

Introspection: Divided and Partly Eliminated

Peter Carruthers

This paper will argue that there is no such thing as introspective access to judgments and decisions. It won't challenge the existence of introspective access to perceptual and imagistic states, nor to emotional feelings and bodily sensations. On the contrary, the model presented in Section 2 presumes such access. Hence introspection is here divided into two categories: introspection of propositional attitude events, on the one hand, and introspection of broadly perceptual events, on the other. I shall assume that the latter exists while arguing that the former doesn't (or not in the case of judgments and decisions, at least). Section 1 makes some preliminary points and distinctions, and outlines the scope of the argument. Section 2 presents and motivates the general model of introspection that predicts a divided result. Section 3 provides independent evidence for the conclusion that judgments and decisions aren't introspectable. Section 4 then replies to a number of objections to the argument, the most important of which is made from the perspective of so-called "dual systems theories" of belief formation and decision making. The upshot is a limited form of eliminativism about introspection, in respect of at least two core categories of propositional attitude.

1 Preliminaries

Before embarking on substantive discussion, some terminological and other elucidatory remarks are in order. I shall understand "introspection" quite broadly, to encompass a variety of potential processes postulated by different types of account. There are just two key ideas. One is that introspection is a higher-order process, issuing in awareness or knowledge of (or at least beliefs about) the occurrence of token mental states. (On some accounts introspection needn't always be reliable, any more than external perception is.) When I introspect a feeling of anger, for example, I become aware of that feeling, and come to know (or at least believe) that I am angry. The other key idea is that introspection is *not* an *interpretative* process. We think that introspective access to our own mental states is epistemically quite different—in kind, and not just in degree—from the access that we have to the thoughts and perceptions of other people (Wright et al., 2000;

Gertler, 2003). The latter occurs via interpretation of people's behavior and circumstances, whether through deployment of theoretical knowledge, or via simulation, or (more plausibly) both (Nichols and Stich, 2003; Goldman, 2006). In contrast, we think that we don't need to notice and interpret our own behavior and circumstances in order to know of our own mental states when we introspect them.

To say that introspection isn't an interpretative process doesn't necessarily mean that it isn't *inferential*, however. Some accounts of introspection maintain that it happens via the operations of *inner sense*, where the latter is modeled on the various outer senses like vision and hearing (Lycan, 1987, 1996). And just as the processes that give rise to a percept of a horse or a tree are partly computational and inferential in character, then so, too, might be the processes that issue in *introspection* of a percept of a horse, or in *introspection* of the judgment that trees absorb carbon. What is crucial is just that these inferences should *not* be ones that appeal to facts about the subject's own behavior and circumstances as premises. For if they did, then there would no longer be any significant, principled, contrast between self-knowledge and other-knowledge.

Notice that the term "introspection" is here deployed quite broadly, to encompass views that are often contrasted by their proponents with introspectionist accounts of self-knowledge (where the latter are understood narrowly, in terms of some or other variety of inner sense). Since the key idea for our purposes is just that introspection issues in higher-order beliefs in ways that don't depend upon self-interpretation, then even neo-Wittgensteinian accounts of self-knowledge that claim a constitutive relationship between verbal expressions of propositional attitudes (so-called "avowals") and the attitudes thereby expressed (e.g. Wright, 2000) will count as introspectionist. A broad swathe of different views will therefore have been ruled out, if it can be shown that our access to our own judgments and decisions is always interpretative.

The correlate of introspection, of course, is consciousness. Everyone will allow that if a mental state is introspected, then it is conscious. But not everyone agrees that introspection is also a necessary condition of conscious status. First-order theorists of consciousness like Tye (1995, 2000), for example, while allowing that humans are capable of introspecting their conscious states, and hence of achieving higher-order awareness of them, will insist that creatures can be subject to conscious states without being capable of introspection. These issues are orthogonal to those that are addressed in the present paper, however. Our focus is on introspection, not consciousness, even if some higher-order accounts of the latter will maintain

that the absence of introspection must entail a corresponding absence of consciousness (Lycan, 1996; Carruthers, 2000; Rosenthal, 2005).^{1, 2}

My goal in this paper is to argue that neither judgments nor decisions are introspectable, but are known only via a process of self-interpretation. I take judgments to be events of belief-formation, and I take decisions to be acts of willing, or the events that create novel activated intentions. Judgments are a kind of active, occurrent, mental event, which when stored give rise to dormant, standing-state, beliefs; and likewise decisions are the mental events that give rise to both standing-state intentions and actions. I have argued elsewhere that standing-state attitudes are only introspectable derivatively (if at all) via introspection of their activated counterparts (Carruthers, 2005). In which case, if activated attitudes aren't introspectable, then neither are beliefs and intentions *tout court*. But I shan't rely on this here. If someone wants to claim that standing-state beliefs and intentions can be introspected even if judgments and decisions can't be, then I shan't gainsay them.

Judgments and decisions aren't the only forms of active, occurrent, propositional attitude, of course. This paper won't say anything about the introspectability of active desires, for example, although I am actually inclined to think that the same sort of negative case can be built.³ Indeed, judgments and decisions aren't even the only forms of activated belief and intention respectively. They are the events through which beliefs and intentions are first formed. But of course long-standing beliefs and intentions can become active thereafter. If someone asks me what I believe to be the date of the battle of Hastings I shall reply, "Ten sixty-six", thereby activating, and expressing or reporting, a belief that I first formed as a teenager. And then the question arises whether activated beliefs of this sort are introspectable.

The arguments presented in Section 3 pertain only to judgments and decisions, and don't directly address the introspectability of activated long-standing beliefs and intentions. (The model put forward in Section 2, and partially confirmed in Section 3, *predicts* that such states

¹ Notice that it follows from these higher-order accounts of consciousness that if we can't introspect our own judgments and decisions (as shall I argue herein that we can't), then there can be no such thing as *conscious* judging, or *conscious* deciding, either. The result would be a limited form of propositional-attitude eliminativism.

² Rosenthal himself uses the term "introspection" in a much more restricted way than I do here, limited to cases where one has *conscious* thoughts about one of one's own mental states.

³ See Damasio (1994), for example, who argues that what we are aware of in introspection are the somatic effects of activated desires and emotions, not those states themselves.

shouldn't be introspectable, however.) And it might seem that an interpretational account of self-knowledge of such states would be singularly implausible. For when I reply when asked what I believe about the date of the battle of Hastings, or about my mother's maiden name, what could possibly be the inputs to the self-interpretation process? I am nevertheless able to answer such questions smoothly and unhesitatingly. This point is by no means decisive, however. For there isn't any reason to think that the verbal expression of a standing-state belief requires that I should first form the higher-order belief that I have that belief. Rather, the search process that activates the standing-state belief in question can make the result available for formulation into speech directly. So answering unhesitatingly when asked what I believe needn't mean that I am capable of introspecting an activated version of that belief. Rather, I might only learn of that occurrent belief by interpreting the utterance (or its counterpart in inner speech) through which I express it. I shall not, however, attempt to defend this here. Our present focus is more narrowly on the introspectability of judgments and decisions.

2 A model of introspection

The theory of introspection that I propose to defend, together with the manner in which introspection fits into the overall architecture of the human mind, is depicted in Figure 1. On this account, there are a range of perceptual systems (visual, auditory, somatosensory, etc.) which broadcast their outputs to a set of conceptual systems. Some of these generate judgments, some create new goals, and some generate decisions and intentions for action. Each of these conceptual systems can store its outputs in memory, and can access and activate those stored representations when reasoning. Included amongst the systems for generating judgments and beliefs is a mindreading faculty, which produces higher-order judgments about the mental states of others and of oneself.

