonetic images. The presence of the concept in reasoning causes these images to enter the stream of consciousness, thus, in a sense, when attention is directed toward them, the metarepresentational system might gain "introspective" access to the first order concept. If, after interpreting ourselves in this manner, we utter the second order expression "I think I shall walk to the bank," this phonetic image might receive attentional focus, re-enter the stream of perceptual data and, perhaps, be further processed by other cognitive systems feeding off the global broadcasting network of perceived information (it may even be momentarily retained in the perceptual "working space," or working memory).

Notice, however, that we not only lack direct conscious access to the first order, amodal representations in a language of thought, we also lack conscious access to the interpretive process itself. Instead, in our standard, common-sense stance, we simply (and falsely) assume that the experienced perceptual and quasi-perceptual states amount to introspecting full-blown first order attitudes, and not just this or that concept, as the ISA theory suggests. Later on, I will say more about Carruthers' account of this commonsensical conception of the transparency of the mind. The main moral of what we've seen of the ISA so far is that, from an entirely objective perspective, subjects do not have conscious access, let alone particularly secure, introspective access, to the thoughts that truly cause, and hence explain, behavior (or, to use Carruthers' own way of characterizing the issue, the thoughts and attitudes that cause behavior "in the right way") (2011, pp. 102-104).

3. Beliefs in an Opaque Brain?

3.1 The Insufficiency of Introspective Knowledge

As we have seen, Carruthers' belief realism is a view that fits rather naturally with the postulation of amodal concepts. Yet, at least as it stands it seems to be problematic. Or so I'll argue in this section. The problem, roughly, is that it doesn't seem to be correct to describe first order amodal representational content in the same terms in which we talk about attitudes from the perspective of common sense.

When confronted with a particular situation in the world, if the mind operates by forming and weighing amodal attitudes, then it has to form beliefs whose component elements represent the aspects of the situation relevant to possible behavior. Carruthers endorses the idea that sensory data are already conceptualized even before entering further cognitive processes, i.e., percepts are, according to this view, "hybrid conceptual-nonconceptual representations" (2015, p. 66). This, in fact, is in line with the way in which we regularly experience perception, for it is merely intuitive that we perceive objects as things in particular. Furthermore, this immediate conceptualization of perception allegedly is very fine-grained. To employ Carruthers' own example of seeing a red tomato falling, we can express the resulting hybrid sensory-conceptual representation as THAT: **red**, RED, **round**, ROUND, **smooth**, SMOOTH, **moving down**, FALLING, TOMATO (where boldface indicates nonconceptual representations) (ibid, pp. 66-67). The hybrid nature of perceptual representations is what makes possible the presence of amodal concepts among the globally broadcasted perceptual data. This, in turn, would explain why we have introspective (in the sense of specially secure and conscious) access to our own perceptual states.

Yet, the true extent to which the aforementioned theory of perception allows us to ground our belief in introspective knowledge of our perceptual states might not be very clear. Intuitively, it seems hard to question that, under normal circumstances, basic object recognition introduces

an amodal representation of what the perceived object is in a very general sense. In the example above, such a basic conceptual unit would be TOMATO. Together with the indexical THAT, the conceptual composite forms the (usually implicit, although directly expressible in a natural language) "perceptually-embedded judgment" THAT: TOMATO. On the other hand, I think there are reasons for doubting that at least some of the other amodal concepts in the example above are as easily introspectible, in the sense of being particularly secure information about experience (Carruthers 2015, pp. 64-68).

In what follows, I'll develop an argument in support of the idea that the ISA faces a difficult *discovery problem*. Paraphrasing Clark Glymour's application of conceptual tools from discussions on the logic of discovery in a methodological analysis of cognitive science (1994, pp. 824-825), I will define a discovery problem as a collection of alternative conceivable strings of amodal representations that cause behavior (strings that, it is important to keep in mind, are the *realization of attitudes* in a language of thought).[7/8] In short, I will argue that, under the ISA's assumptions about the cognitive capabilities that underlie overt behavior and self-knowledge, *plus* a crucial lack of certain knowledge, agents don't have the grounds for solving the relevant discovery problem.

