On knowing your own beliefs: A representationalist account

Peter Carruthers

This chapter first outlines the interpretive sensory-access (ISA) theory of self-knowledge, developed and defended at length in my 2011 book, *The Opacity of Mind*. It then considers and critiques a pair of competitors, each of which regards the relationship between one's beliefs and one's knowledge of them as constitutive rather than relational. The first is a form of dispositionalism about belief. The second builds on the distinction drawn by cognitive scientists between so-called "System 1" and "System 2" reasoning processes.

1. The ISA theory

This section will describe the interpretive sensory-access theory and sketch some of the evidence in its support, before explaining its commitments regarding the nature of belief.

1.1 The theory

Carruthers (2011) maintains that the system that is employed when one identifies and attributes mental states to oneself is none other than the mindreading system that underlies one's capacity to attribute mental states to other people. Moreover, this system only receives sensory input (including visual, auditory, and motor imagery as well as perceptions of the world and of one's own body). It follows, then, that one's mindreading system lacks direct access to one's underlying attitudes. The latter operate entirely in the background, competing with one another to help influence the contents of consciousness, but remaining inaccessible to the mindreading faculty. Yet there is no other system or mechanism that gives one access to one's own propositional attitudes. In order to attribute thoughts to oneself, then, the mindreading faculty is forced to interpret the available sensory evidence. This can concern one's physical circumstances and overt behavior, or it can involve one's own visual imagery, affective feelings, and inner speech. The result is that all access to one's own propositional attitudes is sensory-based and interpretive in nature. Carruthers (2011) calls the ensuing theory the "interpretive sensory-access" (ISA) theory of self-knowledge.

The ISA account builds on a number of well-established findings concerning the architecture of human (and animal) minds. One is the global broadcast of attended sensory representations to a wide range of different systems in the brain, including those for forming memories, for drawing inferences, and for providing affective and evaluative responses. (This was initially proposed by Baars, 1988, but has been confirmed by a great deal of experimental work since then.) Another is the architecture of working memory, which utilizes the same framework of global broadcasting and top–down attentional resources to sustain, rehearse, and manipulate sensory-based representations (Baddeley, 2006; Jonides et al., 2008). Crucially, it appears that there is no *other* workspace in which propositional attitudes can themselves be active and accessible to a wide range of consumer systems.

It should be stressed that the contents of working memory are sensory *involving*, not purely sensory in nature. On the contrary, in the course of normal perception concepts become bound into the incoming sensory representations and (when the latter are attended to) are broadcast along with them. As a result, we see something *as* a house or *as* a horse, and we hear the speech of others as imbued with meaning and communicative intentions. (Since such states can give rise to semantic memories and play causal roles somewhat like those of a judgment, I refer to these as "sensorily-embedded judgments". These will, as such, be available as input to the mindreading faculty and can be self-attributed in a non-interpretive manner, thus requiring a small qualification in the scope of the ISA theory.) Something similar is true of one's own visual imagery and one's own inner speech.

1.2 Support for the ISA theory

Carruthers (2011) reviews a wide range of evidence in support of the interpretive sensory-access theory, from across cognitive science. This includes evidence of the nature and sensory basis of global broadcasting and working memory, the nature and sources of our capacities for metacognitive control of learning and reasoning, alleged dissociations between self-knowledge and other-knowledge in autism and schizophrenia, brain-imaging evidence of the systems involved in self-attribution and other-attribution, and more. In addition, many competing theories of self-knowledge are discussed and critically evaluated. An inference to the best explanation across this entire data-set issues in powerful support for the ISA theory, and a correspondingly

strong case against its rivals.

One especially important strand of support is provided by numerous studies that demonstrate how easily people will *confabulate* about their current or very recently past thoughts, sincerely attributing to themselves judgments, goals, or decisions that we know on independent grounds they lack. The patterning in this data is exactly as the ISA theory would predict: people misattribute thoughts to themselves in circumstances where they have been provided with sensory cues of just the sort that might mislead a third-person mindreader. In contrast, none of the other theories of self-knowledge can explain this patterning (or at least, not in their own right—some can piggy-back on the success of the ISA theory).

There is also evidence that people's speech actions do not directly express their underlying thoughts but rather (like all other actions) are subject to a variety of competing motivational influences. So when people say what they think (either aloud or in inner speech) this provides some *evidence* of their thoughts without by any means providing direct and reliable access to those thoughts, either to others or to themselves. (This remains true even when people's statements are acknowledged to be sincere.)

For example, extensive use has been made of the counter-attitudinal essay-writing paradigm, in which subjects are induced to write an essay arguing for the opposite of what they actually believe. In so-called "free choice" conditions in which it is emphasized to subjects that they are writing their essays freely (and provided they believe that their essays might result in something bad, either for others or for themselves), they will later shift their reported attitudes on the topic quite markedly. For example, they might change from being strongly opposed to a rise in university tuition fees to a position of neutrality, or even to mild approval. Such effects are generally strong and robust, and have been replicated hundreds of times. Carruthers (2011) argues that people are not altering their underlying attitudes, but are appraising their potential speech acts in the manner of Damasio (1994) and are selecting the one that "feels best" to them. In the circumstances, this will be one that presents their previous essay-writing in such a way as to minimize the harm done. Carruthers also argues that such data are deeply problematic for all of the competing theories of self-knowledge.