Insert Figure 1 about here

There is now extensive evidence from a variety of sources that the human mind exemplifies a perception / belief / desire / decision-making architecture (Carruthers, 2006).⁴ And

⁴ Admittedly, this sort of model is rejected by philosophers who endorse "enactive" accounts of the architecture of mind, such as Hurley (1998) and Noë (2004). These authors assume without real argument, however, that action is

there is robust evidence of the “global broadcasting” of (conscious) perceptual outputs to a wide range of concept-using consumer systems (Baars, 1988, 1997, 2002, 2003; Dehaene and Naccache, 2001; Dehaene et al., 2001, 2003; Baars et al., 2003; Kreiman et al., 2003). There is also good evidence that imagery (including the auditory imagery that gets deployed in so-called “inner speech”) re-uses the resources of the perceptual systems, utilizing back-projecting neural pathways to generate patterns of stimulation similar to those that would occur when undergoing a perception of the appropriate sort. These are then processed by the perceptual system in question and globally broadcast in the usual way (Paulescu et al., 1993; Kosslyn, 1994; Shergill et al., 2002; Kosslyn et al., 2006).

There is also robust evidence of a distinct, or partially distinct, mindreading system (Frith and Frith, 2003). This accesses the outputs of perceptual systems and attributes mental states in the light of that information. On some accounts the mindreading system is a module or set of modules, and is to a significant degree innate (Baron-Cohen, 1995; Scholl and Leslie, 1999). On other accounts it is an organized body of knowledge, built up during infancy by processes of learning and theorizing (Wellman, 1990; Gopnik and Melzoff, 1997). For present purposes we don’t need to take a stand on these issues. Most theorists are now agreed, however, that the mindreading faculty needs to operate in close conjunction with other systems, and that the attribution of mental-states to other people also involves processes of *simulation* of various sorts (Nichols and Stich, 2003; Goldman, 2006).

Notice that by virtue of receiving globally broadcast perceptual states as input, the mindreading system will find it trivially easy to self-attribute those percepts. Receiving as input a visual representation of a dog chasing a ball, for example, it will be trivial for it to form the judgment, “I am seeing a dog chasing a ball”. (At least, this will be easy provided that the visual state in question has been partially conceptualized by other mental faculties, coming to the mindreading system with the concepts *dog*, *chasing*, and *ball* already attached. I shall return to discuss the significance of this point in a moment.) This is the way in which introspection of

constitutive of perception and cognition, rather than merely contributing causally to it. See Block (2005) for an extended critique of Noë along these lines. And see Carruthers (2006) for an account that sees action as making important contributions to human cognition—indeed, as being fully determinative of certain forms of cognitive process—while preserving the main elements of the perception / cognition / decision / behavior model assumed in the present article.

perceptual, somatosensory, and imagistic mental events is achieved, I suggest. Given that the mindreading faculty possesses the concepts *sight*, *hearing*, and so forth (together with a concept of self), it should be able to activate and deploy those concepts in the presence of the appropriate sort of perceptual input on a recognitional or quasi-recognitional basis (Carruthers, 2000). Since no appeals to the subject's own behavior or circumstances need to be made in the course of making these judgments, the upshot will qualify as a form of introspection.

As the example of seeing a dog chasing a ball makes clear, the thesis that judgments aren't introspectable requires important qualification. In particular, it should be restricted to judgments that aren't perceptual judgments. According to Kosslyn (1994) and others, the initial outputs of the visual system interact with a variety of conceptual systems that deploy and manipulate perceptual templates, attempting to achieve a "best match" with the incoming data. When this is accomplished, the result is globally broadcast as part of the perceptual state itself. Hence we see an object *as* a dog or *as* chasing something. Since this event can give rise immediately to a stored belief, it qualifies as a (perceptual) judgment. But since it will also be received as input by the mindreading system (by virtue of being globally broadcast), it will also be introspectable. In the discussion that follows, therefore, whenever I speak of "judgments" I should be understood to mean "*non-perceptual* judgments".

There is good reason to endorse the sort of mental architecture depicted in Figure 1, then, together with its various components. And the upshot is that we have introspective access to our own perceptual and quasi-perceptual states. The remainder of this paper will be devoted to justifying what is *not* represented in Figure 1. In particular, I propose to defend the view that there aren't any causal pathways from the outputs of the judgment-generating systems and the decision-making system to mindreading, which would be necessary to allow introspective access to our own judgments and decisions. My thesis is that the mindreading system only has access to perceptual input (in addition to some forms of stored knowledge), and thus that it can only self-attribute judgments and decisions through interpretation of that input, in much the sort of way that it attributes judgments and decisions to other people. (The difference is just that in one's own case the evidential base for interpretation is much greater, including, for example, inner speech and other forms of mental imagery.) As a result, there is no such thing as introspection of judgments and decisions.

What more can be said in support of the architecture depicted in Figure 1, together with

its correlative denials of introspection? A large part of the answer to this question will be given in Section 3, where I shall present a variety of kinds of direct empirical evidence. But there are also more general arguments of an evolutionary and anatomical kind. There exists a good answer to the question why an “outwardly focused” mindreading faculty of the sort represented in Figure 1 (or the capacity to construct such a faculty via learning) might have evolved. This is some or other version of the “Machiavellian intelligence” hypothesis (Byrne and Whiten, 1988, 1998), which points to the immense fitness advantages that can accrue to effective mindreaders amongst highly social creatures such as ourselves. We also have good evidence that the brain is constructed in such a way as to realize the global broadcast of perceptual events, thus facilitating other-directed mindreading *inter alia*, together with introspection of such events as a by-product.

In contrast, there aren't any good proposals concerning the powerful selection pressures that would have been necessary to construct and preserve the brain mechanisms needed to realize introspection of judgments and decisions.⁵ (They would have had to be powerful, because brain tissue is very expensive to build and maintain. See Aiello and Wheeler, 1995. Moreover, increases in head size bring much-increased risks of both maternal and infant mortality during childbirth, and have necessitated an extensive period of infant dependency that is unique to the human species, which is also very expensive. See Barrett et al., 2002.) Nor do we have anatomical evidence of such mechanisms, which would be needed to link the outputs of all the various conceptual systems with the mindreading faculty.

Philosophers, however, are virtually united in thinking that there is introspection for judgments and decisions, just as there is for perceptual and imagistic states (Wright et al., 2000; Gertler, 2003). No doubt this is partly because some philosophers are unaware of the relevant empirical evidence and other empirical considerations. But it is also because philosophers' views tend to be much more driven by intuitions than by empirical evidence. And there is no doubt at all that we have a powerful intuition of the existence of introspective access to our own judgments and decisions. I shall argue, however, that this is a *mere* intuition, without any rational ground.

Note that according to the model presented here, visual and other images (including inner speech) will be amongst the forms of evidence available to the mindreading faculty for interpretation, whenever the latter is engaged in self-ascribing judgments and decisions. For the

⁵ See Carruthers (2006, 2008a) for discussion and critique of some of the few suggestions that have been made.

moment I shall assume, as most cognitive scientists do, that while inner speech and other imagery might be *expressive* of underlying thought processes, they aren't *constitutive* of those processes. So the fact that we can introspect our own inner speech does nothing to support the view that our judgments and decisions are similarly introspectable. In Section 4 I shall return to consider whether any of this changes once we allow that inner speech might be (partially) constitutive of certain types thought, as many kinds of “dual systems theory” of human thinking and reasoning would imply.

3 The evidence against introspection of judgments and decisions

The model presented in Section 2 predicts that there should be no introspective access to judgments and decisions. The purpose of the present section is to marshal evidence in support of the correctness of this prediction.