[7] This is a paraphrase of a formulation by Glymour in his cited article. The original phrase is: "I will say that a *discovery problem* consists of alternative conceivable graphs of normal cognitive architecture" (1994, p. 824). Italics are from the original.

[8] In his cited article, Glymour is assessing a different (and far more general, although in its formal aspects analogous) problem than the one discussed here: under what conditions is the cognitive scientist capable of selecting the "functional diagram" (typically, a network of modules) that most accurately represents the underlying architecture of some cognitive capability. Importantly, on many occasions, cognitive scientists rely on data from injured brains (i.e., injured in such a way that the studied capacity is limited, modified or nullified) in order to narrow down the range of *a priori* possible functional diagrams for the relevant capacity in normal brains. Glymour's argument demonstrates that, given certain plausible assumptions about our cognitive architecture, many discovery problems in cognitive science become unsolvable by our current methods (i.e., the correct functional diagram cannot be singled out from among the many *a priori* possible diagrams).

My argument, then, can be schematized in the following way:

(1) Agents lack the means to determine, *within the possibilities of their commonsensical expressions in a natural language*, all the relevant amodal representations that correspond to non-conceptual perceptual representations.

(2) Given that the implicit amodal *attitude* that causes the corresponding behavior is partly comprised of such perception related amodal representations, agents also cannot determine (again, within the possibilities of their commonsensical expressions in a natural language) the full content of the behavior-triggering attitude.

_____

(3) Hence, under the ISA's assumptions, agents can't solve the discovery problem.

If it were shown that under the ISA's own assumptions it is not possible to determine the amodal components of the perceptually-embedded judgments in the stream of consciousness, then it would follow that it is also not possible to determine the content of the attitudes that cause observable behavior. I believe this is precisely the problem the theory faces.

Granted, we seem to have something similar to introspective access to some perceptual information, just as the ISA itself assumes. Yet, apart from concepts usually expressed by nouns in natural languages (e.g., TOMATO, COIN and STREETLIGHT), and perhaps some other concepts usually expressed by certain adjectives (e.g., RED or COLORED, and I'm tempted to include concepts related to other modalities, such as HIGH PITCH, and maybe some related to the sense of touch), it is not at all clear what else should fall under the category of basic perceptual concepts. Let's, for instance, consider Eric Schwitzgebel's arguments for skepticism about introspective

perceptual knowledge (2011b).[9] As Schwitzgebel notes, there's substantive evidence regarding disagreements, among both philosophers and psychologists, and across different epochs, about introspected experience, including occurrent perceptual experience (which is particularly relevant to my present argument). Some examples are particularly striking, such as the case of disagreements about variance in the apparent shape of objects in relation to the observer's point of view. For example, some consider coins to look, from a certain angle, elliptical and flat, some consider them to look round; another example pertains to the apparently varying size of streetlights in direct proportion to the distance between them and the observer: some argue that, at least in an objective sense, their size doesn't change, while at the same time we might still get the feeling of some variation in the perceived size of some of them (for instance, Tye 2000; Noë 2004). In these disagreements, what is questioned is not the role played by the concepts that provide the most basic form of classification for the perceived object, such as the concepts COIN and STREETLIGHT (these are of the same kind as TOMATO in the example above). At the very least, in normal circumstances we shouldn't have trouble categorizing familiar objects. But concepts related to geometrical shapes and relations, as Schwitzgebel observes, seem to be more problematic.[10]

The cited disagreements among theorists of perception are the consequence of a crucial incapacity. It is plain everyday practice for individuals to consciously distinguish between concepts such as being flat and elliptical, on the one hand, and being round on the other, or between the concepts of something enlarging and of something approaching. But attempts at

[9] Interestingly, like Carruthers, Schwitzgebel denies the existence of a single, specified and particularly secure mechanism of introspection (2011a). In turn, he proposes the possibility of studying self-knowledge as the result of the combination of several second order cognitive processes (2011c).

[10] In fact, here I'm only considering Schwitzgebels' examples about basic geometrical traits of percepts, but his *Perplexities of Consciousness* contains examples of similar problems in the description of other aspects of vision and of representations from other modalities as well.

coming up with a systematic categorization of what's seemingly the same influx of perceptual representations *under commonsensical concepts* fail; such concepts are underdetermined by the introspectible evidence, so that there are no sufficient reasons for choosing one over others in its vicinity as the proper way of characterizing certain aspects of experience. This situation, I contend, doesn't improve when we move into the level of *amodal* representations.

