Advanced Review

Introspection

Mark Engelbert and Peter Carruthers*

Two main questions about introspection are addressed: whether it exists, and whether it is a reliable source of self-knowledge. Most philosophers have assumed that the answers to both questions are positive, whereas an increasing number of cognitive scientists take the view that introspection is either nonexistent (with self-attributions of mental states being made on the same sort of interpretative basis as attributions of mental states to other people) or unreliable. A number of different models of self-knowledge are discussed, and the evidence bearing on the existence and reliability of introspection is reviewed. New experiments are required to tease apart some of the alternatives. © 2010 John Wiley & Sons, Ltd. *WIREs Cogn Sci* 2010 1 245–253

Philosophers have traditionally assumed that the human mind is largely transparent to itself.[1–3] The thesis of mental transparency can be divided into a conjunction of two claims: (1) mental states are *self-presenting* (if one is undergoing a given mental state, then one knows that one is) and (2) our knowledge of mental states is *infallible* (if one believes that one is undergoing a given mental state, then so one is). The scope of this thesis has generally been restricted to current mental events of judging, deciding, reasoning, experiencing, imagining, and feeling, however. This is because it is familiar to common sense that stored knowledge states (e.g., memories) can exist without being presently accessible, and that one can be mistaken about one's long-term motives and qualities of character.

Under pressure from scientific psychology (beginning with the work of Freud but intensified through the rise of cognitive science), most philosophers have backed off from these strong transparency claims. Most will now accept that mental events of all types can occur in ways that are inaccessible to us, and most will accept that we can make mistakes about even the contents of our current conscious mental lives. However, they continue to believe that our access to some of our own mental events is quite different in *kind* from our access to any of the mental events of other people. For, we do not have to engage in *interpretation* when ascribing conscious mental events to ourselves, in the way that we do have to interpret others in light of their circumstances and behavior. This form of noninterpretative access is what is normally referred to as 'introspection.' (Some reserve the term 'introspection' for those cases where one consciously and deliberately pays attention to one's mental states.[4] We will not follow this usage.)

A distinction should be drawn between access to our own mental states that is *inferential* and access that is *interpretive*. Many cognitive processes—including perceptual processes like vision and language parsing—are thought to be inferential in that they employ a computational process together with simplifying assumptions to arrive at a 'best hypothesis' to match the incoming data. Accounts of self-awareness (including those that liken self-awareness to a kind of perception[5]) may thus allow that our access to our own states is inferential in this sense, but such access would still be fully introspective provided that the process employed is genuinely different than that used to attribute mental states to others. *Interpretive* self-attributions, on the other hand, employ a process that is more than just computationally complex, but is similar in significant respects to the sort of interpretive process we use to attribute mental states to others (e.g., relying on information concerning the agent's behavior).

Two questions arise for cognitive science to address: (1) Is there any such thing as introspection? That is, is there any faculty of human psychology that allows for direct, noninterpretive access to one's own experiences, judgments, and other mental events? (2) Whatever psychological mechanisms are used to generate beliefs about one's mental events, how *reliable* are those mechanisms in generating true beliefs about oneself? It is on these two questions that we will focus in this article.

*Correspondence to: pcarruth@umd.edu

Department of Philosophy, Skinner Building, University of Maryland, College Park, MD 20742, USA.

DOI: 10.1002/wcs.4

The logical space for different accounts of self-knowledge is broad. (We use the term 'knowledge,' here, quite loosely, to encompass false beliefs as well as true ones.) Theoretical models may preserve or eliminate noninterpretative access, and, independently of that question, they may place the accuracy of self-attributions at any of a variety of points along a spectrum from practically infallible to hopelessly inaccurate. Thus, in addition to philosophers who assume that introspection is both direct and reliable, there are researchers who hold, e.g., that the only mechanisms of self-attribution are interpretive,[6] but that these mechanisms are quite reliable under most circumstances.[7] One could also remain neutral about the nature of the processes by which we attribute mental states to ourselves while holding that those processes are highly unreliable.[8]

Contemporary discussion of introspection among psychologists and empirically minded philosophers thus takes two forms: there are those who attempt to give psychologically plausible accounts of self-knowledge that vindicate—to a greater or lesser degree—aspects of the traditional philosophical model,[9,10] and there are those whose accounts preserve little or none of what traditional philosophers assume about introspection, arguing variously that there is no introspection[6,7] and that mechanisms of self-knowledge are systematically unreliable.[8,11–13]