A number of cognitive scientists have made similar claims—arguing that we often suffer from the *illusion* that we are introspecting our own judgments and decisions—and the present section is much indebted to their work (Gopnik, 1993; Gazzaniga, 1995, 2000; Wegner, 2002; Wilson, 2002). But some of these writers fail to address the question whether their denials of introspection extend also to perceptual states. Hence it isn't clear whether they would endorse the Figure 1 architecture, with its correlative prediction of a divided result (introspection for perceptual but not for propositional attitude events). Nor, for the most part, do they deny (as I do) that (non-perceptual) judgments and decisions are *ever* introspectable. On the contrary, Wegner and Wilson, at least, are explicit in allowing that such introspection can occur, thus endorsing a form of “dual process” account. This combination of views will be addressed in Section 4, where I shall argue that the introspectable events in question don't have the right sorts of causal role to count as genuine judgments or decisions.

It is worth noting that both Gazzaniga (2000) and Wilson (2002) think that the mindreading system, when turned upon oneself, is doing more than just interpreting. Both think that the system plays additional roles in building a self-narrative and maintaining a positive self-image. (This might even be a secondary adaptive function of the mindreading faculty, acquired subsequent to its initial “Machiavellian” benefits.) One might conceptualize this as a kind of *interpretation within constraints*, not unlike what sometimes happens in science when only hypotheses that are consistent with a particular background theoretical framework are taken

seriously. And something like it arguably occurs in the case of other-interpretation, too—as when the need to preserve one’s marriage makes one (unconsciously) consider only certain types of explanation for one’s spouse’s behavior. In what follows, therefore, I shall not assume that the interpretative role of the mindreading faculty is always “pure”, uncontaminated by the agent’s goals and needs.

3.1 *Split brains*

Gazzaniga (1995, 2000) defends a view similar to that outlined in Section 2, grounded in a variety of kinds of experimental data from commissurotomy (“split brain”) patients, some highlights of which I shall shortly describe. Gazzaniga argues that the left brain houses the main elements of the mindreading system (which he dubs “The Interpreter”), with access to perceptual, somatosensory, and proprioceptive input, but with no access to the judgments, reasoning processes, or intentions of the subject.⁶ The mindreading system continually monitors the evidence available to it (the circumstances, the agent’s own bodily movements or intimations of movement, together with bodily–emotional reactions, inner speech, visual imagery, and so forth) and weaves together an interpretative story. Often enough, the story is actually correct, and the judgments and decisions attributed to the self are accurate. But sometimes the data are misleading or (as in the case of split-brain patients) absent altogether, and confabulation results.

In one famous case (representative of many, many, others of similar import) Gazzaniga (1995) describes how different stimuli were presented to the two half-brains of a split-brain patient simultaneously. The patient fixated his eyes on a point straight ahead, while two cards were flashed up, one positioned to the left of fixation (which would be available only to the right brain) and one to the right of fixation (which would be available only to the left brain). When the instruction, “Walk!” was flashed to the right brain, the subject got up and began to walk out of the testing van. When asked why, he (the left brain, which controls production of speech as well as housing a mindreading system) replied, “I’m going to get a Coke from the house.” This attribution of a current intention to himself is plainly confabulated, but delivered with all of the

⁶ Some theorists dispute the claim that mindreading (like language, to a significant degree) is an exclusively left-brain process (Hirstein, 2005). This is irrelevant for purposes of the present argument. Even if each half brain houses a self-contained mindreading system, each will still need to confabulate when explaining actions initiated by the other half brain.

confidence and seeming introspective obviousness as normal.⁷

On another occasion a picture of a chicken claw was flashed to the right of fixation (available only to the left brain) and a picture of a snow scene was flashed to the left of fixation (available only to the right brain). The subject was then asked to choose an associated item from an array of pictures placed in full view (available to both hemispheres). With his right hand (controlled by the left hemisphere) the subject chose a picture of a chicken, and with his left hand (controlled by the right hemisphere) he chose a picture of a shovel. Thus the left brain had (correctly) matched the chicken with the chicken claw, and the right brain had (again correctly) matched the shovel with the snow scene. (The salience of shovels connected with snow is especially high in the north-east of the United States, where these experiments were conducted.) But when asked to explain his choices, the subject (his left brain) replied, “Oh, that’s simple: the chicken goes with the chicken claw, and you need a shovel to clean out the chicken shed.” The latter part of this explanation had plainly been confabulated, resulting from the fact that the mindreading system that answered the question lacked access to the events that led to the choice of the shovel (to go with the snow scene).

It is important to note that while commissurotomy patients can often have good understanding of their surgery and its effects, they never say things like, “I’m choosing this because I have a split brain and the information went to the right, non-verbal, hemisphere” (Gazzaniga, 1995). On the contrary, they often make their confabulated reports smoothly and unhesitatingly, and their (their left brain’s) sense of self seems quite unchanged following the operation.⁸ Even reminders of their surgery during testing have no effect. On a number of occasions testing was paused and the experimenter said something like, “Joe, as you know you have had this operation that sometimes will make it difficult for you to say what we show you over here to the left of fixation. You may find that your left hand points to things for that

⁷ Note, however, that the attribution of an intention to oneself, once made, can become self-fulfilling, as we shall see in Section 4. Hence I would be prepared to bet that, if he hadn’t been interrupted by the curious experimenter, the subject would indeed have gone into the house and got himself a Coke.

⁸ Admittedly, subjects will also sometimes express their confabulated reports with low confidence, or say that they don’t know why they did something. The existence of such cases is no problem for my account; indeed, they are predicted by it. For if the mindreading system is unable to come up with an explanation fast enough, then subjects will become aware that they are self-interpreting, and will thus no longer have the impression that they are introspecting.

reason, OK?” Joe assents, but then on the very next series is back to showing the interpreter effect once again (Gazzaniga, personal communication). If patients were aware of interpreting rather than introspecting, then one would expect that a reminder of the effects of commissurotomy would enrich the hypothesis pool, and would sometimes lead them to attribute some of their own behavior to that. But it doesn't.

Of course it doesn't follow from the extensive commissurotomy data that normal human subjects never have privileged, immediate and non-interpretative, access to their own judgments and decisions, as Goldman (2006) points out. Gazzaniga's data were collected from patients who had undergone serious brain damage (a severed corpus callosum). Hence it may be that in normal brains the mindreading system does have immediate access to the agent's judgments and intentions. The split brain data force us to recognize that *sometimes* people's access to their own judgments and intentions can be interpretative (much like their access to the judgments and intentions of other people), requiring us at least to accept what Goldman (2006) calls a “dual method” theory of our access to our own thoughts. But one could believe (as Goldman does) that introspection is the normal, default, method for acquiring knowledge of our own propositional attitudes, and that we only revert to self-interpretation as a back-up, when introspection isn't available.

The split-brain data do seem to show decisively that we have no *introspective* warrant for believing that we ever have introspective access to our own judgments and decisions, however. This is because patients report plainly-confabulated explanations with all of the same sense of obviousness and immediacy as normal people. It follows that subjects themselves can't tell when they are introspecting and when they are interpreting or confabulating. So for all we know, it may be that our access to our own judgments and decisions is *always* interpretative, and that we *never* have introspective access to them.⁹

I have argued elsewhere that the best explanation of the mindreading system's inability to

⁹ Given reliabilist conceptions of knowledge and justification, of course, we might count as knowing, and as justified in believing in, the existence of introspection, despite our inability to discriminate cases of introspection from cases of confabulation. This will be so provided that introspection really does exist and is common, and provided that our belief in it is caused by the fact that we often introspect, and caused in the right sort of way. My point here, however, is that our inability to discriminate shows that we have no *subjectively accessible* reason to believe in the existence of introspection. So if we are wondering whether or not introspection is real, we should find the split-brain data deeply troubling.

tell whether or not it is introspecting is that our belief in introspective access is actually one of the simplifying assumptions made by the mindreading system itself (Carruthers, 2008b). Indeed, the mindreading system appears to operate with a model of its own access to the rest of the mind that is essentially Cartesian. It assumes that subjects know, immediately and without self-interpretation, what they are experiencing, judging, and intending. This is what makes it possible for Locke (1690) to write, “There can be nothing within the mind that the mind itself is unaware of.” And it is, no doubt, why the very idea of *unconscious* mental states has seemed so counter-intuitive to most people, and has historically met with such resistance. It will, likewise, be the main source of resistance to the idea that we know of our own judgments and decisions by self-interpretation. For we aren’t *aware* of engaging in any such process of interpreting.