1.3 ISA's realist commitments

The account of the mind presupposed in Carruthers (2011) is unabashedly realist. It assumes, in particular, that propositional attitudes are discrete structured representational states composed out of component concepts. Beliefs, for example, are not just complex dispositions of a certain sort. Rather, they are structured categorical states that give rise to various dispositions in the presence of other such states together with the normal operations of the mind. Someone who believes that there is beer in the fridge, for instance, has a stored state composed of the concepts BEER and FRIDGE that represents that there is beer in the fridge and which, when active, is apt to interact with an occurrent desire for beer so as to issue in fridge-opening behavior, and which is apt to cause surprise if the fridge turns out to be empty.

From this perspective there is an important distinction between *explicit* and *implicit* beliefs, however (Dennett, 1978). Explicit beliefs are stored representations of the above sort. Implicit beliefs are those that one *would* readily form and act on if circumstances arose, given the explicit beliefs one *actually* has. Many of us can be said to believe things that we have never considered and almost certainly don't have representations of stored in our brains. (To mention some of Dennett's examples: that zebras in the wild don't wear overcoats; that one has never danced with a movie star; that one has never been to the moon; and so on.) But these are obvious entailments of things that one does actually believe. Given one's explicit belief-base, one is disposed to add to it such beliefs immediately and unreflectively if asked. One can therefore be said to have believed them (tacitly) all along.

In many cases, of course, it can be hard to tell which of one's beliefs are explicit and which are merely tacit, in the absence of careful controlled experimentation. If you ask me how many planets there are, for example, I shall answer, "Eight." But which explicit belief underlies this answer? Is it that there are eight planets? Or is it that the number of planets is one less than nine (formed when Pluto was demoted)? Or both? Such questions cannot be answered introspectively, and the behavioral dispositions associated with each are quite similar (except that one's reaction time might be slightly longer in the second case, since one needs to *infer* that one less than nine is eight). These difficulties are epistemic, however, not metaphysical. It creates no problems for realism about belief to claim that there is always a fact of the matter, even if the facts can sometimes be hard to know.

It is also important to note that realists about propositional attitudes can allow that beliefs

admit of degrees, in at least two respects. One concerns the attitude of believing. One can believe something *firmly*, or with certainty, or one can believe it *weakly*, or tentatively, or one can occupy any of the points between these extremes. In the case of explicit beliefs, the varying degrees might be realized in the strength of the relevant memory traces, for example, together with the extent to which the belief receives support from one's other explicit beliefs. But the second way in which beliefs can admit of degrees concerns the *content* of belief. On many accounts of the latter, content will depend, in part, on the inferential and conceptual liaisons between a given belief state and others (Millikan, 1984; Block, 1986). If someone lacks a significant number of these (say a young child who has not yet acquired many of the related concepts) then if might be correct to say that she does not *fully* have the belief in question, but rather a simpler, related belief.

(In addition, of course, a realist can allow that there can be *kinds* of mental state whose functional profile places them in between belief and something else—desire, say. Online perceptual judgments of goodness may provide one good example. See Carruthers, 2011.)

Given that the interpretive sensory-access (ISA) theory of self-knowledge presupposes realism about the attitudes, and about belief in particular, it is natural to wonder whether dispositionalism about belief presents a viable threat to the theory. This will form the topic of Section 2.

2. Dispositional theories of belief

If beliefs are complex dispositions, then an account of self-knowledge of belief can be defended that is *constitutive* in nature (Schwitzgebel, 2011). For if part of what it *is* to believe that *P* is that one is disposed to believe of oneself that one believes that *P*, then knowledge (in the sense of reliably caused true belief) of one's beliefs will be partly constitutive of having beliefs at all. On some versions of this account, knowledge of one's own beliefs will not be an epistemic *achievement*, and one might expect as a result that it should be especially reliable and authoritative. This section will first evaluate dispositional accounts of belief in their own right, before turning to their implications for first-person epistemology.

2.1 The metaphysics of belief

Dispositionalists about belief maintain that beliefs are complex multi-track dispositions (Schwitzgebel, 2002). To believe that there is beer in the fridge, on this account, just *is* to have a distinctive syndrome of behavioral, affective, and cognitive dispositions. Someone with that belief is disposed to approach the fridge if desirous of a beer and will assert that there is beer in the fridge if asked; she will be disposed to experience surprise if she opens the fridge to find an absence of beer; and in the same circumstances (depending on her other beliefs) she might be disposed to infer that her partner has consumed it all. Note that among the cognitive components of the dispositions that constitute belief will be a disposition to believe of oneself that one has that belief, and that among the behavioral components will be a disposition to avow that belief in speech, or to assert that one has it.

One apparent problem for dispositionalism about belief is raised by Ramsey et al. (1990), who point out that there can be situations in which a given action or other mental state could be a manifestation of either one of two distinct dispositions. They give the example of Inspector Poirot, who deduces that the butler is lying when the latter says he was absent on the night of the murder because he was staying at a hotel in town and returned to work on the morning train. Poirot has two beliefs that are each individually sufficient to uncover the lie: he believes that the town hotel is closed for the season, and he also believes that the trains were not running because of a drivers' strike. So his conclusion, *the butler is lying*, could be a manifestation of either one (or both) of the disposition-sets that (allegedly) constitute the two beliefs. We surely think that there will be a fact of the matter about *which* of Poirot's beliefs led to the inference. But if beliefs just *are* dispositions (including dispositions to conclude that the butler is lying in circumstances of this sort), then there seems no room for a further fact of the matter.