Any judgment expressed upon introspecting our current perceptual state can represent or misrepresent a first order amodal judgment. Thus, assuming that, on the grounds of their playing distinct functional roles, sometimes we can efficaciously distinguish between judgments such as "this is the image of a flat and elliptical coin" and "that is the image of a round coin," there has to be one distinct amodal counterpart for each of them, e.g., THAT: FLAT ELLIPTICAL COIN and THAT: ROUND COIN. On the other hand, if on such grounds we were to have no particular reasons for choosing one over the other in our explanations of the mind's inner workings, then we should ask ourselves whether we are using the correct categories for describing and explaining the underlying cognitive phenomena. Or, in the case when we do possess the right categories, we should ask whether we are using them correctly.

Now, in accordance with the introspective powers that we can safely acknowledge, we might characterize the perceptual content of our attitude as, for instance, THAT: ENLARGING STREETLIGHT or THAT: APPROACHING STREETLIGHT. Yet we know that "enlarging" and "approaching" have very different meanings, so the corresponding amodal concepts should trigger different behavior. To be clear, it is not the case that I'm assuming that the words "enlarging" and "approaching," on the one hand, and the amodal representations ENLARGING and APPROACHING, on the other, should share the same meanings. This would violate the ISA model's own assumptions, since small caps concepts, such as ENLARGING, are not to be analyzed

as if they were English words, for they are supposed to be elements in the language of thought, and not in some natural language. Rather, the problem is that the same behavior, say streetlight dodging, is *introspectively* associated with different English words that express different explicit concepts. These differences, easily analyzable at the explicit level, do not seem to correlate well with possible differences between concepts at the implicit level (the level of amodal representations in the language of thought). For all we know, what some of us would describe as the behavior of tossing an elliptical coin, and what some of us tend to describe as the behavior of tossing a round coin, could be the result of the unconscious processing of the same set of amodal concepts in both cases. Or, maybe, in some of the occasions where we choose to say "elliptical," the unconscious processing utilized the representation ROUND, and vice versa.

It follows from the previous discussion that, under the ISA's assumptions and the current state of our knowledge about the mind, we the agents run up against two layers of underdetermined representations of mental content. First, occurrent perceptual content admits at least more than one possible description in a natural language. Second, whatever we choose as the correct natural-language description of the occurrent perceptual state, it is going to be our main cue for inferring the amodal (language of thought) content that truly explains behavior. Yet at this second level of interpretation, we also face the problem of choosing between possible strings of (amodal) concepts with insufficient evidence. The ISA's agent appears to be at a loss, for it is not clear at all what the criteria are for determining whether she is successfully self-ascribing attitudes.

So far, I have focused on drawing consequences for agents in general from the ISA plus our limited knowledge regarding certain respects of our mental life. But, importantly, analogous consequences can be drawn in relation to a cognitive scientist who endorses the ISA. The central

purpose of the theory is to function as a framework for characterizing self-knowledge in the context of our limited conscious access to the inner workings of the mind. Following its assumptions, the cognitive scientist should correlate the occurrent perceptual experience and behavior of a subject with some string of amodal representations in a language of thought, yet she will run into an underdetermination problem of the kind that I have stressed in this section of the paper. For instance, let's consider an agent who, upon reflection, falsely rationalizes her behavior through the self-ascription "I thought [or 'believed'] that I had to avoid that approaching tree." Given that our capability for reliably accessing basic first-order object-recognition information hasn't been questioned here, we can state that a cognitive scientist is entitled to postulate that the concept TREE did play a role in the agent's behavior. Thus, it is safe for the scientist to postulate that, in English, the truly efficacious attitude could be expressed as something that fits the blueprint "I thought that… tree…" Nonetheless, if the underdetermination arguments developed here are sound, the scientist shouldn't be able to determine what the missing components of the attitude, *as they are expressed in some model of a language of thought*.