The Cartesian assumption has an obvious heuristic value, greatly simplifying the mindreading system’s computations. If we grant that the mindreading system is for the most part reliable, producing true beliefs about the subject’s own judgments and intentions, then it will make the operations of that system a great deal simpler if it assumes that those beliefs are always a result of introspection. For from the point of view of what matters (such as judging whether someone is being sincere when they report an intention of theirs, or report a belief), it is unimportant whether people’s beliefs about their own states result from introspection or rather from self-interpretation, provided that the latter process is sufficiently reliable.

What the split brain data strongly suggest, then, is that our common-sense belief in the existence of introspective access to judgments and decisions is without epistemic warrant, and that sometimes, at least, our access to our own attitudes is actually (but unconsciously) interpretative. In consequence, since the “interpretation only” theory is simpler than its “dual method” rival, this gives us some reason to think that we *never* have introspective access, and that our beliefs about our own judgments and intentions are *always* a result of self-interpretation. Arguments from simplicity are relatively weak ones, however, especially in the biological domain, where we should expect to find systems that are messy and complicated. We need to ask, therefore, whether there are other considerations favoring the “interpretation only” approach. I shall argue that there are.

3.2 *Confabulation induced by brain stimulation*

If subjects always arrive at beliefs concerning their own judgments and decisions via an

interpretative inference from the data available to them, rather than via introspection, then a number of predictions can be made. One is that if actions can be induced directly, in ways that bypass the decision-making process, but in ways that are unknown to the subjects, then people should still claim ownership of those actions, and should claim to have been aware of deciding to perform them. This is just what Brasil-Neto et al. (1992) found, using focal magnetic stimulation of areas of motor cortex.

Subjects were instructed that when they heard a “click” (actually the sound of the magnet being turned on) they should lift one or other index finger, and that they were free to choose which finger to move. Areas of motor cortex were then stimulated, either on the right or the left. In such cases subjects showed a marked tendency to raise the index finger of their contralateral hand (provided that they made the movement close enough in time to the onset of the magnetic burst), but in each case they claimed to have been aware of *deciding* to lift that finger. Control experiments in which the magnets were directed away from the subject’s head, or in which other areas of the brain were stimulated—including pre-motor cortex—produced no effects.

Everything that we know about the organization of the brain suggests that motor cortex isn’t the place where decisions themselves are located, nor is there any plausible route via which stimulating motor cortex could cause a decision to be made.¹⁰ So the cranial stimulation is unlikely to have caused an intention to lift the contralateral finger to come into existence, which is then introspected. Rather, the cranial stimulation, in the presence of a background intention to lift one finger or another on hearing the click, directly caused the lifting of the contralateral index finger. And the subject, knowing that he was intending to lift one finger or another when he heard the click, and knowing nothing of the likely effects of magnetic stimulation, therefore deduced that he had lifted that finger intentionally. Since from the subject’s perspective the best explanation of the data available to him is that he chose to lift the index finger that subsequently moved, that is what he reports. But he is unaware that he makes this report as a result of an interpretation. Rather, he thinks that he is *aware* of his decision.

These data are, admittedly, consistent with the “dual method” account of self awareness. It may be that on other occasions people do introspect their decisions. But the costs of adopting

¹⁰ Notice that it is unlikely that stimulation of motor-cortex should have caused a decision to be made within the frontal cortex via the back-projecting neural pathways that connect the two. For in that case one would have expected stimulation of pre-motor cortex to have the same effect.

such a view are, now, significantly increased. For it will have to be allowed that subjects don't *only* resort to interpreting their own behavior in those rare instances where they are physically cut off from the usual sorts of introspectable mental events (as in cases of commissurotomy), or in those cases where introspection isn't available. Rather, more generally, it will have to be claimed that subjects resort to interpretation in all cases where they happen not to introspect a relevant intention, but where they take themselves to have good reason to believe that an intention was involved. Hence when a subject in Brasil-Neto et al.'s experiments fails to detect by introspection a decision to move a particular finger, but (being unaware of the effects of magnetic brain stimulation) thinks that such an intention must have been present (remember, he was set up by the experimenters to believe that he should be freely choosing which index finger to lift on each trial when he hears the click), he interprets himself accordingly.

3.3 *Confabulation following hypnosis*

The data on confabulation concerning actions undertaken as a result of post-hypnotic suggestion are even harder to explain away. Sometimes when hypnotized subjects are instructed to perform some action later, after they have come out of their hypnotic state, they do just that: they perform the action suggested, often with no memory of the instructions having been given or of the hypnotic episode itself (Edwards, 1965; Sheehan and Orne, 1968). But when asked to explain their actions they will (if they have no memory of the instructions, and especially if they are unaware of having been hypnotized) offer one that seems plainly confabulated. For example, a hypnotized subject might be told, "When you awake you will take a book from the table and place it on the shelf." When he later performs such an action and is asked why he decided to do it, he might respond, "I don't like to see things untidy; the bookshelf is the place for books, so that is why I am placing it there" (Wegner, 2002). What he reports, in effect, is a decision to tidy the room.

The most plausible explanation of the episode is as follows. The subject performs the action for the same reason that he does whenever following instructions from another person, and just as he does during the hypnotic episode itself: he acts as he does because he has been requested to do so, wants to comply, and has no contrary motive; so he decides to do what he has been asked. But lacking any knowledge of these reasons, or of the ensuing decision, his mindreading system sets to work, when prompted, to construct the most plausible explanation

that it can, and the result is then self-attributed with the usual sense of introspective obviousness.

Notice that the subject's explanation, here, can be offered either during or immediately following the action itself. So it isn't plausible to claim that the confabulated explanation results merely from lapses of memory. That is, it isn't plausible to claim that while the subject's actual judgment and decision were once introspectable, those events have been lost from memory by the time that the question is asked. Nor is it easy to discern any alternative route via which the decision to tidy the room might have been caused by the hypnotist's suggestion, thus rendering the introspective report veridical. For why should the instruction to place a book on the shelf cause a subject to believe that the book is out of place, and to form the intention of tidying it away?

Moreover, the patterning in the data across the full range of cases of hypnotic suggestion supports the self-interpretation model, as Wegner (2002) points out. For subjects performing an action resulting from post-hypnotic suggestion are much less likely to confabulate an explanation for their behavior if they are aware that they have been hypnotized (since the well-known phenomenon of post-hypnotic suggestion provides them with an alternative explanation). And they are much less likely to confabulate an explanation if their behavior is especially bizarre. This is because it is thereby harder to construct an explanation that will seem intuitively plausible; and if it takes too long to think up an explanation, subjects will become aware that they are reflecting, and then it will no longer seem to them that they are introspecting.

A "dual method" theorist such as Goldman (2006) can perhaps explain (away) these data, but only on the assumption that the decisions caused by the hypnotic instructions are located in some sub-system of the mind that is inaccessible to introspection, or if it is assumed that a secondary effect of the hypnosis is to somehow block introspective access to the mental events that cause the action, which would otherwise have been introspectable. Such suggestions might conceivably be correct. But they lack any independent motivation. (And remember, the "split brain" data show that we lack any subjectively-accessible warrant for the common-sense belief in the reality of introspection. So we really do need independent evidence at this point.) It is simpler and more plausible to suppose that there is no such thing as introspection of judgments and decisions, and to conclude that self-interpretation is always the mode in which such events are self-ascribed.