COGNITIVE ACCOUNTS OF INTROSPECTION: REDEPLOYMENT AND RECOGNITION

Several philosophers have attempted to give psychologically informed accounts of introspection that preserve its status as a special method for gaining knowledge about ourselves. For example, Peacocke[14] and Rey[15] both suggest that introspection may work in a way that is analogous to the 'efference copies' posited in motor planning and motor control.[16–18] When the nervous system executes a motor plan, copies of the motor intention are transformed into a 'forward model' of the expected sensory consequences of the action for purposes of monitoring and swift online correction. Likewise, the systems responsible for producing propositional attitudes might send copies to an introspection system so as to make them available to the agent. Unfortunately, details of how such a model might work are not provided by either Peacocke or Rey, nor does there appear to be any positive evidence for this view. Bayne and Pacherie[19] have offered an account of how efference copies might provide introspective access to one's *motor* intentions, but they acknowledge that their model would be inadequate as an account of introspection for propositional content even for intentions, let alone for the full range of propositional attitudes.

Since introspection is taken to yield second-order metarepresentations of existing first-order mental events, one question a theory of introspection must answer is, 'How do first-order representations come to be represented in second-order representations?' Most theories of introspection postulate some form of redeployment—i.e., the contents of self-attributions are copies of first-order representations. Redeployment accounts of introspection, then, are accounts of how first-order representations are copied, and the copies deployed within second-order representations. Thus if a subject sees a cat on a mat and judges [the cat is on the mat], then this same representation can be copied and redeployed in a second-order thought, 'I believe that [the cat is on the mat].'

One version of this idea is proposed by Nichols and Stich,[10] who suggest that a monitoring mechanism copies representations from the 'belief box,' prefixes the attitudinal ascription 'I believe that,' and places the resulting second-order representation back into the belief box. This account seems plausible enough for beliefs, but it is unclear how it should be generalized to cover other propositional attitude types. Nichols and Stich must either postulate distinct monitoring mechanisms for each attitude type (i.e., a different monitoring mechanism would copy a representation from the 'desire box,' prefix it with 'I desire that,' before placing the resulting second-order belief into the belief box), or they must posit a single monitoring mechanism that retrieves token attitudes from the multitude of attitude 'boxes'—all of which, presumably, will be differently neurally realized. Either way, the Nichols and Stich account predicts that each of the monitoring mechanisms, or each of the channels by which a singular monitoring mechanism is connected to the various attitude boxes, might become damaged independently of damage occurring to the others. Hence we should expect to find people who can self-attribute beliefs but not desires, or who can self-attribute visual experiences but not auditory ones, and so forth. There is little evidence of such multiple dissociability.

Goldman[9] offers a similar but distinct redeployment account, one in which the contents of self-attributions are again given by redeploying first-order representations, but the *attitude type* is given by recognition of the neural properties that instantiate the representation. Drawing on Craig's[20] account of how neural properties (i.e., the activations of particular classes of neurons) allow for recognition of perceptual states like pains and itches, Goldman suggests that there may similarly be neural properties that act as signatures of various propositional attitudes. The introspective faculty thus *redeploys* the content of self-attributions, and *recognizes* the attitude type via its neural properties. Given this account, Goldman concludes that 'introspection is a perception-like process' (p. 253). While Goldman plainly thinks of introspection as a unitary faculty, however, it is doubtful whether he is entitled to such a view. To the extent that the neural realizers of different attitude and percept types are located in different regions of the brain, each attitude or percept type will rely on a distinct channel to deliver information to the introspection faculty. Thus, Goldman, too, should predict multiple dissociations in the capacity to introspect.

SELF-MONITORING, METACOGNITION, AND MEMORY

Each of the foregoing cognitive accounts of introspection makes introspection a fairly direct, noninterpretative, process, while also preserving a significant degree of reliability for that process. But what evolutionary pressures might have shaped the emergence of a set of introspective mechanisms? One natural and very popular suggestion is that they are designed to have a supervisory role with respect to regular, first-order, cognitive processes—trouble-shooting and intervening in those processes in cases of difficulty, initiating new strategies, checking that tasks are proceeding as expected, and so on and so forth.[21] 'Metacognition' is a very broad concept, encompassing a wide range of cognitive procedures. Here we review findings for one prominent variety of metacognition, metamemory, which cast doubt upon the idea that this metacognitive process, in particular, monitors first-order cognition in the manner commonly assumed. If the evidence we review below generalizes to other kinds of metacognitive process, then it appears that while there is indeed a supervisory role for metacognition, it is one that does not require any introspective capacity distinct from a third-person mindreading system. It would seem, moreover, that our metacognitive interventions are not capable of the sort of direct impact on cognitive processing that would be predicted if introspection had, indeed, evolved for the purpose.