To summarize the previous results: according to the ISA, agents attribute to themselves, upon interpreting perceptual cues, some attitude, say p, expressed in a natural language, say NL, but there is a good chance that said cues and the corresponding overt behavior are truly caused by the amodal attitude q. q is generated and processed in a language of thought, say TL. But now the cognitive scientist has to figure out what the expression of q in TL looks like, *inasmuch as the tools that we have at our disposal in a natural language allows her to model*. If p, in NL, was something along the lines of "I have to avoid that approaching tree," as we have seen, the scientist will probably be entitled to say that an element corresponding to the noun "tree" figures

in the TL expression of q (we express that element as TREE). What else figures in the TL expression of q? I have tried to make a case for the idea that we, as well as the cognitive scientist, just don't know. For all we know, we may have several competing candidates for filling those roles, yet it may also be the case that none of our candidates fit the bill.

3.2 Explaining Some Problems Away

The discussion in the previous section provides insight on certain implausible consequences of utilizing the ISA for explaining the opaqueness that affects self-knowledge. In short, the ISA's agent faces an unsolvable discovery problem, according to which she is absolutely incapable of determining the correctness of any of her beliefs about her own mind. But such a conclusion clashes with our regular experience regarding the self-ascription of attitudes: it is not just that it seems hard to imagine how we could perform any sort of task, regardless how minimal, if we were to lack any resource for, at the very least, successfully keeping track of our own thoughts on more occasions than those in which we fail to; we even have the experience of revising beliefs and of becoming aware of our own confabulatory practices. I think that the implausibility of these results is the consequence of conflating two different senses in which, in the context of the philosophy of cognitive science, we utilize the terms that denote attitudes, such as "belief." Furthermore, I think that if this confusion is adequately solved, then, in the context of a theory such as the ISA (or of a modified version thereof) another distinction is in place, i.e., a distinction between two forms of opaqueness of the mind.

When we correct our own or others' ascriptions of attitudes, we do it on the basis of the same kind of cues that we use for making the wrong ascriptions. From our *commonsensical* standpoint, the corrected ascription does not include any belief that we couldn't in principle have ascribed to others or ourselves in the first place. Amodal attitudes in a language of thought do not

follow this logic for the simple reason that in order to pick them out as plausible *explanans* for occurrent behavior and cognitive states, we rely on different indices (the ones provided by a theory in cognitive science and the relevant empirical data). There seem to be, then, two senses in which we can talk of attitudes when discussing self-knowledge in the context of the ISA. One corresponds to our everyday practices of explaining behavior only with the aid of our commonsense. The other corresponds to the amodal representations that the ISA, as well as other theories that interpret data from cognitive science, postulates as the true causes of behavior, i.e., amodal representations in a language of thought. The theory's assumed *belief realism*, nonetheless, collapses these two senses of the word "belief", since it stipulates that that which we consider to be beliefs, as well as attitudes in general, at the explicit, commonsensical level, is actually realized as amodal representations at the implicit level.

As I have argued here, from the ISA's own assumptions there ensues a serious epistemological problem. We simply can't solve the problem of determining the beliefs that truly explain behavior (or maybe even the more basic problem of determining our own beliefs about the stream of perceptual representations) at the implicit level, where amodal representations are supposed to be processed. This incapacity can be characterized as a form of radical, insurmountable opaqueness of the mind.

Now, avoiding the conflation of the two senses of "belief" (and the other attitudes) outlined above could allow us to draw a distinction between two ways in which the mind is opaque to us. Moreover, I think there is room within Carruthers' own views to accommodate these distinctions, provided that he is also willing to drop the ISA's assumed belief realism (which, of course, doesn't prevent one from endorsing other versions of realism about significant portions of our *folkpsychological* conception of the mind). As he notes, the

mindreading/metarepresentational system has a model of its own seemingly transparent, i.e., introspective, access to the first order mental contents of the mind, which seems to be widespread among humans, across different cultures and times (2009, pp. 126-127; 2011, pp. 25-32). In fact, Carruthers hypothesizes that the development and persistence of this (in general) *erroneous* self-conception is due to its enabling a more efficient process of self-interpretation. If the metarepresentational system were to project a self-image where the subject is interpreting herself just as if she were interpreting others, its own operations would be seriously hampered, since it would constantly be forced to engage in reflective revision (for we know that interpreting other agents is a highly fallible task). To put it bluntly: we are better off living under a false conception of our access to our own minds, a kind of Cartesian illusion of transparency, because it is cognitively more efficient.