3.4 *Experimentally manipulated confabulation*

If people always form their beliefs about their own judgments and decisions via an interpretation of the available data (both overt and introspective), then it ought to be possible to manipulate people's sense that their actions were or weren't intended by manipulating the cues that they use when interpreting. This is just what Wegner and Wheatley (1999) set out to do. They reasoned that people's sense of having willed an action should be especially sensitive to two factors. One is the existence of cues prior to the action that are semantically related to the action or outcome, such as the occurrence of a word or phrase in inner speech describing the action, or a visual image of the outcome. The other is that semantically related cues that occur close to the action in time should be more likely to lead to an interpretation of intentionality, just as precursor events in the physical domain are more likely to be seen as causal when they occur just prior to some effect.

Wegner and Wheatley (1999) asked subjects to cooperate with another person (actually a confederate of the experimenter), jointly moving a large computer mouse in order to shift a cursor around on a computer screen, on which a variety of images of familiar objects were displayed. It was explained to subjects that the experiment was to investigate how people's feelings of intention and control come and go. Both the subject and the confederate wore headphones, and were asked to move the cursor around together for about 30 seconds, during or shortly after which they would each hear a word spoken aloud, ostensibly as a distracter. The subject was led to believe that the confederate would hear a different word (in fact the confederate received instructions from the experimenter). They were told that after 30 seconds some music would begin playing, and subjects were asked to wait a few moments before taking a decision to stop moving the mouse (and hence the cursor) at a point of their choosing thereafter. Meanwhile the confederate was receiving instructions intended to guide the cursor to be in contact with an image of a particular object of the type named on the subject's tape (e.g. an image of a swan, when the subject would hear the word "swan"), and to bring the mouse to a stop with the cursor near that object as soon as possible after the music began to play. (The spoken word, of course, was intended to prime for thoughts of the corresponding object.)

The variable manipulated through the experiment was the time that elapsed between the subject hearing the word and the cursor being brought to a stop. With a 30 second delay between the former and the latter, subjects reported only *some* sense that they had intentionally brought

the mouse to a stop beside the appropriate object. This impression increased linearly to a maximum when the word was heard between five seconds and one second before the stop, however, and collapsed again when the word was heard shortly *after* the cursor had stopped. In control experiments designed to see whether hearing the word would actually cause subjects to form an intention to stop near the appropriate object, the confederate was instructed not to initiate a stop, leaving it to the subject to do so. There was found to be no statistical relationship between the points on the screen where the subject brought the cursor to a halt and the position of the named object. So it is unlikely that subjects were accurate in reporting a decision to stop beside the named object under conditions of confederate control.

This experiment provides dramatic confirmation of the anti-introspection position.¹¹ We know that subjects were confabulating their reported decisions to make the cursor stop at the point beside the object that had previously been named, because they showed no such tendency when actually given control of the stops. And subjects' tendency to confabulate such an explanation could be manipulated by the simple expedient of varying the temporal interval between hearing the word and the time of the stop, just as would be predicted if subjects were arriving at judgments of mental causation in much the same sort of way that they arrive at judgments of physical causation—by interpreting, with temporal contingency being one important interpretative factor.

3.5 *Interim discussion*

I have argued that the best explanation of the evidence surveyed in this section so far is that our access to our own judgments and intentions is always grounded in our interpretation of the data available to the mindreading faculty (where the process of self-interpretation isn't itself

¹¹ Wegner (2002) reports a related experiment in which subjects stood in front of a mirror with a confederate standing invisible behind them. The confederate's arms were inserted through the sleeves of the gown worn by the subject, appearing in place of the latter's own (and with gloves on the hands to obscure identification). When the confederate moved his arms about in accordance with instructions played privately through headphones, subjects reported that the experience was a little eerie, but felt no authorship of the perceived movements. However, when subjects, too, heard the instructions, their sense of producing the movements for themselves increased significantly. They also acquired an emotional attachment to the confederate's hands. Under these conditions (and only under these conditions) subjects underwent a sharp skin-conductance response when viewing one of the hands being snapped painfully by a rubber band at the end of the performance.

conscious, of course), and hence that such access doesn't qualify as introspective in character.

Someone might object, however, that all of the evidence that has been adduced is based upon pathological, unusual, or highly manipulated cases. An argument grounded in such evidence then seems a bit like the argument from illusion in the philosophy of perception against the idea that we have direct perceptual contact with external objects. And this argument is generally reckoned to be fallacious. The existence of illusions deriving from clever deceptions and brain manipulations doesn't show that I lack direct perceptual contact with my coffee cup in the normal case (at least not without considerable further argument). Likewise the occurrence of confabulation deriving from clever deceptions and brain manipulations doesn't, without considerable further argument, show that I am not in direct introspective contact with my judgments and decisions in the normal case.

However, the form of argument that I intend isn't this: "We often go wrong without knowing it, and we can't introspectively distinguish the cases where we go wrong from the cases where we get it right, so when we do get it right our knowledge isn't direct and unmediated." (This is the argument from hallucination in the philosophy of perception.) The claim is rather that the specific ways in which we go wrong without knowing it can show us something about the manner in which the self-knowledge faculty operates. It is thus much more similar to arguments from illusion in cognitive science, which are rightly taken to reveal important facets of the way that the visual faculty works.¹²

Perceptual illusions are regarded as vital data by vision scientists, revealing key facts about the assumptions that are built into the visual faculty, such as that light shines from above, or that moving objects are locally rigid (Gregory, 1978; Palmer, 1999). But the resulting account of vision as inferential in character is perfectly consistent with the philosopher's idea of vision as involving direct perceptual contact with external objects, given the way that the latter account is intended. For the assumptions and inferences in question are all "sub-personal", taking place within the visual faculty in ways that don't involve the agent's beliefs. In contrast, the

¹² Seen in this light, a "dual method" theorist about self-knowledge would be like a vision scientist who concluded from the data on hallucination and visual illusion that there are actually two distinct visual faculties, rather than just one. This would be absurd. And note that it would *much* stronger than anything claimed by philosophers who have proposed disjunctive theories of vision to accommodate such data (Snowdon, 1990; McDowell, 1994). Their claim is only that the visual faculty issues in two different types of perceptual *state*.

implications of the confabulation data reviewed above are much more destructive of our ordinary conception of self-knowledge. There are two main differences. The first is that the assumptions that are shown to be operative when we attribute propositional attitudes to ourselves aren't just universal ones (such as "light shines from above"), but also involve specific facts about our own circumstances or current or recent behavior. And the second is that these assumptions aren't generally sub-personal, but are often reportable beliefs. (For example, a subject in the focal magnetic stimulation experiments discussed in Section 3.2 will know, and can report, that he has just moved his left index finger.) These differences seem sufficient to undermine our common-sense view that our access to our own mental states is direct, and radically different from our access to the attitudes of others.

I claim, then, that when it is understood properly the argument from "unusual cases" against introspection of judgments and decisions is sound. But in any case, however, it would be false to claim that all of the evidence derives from cases that are pathological, unusual, or highly manipulated. On the contrary, there is a wealth of evidence from social psychology to the same effect, deriving from perfectly ordinary situations. Some of this evidence will be reviewed in the sub-section that follows.