A dispositionalist can reply that distinct disposition-sets can be distinguished by the differing counterfactuals to which they give rise. For example, suppose it were true that had Poirot been asked just before concluding that the butler is lying, "What are you thinking?", he would have replied, "That the hotel is closed for the season." This would then show that it is the belief expressed by the latter that led him to question the butler's truthfulness. And if he would *not* have acknowledged thinking about the train strike, then this can show that the latter belief did not, on this occasion, manifest itself in the conclusion that the butler was lying.

If the objector continues to press, demanding to know in virtue of what (on a

dispositionalist account) these counterfactuals are true or false, it can be replied that distinct disposition-sets will have different categorical bases. These can serve to ground the counterfactuals in question, since there should be a fact of the matter about which categorical base was causally involved when Poirot drew his inference. This reply should be available to dispositionalists whether or not they believe (as the realist does) that the categorical bases of belief are discrete structured representational states. But on this view, note, the categorical base is not itself the belief. The disposition *of which it is* the categorical base is that. The appeal to a categorical base is evidential rather than constitutive, enabling us to discriminate between beliefs that overlap with one another in their dispositions.

A bigger problem for dispositionalism might seem to be that the dispositions that would need to constitute any given belief are indefinitely malleable and open-ended. This follows from a point noticed by early functionalists about the mind, that what a belief will dispose you to do depends entirely on your desires and your other beliefs. As these latter states are varied, a whole new suite of dispositions-to-behave (as well as new dispositions-to-think and dispositions-to-feel) will come into view. This makes it hard to believe that our conception of belief is that of a complex dispositional state, since the relevant disposition is impossible to specify.

It might be replied that our conception of any given belief presupposes a *normal* background of goals and other beliefs; hence the dispositions in question are limited to those that obtain in normal circumstances. It seems implausible that this maneuver can work, however, since what counts as "normal" is itself so variable and flexible. What people will *normally* believe depends a great deal on their varied circumstances, and likewise the range of normal desires is quite wide and varied. Moreover, even when beliefs are far from normal we often have no difficulty in predicting what someone will do. If someone believes that the quickest way from his kitchen to his car is through the front door, for example, then in normal circumstances we can predict that he will take that route if he wants to drive to the shops. But if he believes, quite unusually, that there is a yawning chasm just outside his front door, then we can predict that he will seek some other way of leaving the house.

At this point dispositionalists should emphasize the distinction between what beliefs *are* (multi-track dispositions) and how we have knowledge of them. It can be allowed, in particular, that the latter depends importantly on our own dispositions. In fact dispositionalists can buy into

at least a limited form of *simulationism* about our knowledge of the dispositions that constitute a given propositional attitude (Nichols and Stich, 2003). In order to know what someone with a given belief will do, or feel, or think, on this view, one entertains that belief oneself as a supposition and lets one's own planning, affective, and cognitive systems respond accordingly. Other things being equal (that is, unless one has reason to vary more than one component of the target subject's attitudes), we can attribute the resulting dispositions to the belief in question. On this view, then, we can come to know *which* dispositions are constitutive of a given belief by simulating the possession of that belief.

So far it might seem that dispositionalism is standing up well under attack. But contrast beliefs with character traits like irascibility or generosity, which really are dispositions. Someone who is irascible is someone who is disposed to become angry easily, and someone who is generous is disposed to make sacrifices for others. Now, explaining someone's behavior by appealing to a trait ("He snapped at you because he is an irascible person") is merely explanation by subsumption under a generalization. The explanation says, in effect, that the reaction was of a sort that one might have expected, because that is the way in which the agent *generally* responds. Explanation by appeal to belief, in contrast, is much more meaty. Such explanations seem, in fact, to cite a token cause of the explanandum. (Note that this point is independent of the fact that beliefs are supposed to be multi-track dispositions whereas traits are single-track dispositions.) For example, consider explaining why Poirot came to believe that the butler is lying by citing the former's belief that the local hotel is closed for the season. This tells us *why* he drew the conclusion. It does not merely tell us that the conclusion was of a sort that Poirot normally draws.

This point can be expanded into a much broader critique of dispositionalism. The problem is that the theory lacks the resources to explain the systematic patterning in the dispositions that are alleged to constitute any given belief. Rather, the belief just *is* that pattern of dispositions. On a realist account, in contrast, the patterning can be explained in terms of the component structure of the state and its interactions with other componentially structured attitudes. Since our ordinary view is that an appeal to someone's belief can genuinely explain why he acts, feels, and thinks as he does (as opposed to being *constituted* by dispositions to act, feel, and think in just those ways), this is a significant problem for dispositionalism.

Given these problems, we need to be provided with positive arguments in support of dispositionalism. The main argument offered in Schwitzgebel (2002) is that the account can help us make sense of cases of "in between" believing. But recall that realists can likewise allow that beliefs admit of degrees, along two dimensions—the attitude component, which can be stronger or weaker, and the content component, which can vary in its conceptual liaisons more or less from the paradigm of a content of a given sort. Schwitzgebel therefore needs to confine himself to examples of in-between believing that fall into neither of these categories.