There is a large body of experimental data on metamemory.[22,23] But for the most part such processes appear to operate without the capacity to intervene directly in the states and events represented. For example, most metamemory capacities only require an ability to initiate or intervene in *behavior*. Thus a child might select one memorization task rather than another on the grounds that it contains fewer items (thus implicating knowledge *about* memory, but not intervening in the process of memory itself). And likewise someone might mentally rehearse items in 'inner speech' as an aid to memorization, which is an indirect behavioral influence on memory, not a direct intervention. It should also be noted that while the intention to learn has an effect on study patterns, it has no effect on learning and recall once study patterns are controlled.[24] This is not what one would predict if metamemory involved an introspective capacity that had evolved for purposes of executive control, enabling subjects to intervene directly in the processes of memorization or memory retrieval. (Guiding behaviors that tend to issue in memorization or retrieval, in contrast, can equally well be done by a mindreading system.)

Koriat et al.[25] review much of the extensive literature on metamemory and experimentally contrast two competing models. One is that metacognitive monitoring serves the function of controlling and directing the underlying cognitive processes. (Plainly this would be consistent with the evolutionary explanation of introspection sketched above.) The other is that metacognitive judgments are evidence-based, cued by experiences that are caused by the cognitive processes in question. (This would be more consistent with a self-interpretative position.) While they do find metacognitive phenomena that fit the former profile, none of these suggests any real role for introspection of attitudes. Rather, they include such phenomena as allocating greater study time to items that attract a larger reward. In contrast, there is extensive evidence of cue-based metacognitive judgments. Thus feelings of knowing are often based on the ease with which one can access fragments of the target knowledge[26] or items related to the target.[27] And judgments of learning made during or after study are based on the 'fluency' with which items are processed during study itself.[28–30] Again, this is not what one would predict if a capacity for introspection of attitudes had evolved for purposes of metacognitive control. For why, in that case, would one need to rely on indirect cues of learning?

It might be replied that we only need to rely on indirect cues when direct ones fail. If we can directly monitor that we know that Columbus landed in 1492, for example (as seems to be the case), then there is no need of indirect cues. The latter are only required, perhaps, in cases where retrieval seems to be failing, although there is an impression that information is available (e.g., in tip of the tongue phenomena). It should be conceded, of course, that we can *express* our knowledge that Columbus landed in 1492 directly and confidently in speech, without needing to rely on any indirect cues. But it is doubtful whether this capacity is metacognitive in character. For no one thinks that language production, in general, requires metacognitive thought. Rather, executive systems in collaboration with the language faculty conduct a search of memory, encoding the information retrieved into language. Perhaps we only acquire the knowledge that we know from the result.

If what seems to be true of metamemory turns out also to be true of other metacognitive processes, then it is doubtful that introspection evolved for self-monitoring, as is so often supposed. But if introspection did not evolve for self-monitoring, then what *is* the explanation of the human capacity to self-attribute mental states? Carruthers' proposal (Ref 7, discussed in 'Self-Interpretive Models') is that self-attribution is conducted by the same cognitive system that we use to attribute mental states to others. Thus the evolutionary explanation of knowledge of our own minds is the same as the story for knowledge of others' minds. (See Refs. 31 and 32 for accounts of how and why such mindreading abilities evolved in higher primates and humans.)

While evidence for a self-monitoring account of introspection is lacking (at least in the case of metamemory), it is important to stress the implications of these models of self-knowledge for the question of memory of one's own mental states, since this will become an issue later in our discussion. If the function of introspection is to monitor one's first-order mental processes, intervening and trouble-shooting where necessary, then two predictions can be made. The first is that some sort of temporary record of immediately prior mental events will need to be kept, so that the monitoring mechanism can represent each stage as an event in an on-going process. Hence subjects should be capable of reporting their immediately past mental states. And the second prediction is that representations of one's own mental states should *not* be stored in long-term memory, unless for some reason they are rehearsed and/or consciously attended to. For this is not necessary to support the trouble-shooting function, and would serve no useful purpose. Rather, knowledge of our immediately past mental events should fade away rapidly, just as dreams do. And indeed, consistent with this prediction, subjects who use Hurlburt's[33,34] introspection-sampling methodology (which will be described briefly in 'Future Research') make many surprising discoveries about the patterns in their inner experience over time, suggesting that long-term memories of such experience are not routinely created.