3.6 *The social psychology literature on confabulation*

Over the last fifty years an extensive literature has been built up concerning confabulation in normal individuals in everyday life, sometimes traveling under the name, "cognitive dissonance", sometimes under the title, "self perception" (Festinger, 1957; Bem, 1967, 1972; Wicklund and Brehm, 1976; Nisbett and Wilson, 1977; Eagly and Chaiken, 1993; Wilson, 2002). Unfortunately for our purposes, much of this literature isn't very tightly focused on self-knowledge of judgments and decisions, but rather (for example) on how much enjoyment someone reports from playing with a particular game, pitted against how long they will *actually* spend playing with it in their free time (Kruglanski et al., 1972). And sometimes (as here) the reports are given long enough after the fact that the data might be explained in terms of failures of memory, rather than by failure or absence of introspection.

In addition, some data might be said to relate to mistakes about the causes or effects of our mental states, rather than to confabulation of the occurrence of those states themselves. This might be true, for example, of the famous "love on a bridge" experiments, in which members of

one set of male participants reported greater physical attraction after being questioned by a young woman while swaying on a dangerous-seeming rope bridge above a chasm than did another set who were questioned by her after they had crossed the bridge (Dutton and Aron, 1974). For it is possible that heightened danger might have led to heightened attention and awareness, which in turn could have caused the young woman to seem especially attractive.

Yet other data may be explicable by postulating the right kind of unconscious mental process leading to the existence of an introspectable mental event. This might be true in connection with some of the data collected by Nisbett and Wilson (1977), whose subjects displayed a marked right-hand bias when asked to select an item from an array of identical objects (nylon panty-hose, say), but who when questioned immediately afterwards about the reasons for their choice said that they chose as they did because (for example) that item looked softer. While it might seem that this report of their judgment must be confabulated, an alternative explanation can be constructed. Notice, first, that the reason for right-hand *choice* bias is likely to be right-hand *attention* bias: people generally pay more attention to what is on their right, and this is what leads them to choose what is on their right. But there are then two different ways that the causal story can be told. Either paying attention to the right-most item leads to choice without any relevant judgment as to quality, and the subject's report of a judgment of superior softness is then confabulated. Or paying attention to the right-most item leads to the judgment that it is softer (where this judgment is perhaps created unconsciously by the mindreading faculty to explain the greater attention that is being paid to the right-most item), which in turn causes choice. In this case, the report is veridical, and could result from introspection.

The social psychology literature also includes, however, an extensive set of studies on the effects that people's own behavior can have on their reports of their current judgments (Eagly and Chaiken, 1993).¹³ Hundreds of experiments have been conducted over a fifty year period replicating such effects and exploring their parameters. One robust finding in this literature is

¹³ One reason why the significance of these experiments may have escaped the notice of philosophers is that social psychologists themselves often use the language of belief *change* rather than belief *confabulation*. This is because they operationalize beliefs to be linguistic dispositions, and because the effects of the experimental manipulations upon both verbal report and subsequent behavior can be long-lasting. Plausible mechanisms underlying these changes will be explored in Section 4. But it should be plain that the first report of a changed occurrent judgment in these experiments is confabulated, even if the result of that report is to create something very much like a new belief.

that people who have been cleverly manipulated into writing an essay for a paltry sum of money in defense of something that they initially disagree with will end up, after the fact, expressing much more sympathy for the position that they have defended than will other people who were paid a decent amount (Cohen, 1962; Linder et al., 1967). It seems that the former group, rather than introspecting their current judgment or degree of belief, reasoned somewhat as follows: “Since I spent all that time writing the essay for such a small sum of money, and since it was my choice to do so, it must be the case that I thought it worthwhile to defend the position in question. So I must believe it.” The well-paid group, in contrast, had a ready explanation for their behavior: they were paid a significant sum to write the essay. So they could just answer the question about what they believe as they normally would, considering only the proposition in question, and asking whether or not they find themselves inclined to assert it.

How might a defender of introspection respond to this sort of case? Since it is mysterious why the mere fact of being asked about one’s judgment of the issue should somehow block the normal introspective process in these circumstances, it looks like the answer will have to be that the process of writing the essay for inadequate pay somehow caused a greater degree of belief in its subject matter. So the report given at the end can be both introspection-based and veridical. But this is puzzling: why should writing an essay under conditions of inadequate payment lead to heightened belief, when going through the very same process for adequate pay doesn’t? (Note that it can’t be the mere fact of thinking up good arguments, and so forth, that produces belief, unless for some reason those who are paid less should argue better!) It seems that the only explanation is that while writing the essay the subjects must tacitly have been asking themselves, “Why am I doing this for so little reward when I don’t have to?”, to which the answer delivered by their mindreading faculty was, “Because I believe in it.”

There are two problems here for defenders of introspection. One is that it just pushes the failure of introspection back into the judgments through which the concluding belief is formed. For by hypothesis, if the person had been asked, “Do you believe this?” while writing the essay, he would have expressed a greater degree of belief than would someone who was paid significantly more. So the self-attribution of this earlier judgment would have been confabulated, even if the later one wasn’t. But the second problem is that it is in any case mysterious how an increased propensity to *attribute* a belief to oneself should lead to greater *actual* belief. How

does the higher-order attribution process “filter down” into the first-order judgment itself?¹⁴ But without this, it cannot be claimed that the report made at the end of the experiment resulted from a veridical introspection of a first-order judgment on the topic.

Moreover, the confabulation-based account of these data makes an obvious further prediction, which turns out to be accurate. The relevant claim is that subjects are self-ascribing belief, not on the basis of introspection, but rather through the application of a folk theory to explain the behavioral data. For example, they might be utilizing the generalization: “If people do something for inadequate reward, but do so freely, then they must be intrinsically motivated to do it.” In the present instance, this amounts to saying that the essay is being written because the person believes the proposition that the essay is defending. If this is right, then someone who doesn’t participate in one of these experiments but is just told the relevant details (not about the initial contrary belief, but just about the task and the extent of the pay) should make the very same attributions. This has been tested, and turns out to be the case (Bem, 1967).

Lest it be thought that I have cherry-picked just a single set of studies to make my case, let me briefly describe one other. It has long been known that asking subjects to nod their heads while listening to a message (ostensibly to test the headphones that they are wearing at the time) increases their reported degree of belief in the message, whilst asking them to shake their heads while listening will increase their expressed disagreement (Wells and Petty, 1980). One explanation of this result is that the mindreading faculty interprets nodding as a signal of agreement, and hence confabulates a heightened degree of belief, and that it interprets head-shaking as a sign of disagreement, leading it to confabulate accordingly. But other explanations have also been proposed. For example, it may be that nodding biases positive thoughts about the contents of the message and inhibits negative ones, whilst head-shaking has the opposite effect. This would naturally lead to changes in degrees of belief which could be introspected and veridically reported. However, Briñol and Petty (2003) devised an elegant series of experiments to test between these and other alternatives, and provide decisive evidence in favor of the self-interpretation explanation.

Briñol and Petty not only manipulated the degree of persuasiveness of the message

¹⁴ This is a question to which we will return in Section 4. But the processes that we will investigate there aren’t ones that would lend any support to the introspectionist at this point. For they operate only following explicit verbalized self-attributions of belief, not unconscious ones of the sort that are envisaged above.

listened to (as evaluated by independent raters), but they also asked subjects to recall and report what was passing through their minds while they listened (e.g. what they visualized or said to themselves). What they found was that when the message was persuasive, nodding increased belief while head-shaking decreased it. But when the message was unpersuasive they got the opposite result: nodding *decreased* belief while head-shaking *increased* it. This is immediately inconsistent with the hypothesis that nodding increases belief by priming for positive thoughts and inhibiting negative ones. The true explanation emerges when the reported concurrent thoughts are examined. When the message is persuasive subjects tend to think positive thoughts (“That would be great!”), whereas when it is unpersuasive they tend to think negative ones (“What a terrible argument!”). It seems that subjects interpret their own nodding behavior as affirming or agreeing with the thoughts passing through their minds at the time, and that they interpret their head-shaking as disagreement. Hence when thoughts are positive, nodding increases belief and head-shaking decreases it; whereas when thoughts are negative, nodding decreases belief and head-shaking increases it.