In fact some of the examples that Schwitzgebel (2002) gives are not well characterized as forms of belief at all, but rather as kinds of know-how, or skill (what psychologists call "procedural knowledge"). This is true of the Spanish-language student Ellen, for example, who *says* that all nouns ending in "a" are feminine, but who uses words like "bolchevista" correctly as masculine when required. This is not an instance of in-between belief, but of someone whose explicit belief doesn't match her practical ability. The example of Geraldine, in contrast (who sometimes acknowledges that her son smokes marijuana and sometimes sincerely denies it), is best explained in terms of the disconnect that can exist between belief and assertion. For as we noted in Section 1.2, speech is an action, and many factors can influence people into saying things other than they believe.

What is true, of course, is that those who *interpret* others can often be "in between" on the question of whether or not a subject believes something. Some considerations can favor saying that he does, whereas others favor saying that he does not. But this is epistemic, not metaphysical. All it shows is that our grounds for *ascribing* a belief to someone can often be "in between". It does not show that the actual mental state of the subject can be in between belief and nonbelief (setting aside the two ways in which beliefs admit of degrees noted earlier, which are equally consistent with a realist account).

I suggest that Schwitzgebel's appeal to in-between believing may actually backfire on him, indeed. For as we noted, among the dispositions that are allegedly partly constitutive of believing that *P* are dispositions to assert that *P* and dispositions to believe of oneself that one believes *P*. It follows, therefore, that anyone who lacks such dispositions does not fully believe that *P*. In particular, it follows that someone suffering from severe aphasia, whose core language abilities are destroyed, cannot fully believe *anything*. This conclusion is hard to accept. For such

people can lead otherwise normal lives, they can be adept at communicating by pantomime, they can continue to be responsible for the family finances, and they can reason successfully about causes and effects (Varley, 1998, 2002).

Likewise, if Schwitzgebel is right then it follows that nonhuman apes cannot fully believe anything, either, given that they lack a capacity for language and lack the concept of belief (as most comparative psychologists currently accept). This conclusion is implausible. Granted, apes may not share any of our more sophisticated concepts. But there is little reason to doubt that they can share some of our basic concepts like *grape* and *ground*. If an ape believes that there is a grape on the ground before her, I can see no reason to insist that she does not *really* and *fully* believe this, because she lacks the capacity to express her belief in speech and is incapable of ascribing that belief to herself.

2.2 The epistemology of belief

I have argued that dispositionalist theories of belief face significant problems. But let us suppose, for argument's sake, that they did not. Setting aside the problems, let us consider whether, by claiming that a disposition to attribute a belief to oneself is partly constitutive of possessing it, dispositionalism provides a viable alternative to the interpretive sensory-access (ISA) theory of self-knowledge. I shall argue that we can set such accounts a dilemma, depending on the nature of the relationship that is thought to obtain between the categorical base of the disposition and one of its manifestations: namely, ascribing the belief in question to oneself.

On the one hand, one might think, as does Shoemaker (1994), that the relationship is direct and immediate. On this view, the relationship is not inferential, but is part of the wider functional role of belief itself. If true, this might warrant a claim of special reliability and authority over our own beliefs. But this model is empirically inadequate given the extensive evidence of confabulation in self-reports of belief. If it is an intrinsic part of believing that *P* that one should immediately believe of oneself that one believes that *P*, then it becomes quite mysterious why people should so easily be induced to ascribe some other belief to themselves instead. In fact, the critique of alternative theories of self-knowledge mounted by Carruthers (2011) from this direction is just as powerful when targeted at Shoemaker's account.

The other possibility is to allow that the causal route to the manifestation of the

disposition to ascribe a given belief to oneself is indirect, and depends upon the interpretive work of the mindreading faculty (Schwitzgebel, 2011). This can render dispositionalism consistent with Carruthers' interpretive sensory-access (ISA) theory of self-knowledge, and also with evidence of frequent confabulation. But now the claim that a disposition to ascribe a belief to oneself is partly constitutive of possessing such a belief is epistemically inert. No new source of reliability, justification, or authority is introduced by the constitution claim. On the contrary, all of the epistemic work is done by the processes of self-interpretation that underlie the disposition.

(Of course it might also be claimed, as Schwitzgebel, 2011, does, that there are *multiple* mechanisms underlying the self-ascription of belief, of which self-interpretation is only one. But this now inherits all of the problems that attach to "dual method" theories of self-knowledge, laid out at some length in Carruthers, 2011.)

More might be said about the circumstances in which self-ascriptions of belief are likely to prove reliable, of course. Thus Schwitzgebel (2011) suggests quite plausibly, for example, that self-reports are more likely to be true in circumstances that are evaluatively neutral. But the plausibility of such suggestions owes nothing to dispositionalism about belief, and can just as easily be embraced by realists who endorse the ISA account. Hence they provide no support for the former.

I conclude, therefore, that not only is dispositionalism implausible as a theory of the metaphysics of belief, but it is, in any case, incapable of mounting a viable epistemological challenge to the interpretive sensory-access theory of self-knowledge.