EVIDENCE AGAINST INTROSPECTION FOR ATTITUDES

Despite the *prima facie* plausibility of accounts like Goldman's, there is considerable evidence from experimental psychology that self-attributions often proceed in an interpretive manner, and are frequently erroneous as well. Much of this data involves retrospective reports that occur too long after the events for subjects to have any memory of them, however. Since subjects would have no option but to self-interpret their remembered behavior in such circumstances, this evidence does not count directly against the existence of introspection. (It remains interesting, however—and something that a defender of introspection may be challenged to explain—that people nevertheless have the *impression* that they are merely remembering their past mental states in such cases, and are unaware of engaging in self-interpretation.) However, subjects will also confabulate answers when asked—implicitly or explicitly—what decisions or judgments they have *just* made following a given behavior. This is, of course, not what one would expect if introspection provided direct and reliable access to one's own mental states.

Experiments have demonstrated confabulation effects for both judgments and decisions. In the case of judgments, Linder et al.[35] found that when subjects were made to write an essay defending a proposition with which they disagreed, those subjects who were paid poorly were more sympathetic to the proposition after the experiment. (Such effects have been replicated many times.) This suggests that the subjects were, at some level, trying to explain behavior that ran counter to their considered judgments, as proponents of so-called 'self-perception' accounts of dissonance phenomena suppose.[13] Those who were paid well were able to appeal to their financial benefit, while those paid poorly were forced to attribute to themselves a higher degree of belief in the proposition they were defending.

Wells and Petty[36] have also demonstrated self-interpretation for judgments, showing that subjects who nod their heads while listening to a persuasive message find the message more convincing than those who shake their heads while listening. In response to the objection that nodding may have caused more positive thoughts about the message (resulting in a favorable judgment that is then introspected), Briñol and Petty[37] manipulated the persuasiveness of the message. Some subjects heard relatively convincing arguments for a proposition while some subjects heard messages with weak and irrelevant arguments. The results for the 'persuasive' condition were the same as in Wells and Petty,[36] but in the 'unpersuasive' condition the trend was reversed: head-shakers expressed greater belief than head-nodders. The explanation proposed is that subjects interpreted their own head movements as expressing either agreement or disagreement with their own internal commentary on the message to which they were listening, adjusting the extent of their agreement with the message accordingly.

Significant confabulation effects have also been observed for decisions. Brasil-Neto et al.[38] asked subjects to make a decision to lift either their right or left index finger upon hearing the sound of a click. Unbeknown to the subjects, the click was generated by a transcranial magnetic stimulation (TMS) machine, which the experimenters used to stimulate either the right or left hemisphere motor cortex, thus causing the subsequent finger movement. Although the movements were caused by magnetic stimulation, subjects nonetheless reported a *decision* to move the finger in question.

Both Wegner[12] and Gazzaniga[11,39] have also found confabulation for decisions. Wegner describes subjects who are given hypnotic suggestions to perform some action, but when asked, post hypnosis, why they are performing the action, they will confabulate a reason. (For example, one subject was hypnotized to take a book from a table and place it on a shelf, and claimed she was tidying the room.) Gazzaniga has found similar effects in split-brain subjects, who have had their corpus collosum (the brain structure that allows for communication between the two hemispheres) severed as a treatment for epilepsy. Since language is generally lateralized to the left hemisphere, information presented to the left side of the body (and thus available only to the right hemisphere) is off-limits to the language-producing centers of the brain. When Gazzaniga[39] presented the instruction 'Walk!' in the left visual field of a split-brain subject, the subject complied. When asked why he was doing so, however, the subject confidently responded 'I'm going to get a coke from the house.' It appears that since the right hemisphere cannot tell the left hemisphere about the instruction, the left hemisphere is forced to generate its own explanation for the behavior (without being aware that it is doing so).

Defenders of introspection reply to these experiments in a variety of ways, which are meant to show that the observed confabulation effects do not support the strong claim that we *never* have introspective access to our own propositional attitudes. One common response is that confabulation is a failure of reasoning about which mental states are the *causes* of one's own behavior.[10,13,15,39] Defenders of introspection point out that no one claims that we should be able to introspect the causal link between actions and the mental states that actually produced them. Rather, introspection requires only that we are able to detect certain of the mental states we actually possess, independent of their status as causes of our behavior. It is thus argued, e.g., that the subjects in Nisbett and Wilson's[40] pantyhose study—who had a marked tendency to select the rightmost item from a set of identical items on display—did actually believe that the rightmost pair of stockings were softer or silkier, but were unaware that these beliefs were caused, somehow, by their right-hand attentional bias.