Within the sphere of our folkpsychological way of keeping track of our own thoughts, there exists a particular form of opaqueness: there are many ways in which we can rationalize our own behavior in a natural language, many of which are underdetermined. We can rectify our self-ascriptions by recognizing that we could have used a better description (in a natural language) of the propositional attitude. But from the scientist's point of view, these descriptions have little to do with the representations that the brain itself produces and manipulates. So, while the metarepresentational system infers the belief "I wanted to eat that red apple" from perceptual and quasi-perceptual cues (and let's remember that the inner speech expression "I wanted to eat that red apple" might be one of those cues), the systems operating at the first order level might be processing the representation … THAT: … RED … APPLE (the blanks are meant to represent the non-introspectible contents), whose process of formation, for all we know, might not resemble at all the process of inferring from multiple perceptual cues (in many occasions, corresponding to

different modalities) at the conscious level.

We can think of our reflective, second order explanations of behavior as linked to first order mental representations and processes in two distinct, but concomitant, ways. One is a linkage in content. As we have seen, the ISA theory allows for introspection of some first order perceptual content that constitutes some of our non-introspectible first order attitudes. These introspected contents, in turn, determine the metarepresentational system's inferences that result in the self-ascription of some attitude, but not necessarily the correct one (i.e., if perceptual content has the embedded judgment THAT: RED APPLE then the inferred attitude will probably be, in some way, about a red apple). The other way in which the two cognitive levels are linked is, of course, causal. The attended perceptual contents cause the interpretive inference (while, at the same time, determining its outcome, even if weakly). The general problem, as I have argued, is that, even provided that the personal-level self-attributed belief shares some of its content, it is not enough to specify the underlying amodal attitude that *causes* behavior "in the right way," at least not when what we are looking for is to represent the cause of behavior in accordance with a scientific theory.

4. Consulted Works

Baars, Bernard J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Block, Ned. (1995). On a Confusion about the function of consciousness. *Behavioral and Brain Sciences*, 18: 227–47.

Carruthers, Peter. (2006). *The Architecture of Mind*. Oxford University Press.

Carruthers, Peter. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences* 32: 121-182.

Carruthers, Peter. (2011). *The Opacity of Mind. An Integrative Theory of Self-Knowledge*, Oxford University Press.

Carruthers, Peter. (2014). On Central Cognition. *Philosophical Studies* 170: 143-162.

Carruthers, Peter. (2015). *The Centered Mind. What the Science of Working Memory Shows Us About the Nature of Human Thought*. Oxford University Press.

Fodor, Jerry. (1975). *The Language of Thought*. Harvard University Press.

Fodor, Jerry. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Glymour, Clark. (1994). On the Methods of Cognitive Neuropsychology. *The British Journal for the Philosophy of Science*. Vol. 45, No. 3: 815-835.

Noë, Alva. (2004). *Action in Perception*. MIT Press.

Prinz, Jesse. (2007). All Consciousness is Perceptual. In: *Contemporary Debates in Philosophy of Mind*, edited by Brian P. McLaughlin and Jonathan Cohen. Blackwell Publishing. 335-357.

Prinz, Jesse. (2012). *The Conscious Brain. How Attention Engenders Experience*. Oxford University Press.

Schwitzgebel, Eric. (2011a). Introspection, What? In: *Introspection and Consciousness*, edited by Declan Smithies, and Daniel Stoljar, Oxford University Press.

Schwitzgebel, Eric. (2011b). *Perplexities of Consciousness*. MIT Press.

Schwitzgebel, Eric. (2011c). Know Your Own Beliefs. *Canadian Journal of Philosophy* 35 (supplement): 41-62.

Shoemaker, Sydney. (1994). Self-Reference and Self-Awareness. In: *Identity, Cause and Mind*. Cambridge University Press.

Sperber, Dan & Wilson, Deirdre. (1986). *Relevance: Communication and Cognition*. Blackwell.

Tye, Michael. (2000). *Color, Consciousness, and Content*. MIT Press.