3. System 2 belief

Other views also entail that knowledge of some kinds of belief is constitutive rather than relational. One, in particular, builds on the literature in cognitive science concerning dual systems of reasoning to claim that some so-called "System 2" events are constituted as beliefs by our own interpretive activity. While this view is criticized at some length in Carruthers (2011), Frankish (2012) has attempted to reply to those criticisms. This section will briefly consider the case for System 2 belief as well as the earlier criticisms, before critiquing Frankish's reply.

3.1 Dual modes of reasoning and believing

Many cognitive scientists have converged on the claim that there are two systems for reasoning and decision making in humans, often called “System 1” and “System 2” (Evans and Over, 1996; Stanovich, 1999, 2009; Kahneman, 2011). System 1 is really a set of systems that are fast, parallel, and unconscious, delivering seemingly-intuitive answers to reasoning problems in ways that operate outside of our awareness. System 2, in contrast, is slow, serial, and conscious, enabling us to reflect on reasoning problems, to implement acquired reasoning strategies, and to access explicit beliefs about appropriate normative standards for reasoning. Increasingly, System 2 is thought to be dependent on the operations of sensory-based working memory, allowing us to ask ourselves questions in inner speech, rehearse previously successful solutions in speech or other forms of imagery, and manipulate alternative representations of the problem (Evans, 2008, 2010).

Frankish (2004, 2009), following Dennett (1978) and Cohen (1992), argues that some System 2 events can have an influence on one’s future reasoning, decision making, and acting just as if they were beliefs. As a result, he thinks, they *are* beliefs, of a System 2 kind. For example, someone who sincerely asserts, “It wouldn’t be bad if tuition were raised”, may thereafter regard herself as *committed* to the truth of what she has asserted, even if the assertion were initially a confabulation of some sort. Remembering that commitment, and feeling obliged to execute her commitments, she may thereafter constrain her actions (including her System 2 reasoning actions) accordingly. This will lead her to reason and act just as would someone who believes that raising tuition wouldn’t be bad. Note that if we accept that she really does have such a belief as a result, then this will be a belief that is *constituted* by the way in which the agent herself construes the initial performance. So these will be beliefs that subjects have a special epistemic authority over.

Interpreting oneself as committed to the truth of a proposition is by no means the only way in which a new System 2 belief can be constituted, on this sort of account. Indeed, while one might naturally construe a public assertion as a commitment, this is much less likely in connection with a sentence in inner speech. But often such inner utterances will be heard as expressing a judgment, or as manifesting a belief. So someone who thinks covertly to herself, “It wouldn’t be bad if tuition were raised”, may take herself to be expressing the corresponding judgment, even if the performance is a confabulation, and she has no such belief. But if she

believes that she has made that judgment, and has a standing desire to think and act reasonably and coherently, then she, too, may thereafter constrain her reasoning and acting just as would someone who believes that raising tuition wouldn't be bad.

One apparent oddity of this view is that the event that is said to become a judgment that *P* in virtue of being taken as such (or in virtue of being taken as a commitment) doesn't have the content *P*, but rather the content, *I am judging that P* (or the content, *I am committing myself to the truth of P*). This is, one might think, the wrong sort of content to be the content of a judgment that *P*. But it can be said in reply that the event in question has two distinct contents, possessing one of them in virtue of possessing the other. Under interpretation, when someone entertains the inner assertion, "It wouldn't be bad if tuition were raised", it *seems* to her that she is judging that it wouldn't be bad if tuition were higher (in this case falsely in the first instance, let us suppose, since she is participating in the "free choice" condition in a counter-attitudinal essay-writing experiment). So the content that attaches consciously to the performance is, *I am judging that it wouldn't be bad if tuition were raised*. This can count as what Carruthers (2011) calls a "sensorily-embedded judgment" with that content. But because she judges that, and because she has a second-order desire to act in ways that are rational in light of her judgments, she thereafter constrains her thinking and acting just as if she believed that a tuition increase would not be bad. As a result, she qualifies as having a belief with the content, *it wouldn't be bad if tuition were higher*. But she has this first-order belief in virtue of unconscious interactions among her second-order beliefs and desires. The initial judgment, then, has *both* of the contents, *it wouldn't be bad if tuition were higher*, and, *I am judging that it wouldn't be bad if tuition were higher*, and it has the former in virtue of having the latter.

Notice, however, that although this means that the subject is not mistaken about *what* she first-order believes, she *is* mistaken about *which event* is the event of her making that judgment. For when she initially entertains the sentence in inner speech and hears this as expressing a judgment, she takes the judgment to be distinct from the episode of inner speech itself, just as she does when she hears herself speaking aloud. So she will believe that her judgment precedes and causes the verbal performance. But in the case in question, there is no such judgment. So although, on a System 2 account of the constitution of belief, she does know *what* she believes (and she knows this authoritatively, since it is her belief about what she believes that—together

with her second-order desires—makes it the case that she believes it), she is mistaken about when and how she believes it. And she makes such mistakes systematically, about all of her System 2 beliefs.