Another objection to confabulation data is that interrupting subjects mid-action and asking them 'Why are you doing that?' (as in Gazzaniga's split-brain studies and the hypnotism studies reported by Wegner) creates a pragmatically complex situation, one in which subjects may feel compelled to *justify* their behavior or attempt to make it appear *rational*.[15] This could trigger modes of self-interpretation that are unnecessary and therefore absent in the vast majority of cases of self-attribution. While there are ways of responding to these objections,[7] it is fair to say that the issues will not be finally resolved without further experimentation that controls for the relevant factors.

SELF-INTERPRETIVE MODELS

Despite the objections to the confabulation data mentioned above, many self-knowledge researchers accept the vast majority of the evidence as valid and as demonstrating a significant role for interpretation in self-attribution. And even the staunchest defenders of introspection accept the validity of at least some of the confabulation data, meaning that almost all theories of self-knowledge allow some role for interpretive mechanisms. This section will focus on those theories that posit some *significant* role for interpretation, including theories that posit interpretation as the *only* means of self-knowledge.

Many of the cognitive scientists who have been at the forefront of work demonstrating self-interpretation in experimental settings stop short of saying that *all* self-knowledge is interpretive. Wegner[12] takes his work to show that we use interpretive mechanisms to infer when we are responsible for (i.e., have made a decision to perform) a given action, but he does not claim that our beliefs about mental states other than decisions are similarly inferential in nature. Gazzaniga[11] posits the existence of a 'left-brain interpreter' whose function is to interpret the agent's own behavior in a way that makes for a coherent story. He does not claim, however, that this precludes any kind of introspective access. Finally, Wilson[13] suggests that we often engage in forms of adaptive self-deception that allow us to maintain a positive self-image, but holds only that the *causes* of one's thoughts and behavior are inferred, while the thoughts themselves are often ascertained through introspection.

Several authors have advanced stronger claims, however, positing models of self-attribution that rely entirely on interpretive mechanisms. Gopnik[6] suggests a symmetry between knowledge of others' mental states through observation of their behavior and knowledge of ourselves. That is, we obtain knowledge of our own mental states by observing our own behavior and applying the same sorts of reasoning we use to infer the mental states of others from *their* behavior. However, faced with the obvious criticism that there must be more to the story, in order to account for the fact that we seem to be considerably better at detecting our own mental states than the mental states of others, Gopnik offers only the suggestion that there is some sort of 'Cartesian buzz' (Ref 6, p. 11) that provides an additional source of evidence as to the states we are in. Researchers have been reluctant to take this view seriously without further specification of what this 'buzz' might be.

Carruthers[7] has, however, offered a detailed account that manages to deny introspection for attitudes while at the same time explaining the superiority of self-knowledge as compared to other knowledge. Carruthers emphasizes the distinction between perceptual states and propositional attitude states, arguing that knowledge of the former is grounded in introspection while knowledge of the latter is based upon self-interpretation. There is, on this account, a single 'mindreading' system that is responsible for generating beliefs about mental states—both one's own and those of others. Carruthers' account of introspection for perceptual states follows Baars'[41,42] global workspace account of consciousness, whereby attended perceptual and quasi-perceptual (e.g., imagistic) information is 'globally broadcast' to a wide range of belief-forming and decision-making systems. This information generally includes incoming sensory data from the environment, as well as information about goings-on internal to the agent, such as mental imagery (including auditory imagery in the form of inner speech), as well as somatosensory and proprioceptive data. The mindreading system, Carruthers suggests, has access to all of this information, but *not* to outputs of the other belief-forming and decision-making systems that receive their input from the global workspace. Thus the mindreading system can directly self-attribute the perceptual states it receives as input to produce attributions of the form 'I am seeing red,' 'I am in pain,' 'I am hungry,' and so on. But it cannot, according to this model, directly self-attribute the propositional attitude states that are the outputs of decision-making and belief-formation systems, for these states are not globally broadcast. Instead, the mindreading system must use the perceptual data available to it to *infer* what the agent's occurrent propositional attitudes are (just as it does in the case of third-person mindreading).

Carruthers' model can thus explain what Gopnik's cannot: why we are able to attribute mental states to ourselves even when we are not engaging in any overt behavior, and why we are much better mindreaders of ourselves than of others. Someone sitting motionless may still have an abundance of information about her current situation in the form of sensory, imagistic, and somatosensory data, and hence should have no trouble attributing mental states to herself. Still, one might worry whether Carruthers' interpretive account is capable of accounting for the full range of cases of self-attribution, including seemingly spontaneous self-reports, especially if it is the case that there exists 'unsymbolized thinking,' or propositional thought that is not accompanied by any visual or auditory mental imagery, as Hurlburt and Akhter[43] and Siewart[44] suggest.