To conclude this sub-section: I have argued that at least some of the studies in the social psychology literature on confabulation speak strongly in support of the “interpretation only” position with respect to self-knowledge of judgments and decisions, and against the “dual method” alternative. But in addition, the overall patterning of the data across the full range of studies supports the anti-introspection position. For in order to explain the data, the defender of introspection is forced to introduce to a variety of different explanations. Sometimes lapses of memory are appealed to, combined with the general principle that when introspection isn’t available, people interpret themselves as best they can (without realizing that they are doing so). Sometimes unusual patterns of causation of judgments or decisions are appealed to (different from case to case). In contrast, the anti-introspection theorist can provide a unified explanation across the different experimental studies. For what is common to all cases is that subjects have inadequate theories that they use when interpreting themselves. Either they lack knowledge of the relevant causal pathways (such as the effects of right-hand bias upon choice), or they have theories that are misleading or false (such as the belief that people doing something for inadequate payment must be intrinsically motivated to do what they do), or they have theories that are generally true but inapplicable in the present case (such as that someone nodding his head is signaling agreement).

3.7 Summary discussion

Taken collectively, the empirical evidence (only part of which has been surveyed here, of course) makes a powerful case in support of the “interpretation only” account of self-knowledge of judgments and decisions. But of course it doesn’t *entail* that account. Rather, the data warrants an inference to the account as the *best explanation* of the available evidence. Hence people can, consistently with the data, continue to insist on a “dual method” alternative if they wish. But they can only do so at the cost of explanatory simplicity, as we have just seen. Moreover, our “interpretation only” account has the resources to explain why a belief in introspection of attitudes should have such a powerful hold on us. This is because it is one of the simplifying assumptions made by the mindreading faculty itself, which doesn’t factor into its calculations or make available in its output its own theorizing activity (Carruthers, 2008b).

Can it be objected that the data surveyed in the present section pertains to mistakes about the *causes* or *effects* of our judgments and decisions, rather than to the latter kinds of event themselves? If so, then the “interpretation only” account can be resisted. For no one now thinks that the causes and effects of our judgments and decisions are introspectable (except where those causes or effects are themselves judgments or decisions). This objection can’t be sustained, however. Consider the case of focal magnetic stimulation (Section 3.2), for example. As we noted, it is very unlikely that subjects should actually have decided to move the correct finger each time, caused by the motor-cortex activity that caused that finger to move. For magnetic stimulation elsewhere on the pathways between motor cortex and frontal cortex fail to cause any such decision. Likewise if we consider the manipulations conducted by Wegner and Wheatley (Section 3.4): it is very unlikely that hearing the word “swan” should *actually* have caused subjects to decide to stop the cursor beside the image of a swan, since in cases where they were given control they displayed no such tendency.

In relation to some of the other data, the thesis that subjects are ignorant of causes rather than failing to introspect is marginally more plausible, albeit still unsustainable. Consider the effects of hypnotic instruction (Section 3.3). It is possible that the subject did indeed make (and introspect) a decision to tidy the table, caused by the previous instruction to place the book on the shelf. But there is no clear account of the causal mechanism involved. Why should the instructions cause an distinct decision that would have the same effect, rather than causing (as

normal) a decision to comply with the instruction? Likewise, consider the effects of essay writing on judgments about the plausibility of the thesis written about (Section 3.6). Although it is possible that subjects did indeed make (and introspect) the judgment that capital punishment is justifiable, caused by an unconsciously-produced higher-order belief that they *believe* capital punishment to be justifiable (which had been produced, in turn, as the best—but confabulated—explanation of their behavior in the circumstances), once again we lack a clear account of the mechanism involved. And in contrast, since confabulation must *already* have taken place in order to this proposal to work, it is much simpler to suppose that subjects' reports of their judgments were the result of confabulation.

Taken all together, then, our denial of the introspectability of judgments and decisions would appear to be amply warranted.

4 Replies to objections

In the present section I shall consider and reply to objections to the anti-introspectionist position that has been developed above.

4.1 Dual systems theory

Our discussion up to this point has taken place under the assumption that introspectable items of inner speech and other imagery aren't, themselves, constitutive of judging and deciding, but serve merely to express an underlying set of thoughts, which thereby become accessible to the subject *indirectly*. But this assumption can be challenged from the perspective of “dual systems theories” of reasoning and decision making, as I shall now explain.

Almost everyone who works on the psychology of human reasoning has converged on some or other version of dual systems theory (Evans and Over, 1996; Stanovich, 1999; Kahneman, 2002). On this account the human mind contains two distinct *types* of system for arriving at new judgments and decisions. System 1 (really a set of systems, arranged in parallel) is fast, unconscious, hard to alter, universal to all thinkers, and evolutionarily ancient. System 2, in contrast, is slow and serial, characteristically conscious, malleable in its principles of operation, admits of significant variations between individuals, and is evolutionarily novel. And a number of authors have emphasized the important constitutive role played by imagery (especially inner speech) in the operations of System 2 (Evans and Over, 1996; Frankish, 2004;

Carruthers, 2006).

On such an account, a representation of the sentence, “Capital punishment is permissible”, or of the sentence, “I’ll open *that* box”, figuring in inner speech, can be partly constitutive of the subject’s (System 2) judgment that capital punishment is permissible, and of the subject’s (System 2) decision to open the box, respectively. In which case, since these imagistic events are introspectable (according to the account of introspection outlined in Section 2), it might seem to follow that the corresponding judgments and decisions are similarly available to introspection. And we would have here a principled version of Goldman’s “dual method” view. Granted, some judgments and decisions (that is, those occurring in System 1) aren’t introspectable (this is what the data reviewed in Section 3 would confirm); but some judgments and decisions *are* introspectable (namely, those that figure in System 2, expressed in globally broadcast imagery of one sort or another).

While we should accept the truth of some sort of dual systems theory, and accept the constitutive role played by visual imagery and inner speech in the operations of System 2 as a whole, we should deny that this gives us a vindication of the claim that some judgments and intentions can be introspected. The first problem is this. Even if the *content* of an “utterance” in inner speech is trivially accessible to the subject (because it can be embedded in a content-report by semantic ascent), the *attitude* that the utterance expresses certainly isn’t. What constitutes something as a judgment, rather than a supposition or a doubt, is its causal role. But the causal role of an utterance in inner speech isn’t accessible to introspection, even if (as I grant) the utterance itself is. It can only be known by inference and interpretation.

Put differently: utterances don’t wear their attitudes upon their sleeves. An overt utterance of, “Capital punishment is permissible”, for example, might be a sincere assertion (an expression of belief), or it might be said in suppositional or pretence mode, or it might be said as an expression of doubt, or even incredulity. (Similar points could be made about the various roles that might be played by a visual image of a box being opened.) Admittedly, these differences are sometimes signaled by tone of voice. But by no means always. And in any case tone of voice can often be used to mislead an audience. The same then holds when an utterance is mentally rehearsed in inner speech. Unless we take for granted the introspectability of decisions (such as the decision to rehearse a sentence as a sincere expression of belief), the role of an utterance as (partially constitutive of) a System 2 judgment just isn’t accessible to

introspection.