3.2 The functional profile of belief

Carruthers (2011) does not challenge the existence of the attitude-like processes that Frankish describes. Nevertheless, he subjects the idea of System 2 attitudes to sustained criticism, arguing that the System 2 events in question don't achieve their effects in the right kinds of way to qualify as attitudes of the appropriate sort. The point is easiest to see in connection with alleged System 2 decisions, but it generalizes to the case of belief and other attitudes. For we think that one of the features distinctive of decisions is that they should *settle* what is to be done (Bratman, 1987, 1999). Once a decision has been taken, practical reasoning about whether or not to act ceases. All that remains, in cases where the act cannot be performed immediately, is to reason about *how* to act, or how to *implement* the decision.

System 2 "decisions", in contrast, fail to fit this profile. Suppose that following a period of System 2 reasoning I say to myself, "So, I shall go to the bank." (And let us suppose for argument's sake that this is a confabulation of some sort, which does not reflect an underlying System 1 decision to go to the bank.) Under interpretation this is heard as expressing a decision. But this then needs to interact with the goal of being a strong-willed person to issue in a *subsequent* decision to go to the bank. The practical reasoning here might look something like this: I have decided to go to the bank; I want to be the sort of person who does what he decides; so I shall go to the bank. This seems to disqualify the event of saying to myself, "So, I shall go to the bank", from counting as a decision of any sort.

Similar points can be made about the way in which decisions should influence and guide subsequent reasoning. Once one has decided to do something, this should constrain one's choices about what else to do and should guide one's reasoning about how to implement the decision. Consider a case where the putative decision, "I shall go to the bank", leads me to think a moment or two later, "So, I need to get the car keys." This looks superficially like a decision guiding reasoning about how to implement the decision. But in reality the reasoning looks something like this: I have decided to go to the bank; I want to be the sort of person who implements his

decisions; if I am to implement my decision to go to the bank, then I need to find some way of getting to the bank; driving would work; so I need to get the car keys. This does not have the functional profile of a *decision* that guides reasoning, but rather that of a belief about a decision (combined with a desire to implement the decision) guiding reasoning.

Carruthers (2011) also takes up a reply made by Frankish (2009) to similar arguments. This is that System 2 beliefs can fit the functional profile of belief *at the System 2 level*. For note that all of additional reasoning detailed in the examples above will generally take place unconsciously, and will be composed of System 1 attitudes. This is fully consistent with the view (endorsed by both Frankish and Carruthers) that System 2 processes are not independent of those of System 1, but are rather implemented in the latter. So it can be said, in particular, that a System 2 decision need only close off further *System 2* practical reasoning, thus fitting the profile of a System 2 decision.

Carruthers (2011) argues that this reply fails, because of the absence, in many cases, of appropriate System 2 events. For example, not only is it part of the functional profile of a decision to close off further practical reasoning, but it is also part of that profile that decisions should result from interactions among suitable beliefs and desires. Now *sometimes* when one thinks, “I shall go to the bank” this might have been preceded by appropriate System 2 events, such as entertaining in inner speech the sentences, “I need cash”, and, “To get cash I should go to the bank.” But this is by no means necessary, or even the normal case. Sometimes all that happens at the System 2 level is that the sight of an empty wallet is followed by, “I shall go to the bank.” So this event doesn’t have the right functional profile to be a decision after all, not even at the System 2 level.

3.3 Functional profiles revisited

Frankish (2012) makes a number of replies to the critique mounted in Carruthers (2011). Thus in response to the point just made, he says that the beliefs and desires that interact to issue in a decision need not be *activated* ones. Rather, they can be *dormant*, or so-called “standing state” attitudes of the sort that one continues to possess while asleep or comatose. But this is surely a mistake. For dormant attitudes cannot be causes. Granted, while active on previous occasions they may have contributed to setting up a habit, say, which might thereafter bypass the normal

attitude-involving functional roles. But that is not what is in question here. The sight of an empty wallet does not activate a *habit* of saying, "I shall go to the bank." Rather, in the example in question it serves to activate beliefs and desires that interact unconsciously, at the System 1 level, to issue in such a performance. So the original objection stands: this event does not have the right System 2 profile to count as a decision, even when attention is confined to the System 2 level.

Frankish (2012) also argues that there are useful generalizations that can be captured in terms of System 2 attitudes that would be lost if we only recognized the existence of System 1 attitudes. In part this is because of the multiple realizability of System 2 processes. For example, in some people (or on some occasions) an assertion in inner speech that *P* might be heard as expressing the judgment that *P*, whereas on others it might be heard as making a commitment to the truth of *P*. And then the ways in which these events lead subjects to constrain their System 2 reasoning and their behavior as if they believed that *P* will differ accordingly—in the one case depending on a desire to reason and act as one's judgments rationally require; in the other case depending on a desire to reason and act in such a way as to execute one's commitments.

It should be stressed, however, that the question is not whether it is *pragmatically* useful to think and speak as if System 2 events were judgments and other attitudes. Since we are often ignorant of the underlying System 1 processes, there is no doubt that it is. If someone asserts that *P*, seemingly sincerely, then we gain explanatory and predictive purchase if we thereafter assume that she believes that *P*. Likewise if I assert that *P* in inner speech, I can usefully take myself thereafter to believe that *P*. Often, no doubt, this is because the person does (and did prior to the utterance) believe that *P*, and our ascription of the belief that *P* will be true. But on other occasions the explanatory and predictive purchase has a different psychological basis in the person's meta-attitudes (such as believing herself to have made a commitment to the truth of *P*). Either way, the person's behavior and their System 2 reasoning in future are likely to be somewhat similar.