ARGUMENTS AGAINST THE RELIABILITY OF INTROSPECTION

We have examined views that deny the reality of introspection, either altogether or with respect to propositional attitudes in particular. Here we examine accounts that allow the existence of introspection while claiming that it is highly unreliable. Just such a combination of views is defended by Schwitzgebel.[3,45] He maintains that we do have immediate (i.e., introspective) and reliable access to our experiences in some simple cases, such as intense pain or confrontation with a vivid shade of red. He does not, however, advance a positive model of the introspective faculty, or address the relative prevalence of introspection versus interpretation, confining himself to arguing against the reliability of whatever is 'the primary method by which we normally reach judgments about our experience' (Ref 3, p. 248). Schwitzgebel thus expresses extreme skepticism about the reliability of most self-attributions of experience, and therefore differs from Carruthers[7] in emphasizing the unreliability of judgments about our own perceptual states (while remaining mostly silent about judgments about our own attitudes).

Schwitzgebel points out how often people are, e.g., mistaken about the clarity and detail of their visual field beyond the fovea. He also notes how difficult it is to know whether or not joy (e.g.) has a single, distinctive, experiential character. And he points out that there is a deep disagreement amongst philosophers over whether or not thinking has a distinctive phenomenology apart from any imagery that might accompany it. Schwitzgebel[8] ultimately concludes that whatever the character of our introspective mechanisms, they are unreliable in two ways: they fail to yield any conclusive judgment about the nature of many experiences (e.g., emotional experience), and when they *do* yield results, those results are frequently inaccurate (as in people's evaluation of the clarity of their visual field).

Let us discuss Schwitzgebel's examples in turn. First consider most people's ignorance of the poverty of experience outside of the very center of their visual field. This seems to us to be more plausibly explained in terms of people's ignorance of their own saccadic eye movements, rather than by any failure of introspection. When we attend to an object of interest—a tree that is bursting with autumn colors, e.g.—our eyes will saccade across it many times each second, with each saccade gathering detailed information from the fovea. There is evidence that information from each saccade is retained in an iconic memory store whose contents are built up over time and made available in consciousness.[46,47] But for the most part we are unaware of our own eye movements. Hence we have the illusion that we are taking in all of the details simultaneously, when in fact the experience is built up sequentially. This is ignorance of our own automatic *movements* rather than a mistake about the character of our experience. And it is *philosophers* who have been misled into thinking that the experience of the visual field is always impoverished. (Granted, it will be so whenever subjects are induced to fixate without saccading.)

Now consider the difficulty of knowing whether or not joy has a consistent phenomenological core. Notice that such knowledge would require us to generalize across a number of introspected experiences. But we suggested in 'Self-Monitoring', Metacognition, and Memory that introspection is unlikely to give rise to many long-term memories distinct from our first-order memories of the objects and circumstances that our experiences are about. Hence we might remember *that* we were joyful on a given occasion, and perhaps what we were doing or seeing or hearing at the time, without having any memory of the experience of joy itself. If there are not many long-term memories of the products of introspection, then, naturally, generalizing about our introspected experiences will be difficult. But this is not a failure of introspection itself.

Finally, consider the question whether thoughts, as such, possess a distinctive phenomenology. It is true that this question is extremely hard to resolve on introspective grounds. But (if we set to one side the dispute over the existence of unsymbolized thinking—see Refs. 7, 43, and 44) that is arguably because answering it would require us to distinguish between *causal* and *constitutive* contributions to phenomenology. And this is not something that can be available to introspection, on any view of the latter. Whenever we entertain a (conscious) thought, this is likely to be accompanied by changes in our phenomenal experience—either visual or auditory imagery, or fleeting emotional feelings, or whatever. But in order to answer the question about the phenomenology of thought *per se*, we would have to be able to tell which of these changes are caused by (or are causes of) the thought in question and which (if any) are constitutive of it. But this is not the sort of thing that an introspective faculty would be capable of detecting. Thus the cases discussed by Schwitzgebel[8] do not provide much reason to think that our access to occurrent perceptual states is unreliable, even if we are poor at generalizing about our experiences or distinguishing causal from constitutive aspects of experience.