The real question is whether recognition of System 2 attitudes is *scientifically* warranted. For this is what the real existence of such attitudes should turn on. In particular, are there law-like generalizations that can only be captured in such terms? If there were law-like generalizations specific to particular attitude-contents, then the answer to this question might

very well be positive. For example, if there were law-like generalizations concerning the belief that it wouldn't be bad if tuition were raised, in particular, then only the ascription of such a (System 2) belief to the subject would enable us to subsume the subject's behavior within those generalizations, enabling us to capture what is common with cases where the subject has a System 1 belief with the same content. It is doubtful whether psychology finds any role for such generalizations, however. For the standard way of predicting what someone with a given belief will think or do is to assume that belief for oneself, and then to reason on one's own behalf (with suitable adjustments for the context, and for other differences from the target), attributing the result to the other person. (This is the core truth in simulationist models of mindreading; see Nichols and Stich, 2003; Carruthers, 2011, 2013.)

We can, of course, *see* something in common between a case where someone acts on their System 1 belief that higher tuition doesn't matter, a case where she acts similarly because she takes herself to have *judged* that higher tuition doesn't matter, and yet another case where she takes herself to be *committed* to it being true that higher tuition doesn't matter. We can opt to say in each case that the subject *believes* that higher tuition doesn't matter. But the question is whether these cases really do have a propositional attitude in common—in particular, whether they figure in the same law-like psychological generalizations. I shall argue that there is good reason to think that this is not the case. For the actual psychological processes and motivations involved are different. If our concern is what really happens in people's minds, rather than everyday predictive and explanatory convenience, then we should decline to recognize the existence of System 2 beliefs.

I suggest that the only law-like personal-level generalizations in psychology (aside from a few cases dealing with particular perceptual illusions and the like) are those that *quantify over* people's attitudes, or over classes of types of attitude (concerning the value of future rewards, for example, as in the finding of temporal discounting; or concerning the size of some numerical quantity, as in the discovery of "anchoring and adjustment"). The central example is the practical reasoning schema: if someone wants something, and believes that there is an effective way to get it, then the person will typically act accordingly. From the point of view of this schema, the particular beliefs and desires in question don't matter.

Notice that an elaboration of the practical reasoning schema should expand it to include

strengths of belief and desire. An action that could be motivated by a given belief–desire pair is more likely to happen the more firmly the belief is held and the stronger the desire. Increasing the strength of either or both should make the outcome more likely, and should make the agent more persistent in pursing that outcome. (Frankish, 2004, claims that System 2 beliefs are all-or-nothing: one has either committed oneself to the truth of a proposition or one hasn't. But commitments can be more or less strong, of course, and their effects on subsequent behavior can be similarly graded.) We can then go on ask what sorts of interventions are apt to cause changes in the strength of belief. If there really were such things as System 2 beliefs then that should mean that they are covered by the elaborated version of the practical reasoning schema, together with any law-like generalizations concerning the factors that are apt to increase or decrease the strengths of people's attitudes.

One such psychological generalization involving belief is that provision of new evidence in support of the belief, or at least considerations that make the truth of the belief seem more likely, are apt to increase the strength of one's conviction. But it is opaque how this would be supposed to happen in connection with System 2 beliefs. For given the way in which System 2 beliefs are realized, there are only two ways in which they can be directly strengthened. One is to increase the strength of the relevant meta-belief (that one judges that *P* or that one has committed oneself to the truth of *P*). Or one can increase the strength of the relevant motivation (to reason and act rationally or to execute one's commitments). But additional evidence that *P* will generally have no bearing on the strength of one's belief that one believes that *P*, nor on the strength of one's belief that one has committed oneself to the truth of *P* (and nor, of course, on one's desires to be rational or to carry out one's commitments).

Only if the subject happens to have some additional beliefs will the provision of new evidence make a difference. For example, if the subject also believes that additional evidence should increase the strength with which a belief is held, and wants to proceed as a *P*-believer should, then she will thereafter act as if she believed *P* more strongly. Likewise, in commitment cases, only if the subject believes that additional evidence that *P* should increase the strength of one's commitment to the truth of *P*, and wants to proceed as she should, will she act thereafter as if *P* were believed more strongly. Not only is there no guarantee that such additional beliefs will always be present, but it is unclear whether they are even *likely* to be possessed in normal cases.

Notice, in addition, that even if such beliefs are present and operative, their activity will still fail to deliver one of the normal effects of increased evidence for the truth of *P*, namely an increase in the confidence one feels at the thought of *P*. Of course, if the person wants to fulfill the commitments of being a *P*-believer and believes that anyone committed to the truth of *P* would feel greater confidence when provided with additional evidence that *P*, then she may *say* that she feels more confident and will attempt to act appropriately. But there is no plausible causal route here to actually *having* a greater feeling of confidence. So one important strand in the functional profile of belief will almost certainly be absent.

Moreover, there are simple interventions that will increase the strength of one's System 2 "beliefs" that do *not* form part of the normal causal profile of belief and influences on belief. This is because increasing the strength of one's *desires* should have no impact on the strength of one's beliefs (except in cases of so-called "motivated believing", where one believes that it would serve one's purposes to hold the belief in question). Yet increasing the strength of someone's desire to think and act rationally, or her desire to execute her commitments, will directly issue in an apparent strengthening of the alleged belief in question. This is the wrong sort of functional profile for belief.

In addition, the two forms System 2 "belief" will exhibit *distinct* functional profiles. Priming for the value of rationality, for example, will enhance the behavior characteristic of someone who believes that higher tuition doesn't matter in the one case but not the other, with the reverse pattern occurring if we prime for thoughts of duty, or obligation, or commitment. This provides us with a reason *not* to treat these two mental states as being of the same type. And of course neither of these types of motivational prime will have any influence on someone who has a regular System 1 belief that higher tuition doesn't matter.

Notice that although priming would be one way of manipulating someone's desire to be rational, or her desire to execute her commitments, this plays no essential role in the argument. The point is just that changes in desires for things other than the truth of *P* should have no impact on the strength of one's belief that *P*; but if there were such attitudes as System 2 beliefs, then changes in one's desires would have just such effects. Moreover, these are desires that the subjects in question may well avow or attribute to themselves, even though they are ignorant of the specific role they play in sustaining belief-like System 2 activity. So these are still personal-

level mental states, and the psychological generalizations in which they figure operate at the personal level too.

I conclude that there are a number of important respects in which alleged System 2 beliefs fail to match the functional profile expected of beliefs. From a realist perspective, then, we should refuse to countenance them as forms of belief. While it may be useful to talk this way for some everyday purposes, we should recognize that such talk is strictly false. As a result, we have failed to identify a class of beliefs about which we have constitutive, authoritative, self-knowledge.

4. Conclusion

I conclude that the interpretive sensory-access (ISA) theory of self-knowledge is not threatened by either of the sorts of view discussed here, both of which regard beliefs-about-one's-beliefs as partly constitutive of believing. Dispositionalism is implausible as an account of belief and does not, in any case, provide a viable competitor for the ISA theory. And although it may be pragmatically useful to recognize System 2 beliefs, there are good reasons to deny their real existence.

Acknowledgements

I am grateful to Nikolaj Nottelmann and Eric Schwitzgebel for their comments on an earlier draft of this chapter.

References

Baars, B. (1988). *A Cognitive Theory of Consciousness*. New York: Cambridge University Press.

Baddeley, A. (2006). *Working Memory, Thought, and Action*. New York: Oxford University Press.

Block, N. (1986). An advertisement for a semantics for psychology. In P. French, T. Euhling, and H. Wettstein (eds.), *Midwest Studies in Philosophy: 10: Studies in the Philosophy of Mind*, Minneapolis, MN.: University of Minnesota Press.

Bratman, M. (1987). *Intentions, Plans, and Practical Reason*. Cambridge, MA.: Harvard

University Press.

Bratman, M. (1999). *Faces of Intention*. New York: Cambridge University Press.

Carruthers, P. (2011). *The Opacity of Mind*. New York: Oxford University Press.

Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, 38.

Cohen, L. (1992). *An Essay on Belief and Acceptance*. New York: Oxford University Press.

Damasio, A. (1994). *Descartes' Error*. London: Papermac.

Dennett, D. (1978). *Brainstorms*. Brighton: Harvester Press.

Evans, J. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255-278.

Evans, J. (2010). *Thinking Twice*. New York: Oxford University Press.

Evans, J. and Over, D. (1996). *Rationality and Reasoning*. Hove, Sussex: Psychology Press.

Frankish, K. (2004). *Mind and Supermind*. New York: Cambridge University Press.

Frankish, K. (2009). Systems and levels. In J. Evans and K. Frankish (eds.), *In Two Minds*, New York: Oxford University Press.

Frankish, K. (2012). Dual systems and dual attitudes. *Mind & Society*, 11, 41-51.

Jonides, J., Lewis, R., Nee, D., Lustig, C., Berman, M., and Moore, K. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, 59, 193-224.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus, and Giroux.

Millikan, R. (1984). *Language, Thought, and Other Biological Categories*. Cambridge, MA.: MIT Press.

Nichols, S. and Stich, S. (2003). *Mindreading*. New York: Oxford University Press.

Ramsey, W., Stich, S., and Garon, J. (1990). Connectionism, eliminativism, and the future of folk psychology. In J. Tomberlin (ed.), *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, Ridgeview, CA.: Ridgeview Publishing.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, 36, 249-275.

Schwitzgebel, E. (2011). Knowing your own beliefs. *Canadian Journal of Philosophy*, 35, 41-62.

Shoemaker, S. (1994). Self-knowledge and "Inner Sense". *Philosophy and Phenomenological Research*, 54, 249-314.

Stanovich, K. (1999). *Who is Rational?* Mahwah, NJ.: Erlbaum Press.

Stanovich, K. (2009). *What Intelligence Tests Miss*. New Haven, CT.: Yale University Press.

Varley, R. (1998). Aphasic language, aphasic thought. In P. Carruthers and J. Boucher (eds.), *Language and Thought*, New York: Cambridge University Press.

Varley, R. (2002). Science without grammar: scientific reasoning in severe agrammatic aphasia. In P. Carruthers, S. Laurence, and S. Stich (eds.), *The Cognitive Basis of Science*, New York: Oxford University Press.