FUTURE RESEARCH

A great deal of research effort over the last 50 years has gone into documenting confabulation effects. It is now solidly established that people will often interpret their own behavior without awareness that they are doing so, and will, as a result, frequently make false self-attributions of mental states to themselves. Unfortunately, comparatively little effort has been devoted to investigating whether, in addition to self-interpretation, people also have the capacity to attribute mental states to themselves via a noninterpretative, introspective, route. What is now needed is a concerted research effort to establish whether or not introspection exists.

No doubt many different forms of experiment can be envisaged. But any that rely upon subjects' reports of their mental states need to be careful to devise appropriately brief timescales for the reporting. This is because it is likely that introspection, if it exists, will only issue in a short-term memory of the introspected events that lasts for a few seconds before being lost. Experimenters will also need to be sensitive to the dangers inherent in probing subjects with 'Why?' questions. For these may have the effect of diverting subjects' attention while also placing a pragmatic onus on them to provide rationalizations of their behavior.

We think that one promising approach for investigators to pursue would be to adapt Hurlburt's[33,34] introspection-sampling methodology. Subjects wear a modified paging device throughout the day, which delivers a 'beep' through an earpiece at random intervals. Subjects are instructed to 'freeze' the contents of their consciousness at the moment they hear the beep, and to straight away make brief written notes of those contents. This means that subjects are probed for their reports within the brief introspective memory-window, and in a way that largely bypasses the dangers inherent in 'Why?' questions.

Because Hurlburt is interested in characterizing and establishing generalizations about the details of subjects' inner lives (rather than in establishing the reality of introspection), the written notes are elaborated in a follow-up interview with an experimenter within the following 24 hours. We are much more doubtful about this aspect of the procedure, given the known constructive nature of memory, and given that any actual memory of the inner experience in question is likely to have been lost in the interval. But these retrospective elaborations could be abandoned for our purposes. And for similar reasons, the beeper need not be timed to go off at random intervals during the day, but rather at specific (but unexpected) points during an experimental procedure. The beeper methodology could be used to probe subjects' awareness at crucial times during any of the usual confabulation experiments, e.g., or while previously hypnotized subjects perform actions that they had been instructed to carry out while hypnotized.

CONCLUSION

Recent attempts to provide psychologically plausible accounts of an introspective faculty, which would grant subjects direct access to their own mental states, have been unsatisfactory. But the debate is far from over. Although experimental psychology has established that we *sometimes* engage in self-interpretation, sometimes does not entail always. Thus much contemporary theorizing about self-knowledge is aimed at characterizing the *extent* to which we rely on interpretive mechanisms, and whether that leaves any room for something resembling traditional introspective access. Advances in the area will require the development of new techniques for testing the mechanisms of self-attribution, ones that control for the various confounds inherent in earlier experiments. A concerted research effort in experimental psychology is necessary to give philosophers and cognitive scientists an evidentiary basis to adjudicate among the many possible mechanisms by which self-attributions might be made.

REFERENCES

1. Descartes R. *Meditations on First Philosophy*. Indianapolis: Hackett Publishing Company; 1641/1993.
2. Kant I. *The Critique of Pure Reason*. New York, NY: Cambridge University Press; 1781/1998.
3. Metcalfe J, Shimamura A, eds. *Metacognition: Knowing About Knowing*. Cambridge, MA: MIT Press; 1994.
4. Schwartz B, Smith S. The retrieval of related information influences tip-of-the-tongue states. *J Mem Lang* 1997, 36:68–86.
5. Nelson T, ed. *Metacognition: Core Readings*. Boston, MA: Allyn and Bacon; 1992.
6. Gopnik A. The illusion of first-person knowledge of intentionality. *Behav Brain Sci* 1993, 16:1–14.
7. Carruthers P. How we know our own minds: the relationship between mindreading and metacognition. *Behav Brain Sci* 2009, 32:2.
8. Wells G, Petty R. The effects of overt head movements on persuasion: compatibility and incompatibility of responses. *Basic Appl Soc Psychol* 1980, 1:219–230.
9. Goldman A. *Simulating Minds: the Philosophy, Psychology, and Neuroscience of Mindreading*. New York, NY: Oxford University Press; 2006.
10. Peacocke C. *Truly Understood*. New York, NY: Oxford University Press; 2008.
11. Gazzaniga M. Consciousness and the cerebral hemispheres. In: Gazzaniga M, ed. *The Cognitive Neurosciences*. Cambridge, MA: MIT Press; 1995.
12. Wolpert D, Kawato M. Multiple paired forward and inverse models for motor control. *Neur Net* 1998, 11:1317–1329.
13. Wilson T. *Strangers to Ourselves*. Cambridge, MA: Harvard University Press; 2002.
14. Rosenthal D. *Consciousness and Mind*. New York, NY: Oxford University Press; 2005.

15. Shallice T. *From Neuropsychology to Mental Structure*. New York, NY: Cambridge University Press; 1988.

16. Grush R. The emulation theory of representation: motor control, imagery, and perception. *Behav Brain Sci* 2004, 27:377–442.

17. Wolpert D, Ghahramani Z. Computational principles of movement neuroscience. *Nature Neurosci*. 2000, 3:1212–1217.

18. Wolpert D, Kawato M. Multiple paired forward and inverse models for motor control. *Neur. Net*. 1998, 11:1317–1329.

19. Bayne T, Pacherie E. Narrators and comparators: the architecture of agentive self-awareness. *Synthese* 2007, 159:475–491.

20. Craig A. How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Rev Neurosci* 2002, 3:655–666.

21. Schwitzgebel E. The unreliability of naïve introspection. *Philos Rev* 2008, 117(2):245–273.

22. Nichols S, Stich S. *Mindreading: An Integrated Account of Pretence, Self-awareness, and Understanding Other Minds*. New York, NY: Oxford University Press; 2003.

23. Nisbett R, Wilson T. Telling more than we can know. *Psych Rev* 1997, 84:231–295.

24. Anderson J. *Learning and Memory: An Integrated Approach*. New York: John Wiley & Sons; 1995.

25. Koriat A, Ma'ayan H, Nussinson R. The intricate relationships between monitoring and control in metacognition. *J Exp Psychol Gen* 2006, 135:36–69.

26. Koriat A. How do we know that we know? The accessibility model of the feeling of knowing. *Psych Rev* 1993, 100:609–639.

27. Wegner D. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press; 2002.

28. Begg I, Duft S, Lalonde P, Melnick R, Sanvito J. Memory predictions are based on ease of processing. *J Mem Lang* 1989, 28:610–632.

29. Benjamin A, Bjork R. Retrieval fluency as a metacognitive index. In: Reder L, ed. *Implicit Memory and Metacognition*. Hillsdale, NJ: Erlbaum; 1996.

30. Koriat A. Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. *J Exp Psychol Gen* 1997, 126:349–370.

31. Byrne R, Whiten A, eds. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. New York, NY: Oxford University Press; 1988.

32. Dunbar R. On the origin of the human mind. In: Carruthers P, Chamberlain A, eds. *Evolution and the Human Mind*. New York, NY: Cambridge University Press; 2000.

33. Hurlburt R. *Sampling Normal and Schizophrenic Inner Experience*. New York, NY: Plenum Press; 1990.

34. Hurlburt R. *Sampling Inner Experience with Disturbed Affect*. New York, NY: Plenum Press; 1990.

35. Locke J. *An Essay Concerning Human Understanding*. Amherst, New York: Prometheus Books; 1690/1995.

36. Wells G, Petty R. The effects of overt head movements on persuasion: compatibility and incompatibility of responses. *Basic Appl Soc Psychol* 1980, 1:219–230.

37. Briñol P, Petty R. Overt head movements and persuasion: a self-validation analysis. *J Neurol Neurosurg Psychiatr* 2003, 84:1123–1139.

38. Brasil-Neto J, Pascual-Leone A, Valls-Solé J, Cohen L, Hallett M. Focal transcranial magnetic stimulation and response bias in a forced choice task. *J Neurol Neurosurg Psychiatr* 1992, 55:964–966.

39. Gazzaniga M. *The Mind's Past*. Berkeley, CA: California University Press; 1998.

40. Rey G. (Even higher-order) intentionality without consciousness. *Rev Int Philos* 2008, 52:51–78.

41. Baars B. *A Cognitive Theory of Consciousness*. New York, NY: Cambridge University Press; 1988.

42. Baars B. *In the Theatre of Consciousness*. New York, NY: Oxford University Press; 1997.

43. Hurlburt R, Akhter S. Unsymbolized thinking. Conscious. *Cognit* 2008, 17:1364–1374.

44. Siewert C. *The Significance of Consciousness*. Princeton, NJ: Princeton University Press; 1998.

45. Hurlburt R, Schwitzgebel E. *Describing Inner Experience? Proponent Meets Skeptic*. Cambridge, MA: MIT Press; 2007.

46. Linder D, Cooper J, Jones E. Decision freedom as a determinant of the role of incentive magnitude in attitude change. *J Personality Social Psych* 1967, 6:245–254.

47. Wolpert D, Ghahramani Z. Computational principles of movement neuroscience. *Nature Neurosci* 2000, 3:1212–1217.