Someone might concede that we can't know, purely by introspection, that we have made a judgment or a decision; for the causal role of the introspectable event in question isn't accessible to introspection. But it might be insisted that we can nevertheless introspect *the judgment* or *the decision*. (Compare: you might be able to see *the cat* on a dark night, even if you can't see *that it is a cat*.) Even this much-weakened introspectionist position isn't defensible, however. For what we access by introspection is a rehearsed sentence in inner speech, or some other relevant sort of image. And this doesn't, and can't, have the right sort of causal role to *be* a judgment or to *be* a decision (as opposed to contributing to some larger process that issues in a new judgment, or in a new decision).

Our idea of a judgment is surely the idea of a mental event that *is* the formation of a new belief with the same content. Hence the resulting belief should arise immediately from the judgment, without the contribution of any further thinking or reasoning. Judging that *P* settles the matter, we think: thereafter one believes that *P* (unless one changes one's mind). Likewise, our idea of a decision is the idea of a mental event that *is* the formation of a new intention with the same content, which may lead to appropriate action in the right circumstances without the need for any further practical reflection. A decision is something that is supposed to *settle* what one does (unless something significant changes; see Bratman, 1987, 1999). But our best theories of the operations of System 2 processes don't confer on inner speech or other forms of imagery these crucial properties.

Frankish (2004), for example, provides an account according to which the operation of System 2 depends upon us coming to have certain other beliefs. In particular, we need to believe that the events in inner speech (or other imagery) that are accessible to introspection constitute *commitments* of one sort or another. Here is how it might work in one simple case. Rehearsing to myself in inner speech the sentence, "Capital punishment is permissible", I come to believe (unconsciously, at the System 1 level) that I have committed myself to the acceptability of capital punishment. I also want to fulfill my commitments. Thereafter, then, if I remember what I have committed myself to, I shall do my best to think (in further episodes of inner speech) and to act as if it were true that capital punishment were permissible. Here the initiating event—the mentally rehearsed sentence—although introspectable, isn't, itself, a "making up of mind", or the formation of a new belief about the acceptability of capital punishment. On the contrary, it only

has its effects via the intervention and activation of further (System 1) beliefs and desires.

The same point holds in connection with the account defended by Carruthers (2006), who suggests a different route via which a mentally rehearsed sentence can give rise to a new belief. On this account, the rehearsed sentence, “Capital punishment is permissible”, might get evaluated by whatever processes would normally check and evaluate the testimony of another person, before storing the content of that person’s utterance as a new belief. So here, too, the rehearsed utterance isn’t itself the formation of a new belief; and a new belief only gets acquired via further processes of (unconscious) thinking and reasoning.

Yet a third account can be extrapolated from the views of Velleman (1989). The suggestion would be that the data available to introspection (inner speech and the like) might give rise to a higher-order belief that I have just formed a belief in the permissibility of capital punishment. And then this combined with my desire for *consistency* will explain my future patterns of thinking and acting. Hence my self-attribution of belief becomes self-fulfilling, even though it may originally have been formed from a confabulated interpretation of introspective and other data.

The upshot, then, is that there is no single event that is both introspectable, on the one hand, and is a judgment or decision, on the other. There are introspectable events that sometimes give rise to judgments and decisions (items of inner speech, or other forms of imagery); but these aren’t, themselves, the judgments and decisions. And there are, of course, such things as judgments and decisions; but these aren’t introspectable.

4.2 *Purely propositional thinking*

Some people might allow that neither inner speech nor visual imagery is constitutive of acts of judging and deciding. Hence the introspectability of inner speech and other forms of imagery does nothing to support the introspectability of judgments and decisions. But they might insist that thoughts can also occur to us wordlessly, and in the absence of visual imagery, whilst also insisting that such purely propositional thoughts can be introspected. For how else, after all, can we be supposed to know of their occurrence? Siewert (1998), for example, describes a case in which he was standing in front of his apartment door having just inserted his hand into his pocket where he normally keeps his key, finding it empty. Although he neither verbalized nor visualized anything at the time, at that moment he was (he says) wondering where the key could be. And his

knowledge of this act of wondering was (he says) immediate, resulting from introspection.

There is no doubt at all that many people *believe* that they have, and can introspect, purely propositional thoughts. (What is at issue is whether these latter beliefs are *true*.) Systematic evidence is provided by Hurlbert (1990, 1993). He asked subjects to wear headphones during the course of their normal daily lives, through which they would hear a beep at randomly generated intervals. They were instructed that, on hearing a beep, they should immediately introspect and take notice of what was passing through their consciousness at the time, making a brief written note of it to be elaborated in a later follow-up interview. What Hurlbert found is that all normal (as opposed to schizophrenic) subjects reported at least some instances of inner speech (ranging from 10 percent of occasions sampled to 80 percent, with the average being over 50 percent), and that most subjects also reported some occurrences of visual imagery and emotional feelings. In addition, although comparatively rare, some subjects reported instances of purely propositional, wordless and non-imagistic, thinking. For example, one subject reported that at the time of the beep she was standing in a supermarket looking at a box of breakfast cereal on the shelf in front of her. She said that she was aware of wondering—wordlessly—whether to buy the box, while also thinking—again wordlessly—that she didn't normally eat breakfast, and that the expense was therefore very likely to be wasted (Hurlbert, 1993, p.94).

Such examples provide no real evidence in support of introspection of judgments and decisions, however. For their existence is exactly what a defender of the anti-introspectionist position would predict. This is because, in cases where someone is doing something without concurrent verbalization or imagery, the mindreading faculty will nevertheless set to work attributing judgments and decisions where possible. And because the mindreading faculty doesn't model its own interpretative activity, it will seem to subjects that the judgments and decisions thereby attributed are known immediately, by introspection (provided that the interpretation process occurs smoothly and swiftly). Indeed, the crucial point to note about the examples above (and others like them) is that the thoughts attributed are exactly those that a third-party observer with the same background knowledge might ascribe. Thus anyone seeing Siewert standing in front of his door fumbling in his pocket, knowing that it is the pocket in which he normally keeps his key but that the pocket is empty, might predict that he is wondering where the key might be (especially if the observer also knows that Siewert has just begun to feel

anxious, as he reports that he had). And anyone seeing the lady standing looking at a box of cereal on a supermarket shelf might predict that she is wondering whether to purchase it. And if they also knew that she doesn't normally eat breakfast at all, and is cautious about spending money, then they might predict that the corresponding thoughts would also be occurring to her.

It should also be stressed that it is one thing to argue for introspection of purely propositional, un verbalized, thought (this is something that my sort of anti-introspectionist will deny), and it is quite another thing to argue for introspection of the *contents* of thoughts that *are* verbalized. (Siewert, 1998, defends both—the former partly via the latter—and Pitt, 2004, defends the latter at length.) For introspection of the contents of inner (and outer) speech is just what the model of the mindreading faculty outlined in Section 2 would predict.

Recall the example of seeing a dog chasing a ball. Various conceptual systems get to work on the perceptual input, classifying what is seen as a dog, and as chasing. And the results are globally broadcast as part of the perceptual state itself, being made available as input, *inter alia*, to the mindreading faculty. Hence we can immediately self-attribute that we are seeing a dog chasing a ball, without engaging in self-interpretation. Likewise in the case of speech (both overt and inner): the language comprehension system gets to work on the auditory input, interpreting it and attaching a content. The latter is globally broadcast along with the representation of the sounds heard. Hence we can introspect, not just the phonology of inner speech, but also its content or meaning. But of course this does nothing to show that we can introspect the judgments or decisions that might thereby be expressed or caused. (And indeed, Pitt, 2004, is quite explicit that he has done nothing to defend the introspectability of *attitudes*, as opposed to the introspectability of thought *contents*.)

4.3 *Violence to intuitions*

Despite everything that has been said so far, the denial of introspection for judgments and decisions will no doubt strike many readers as hugely counter-intuitive. Consider an everyday example. While working in my study I might take a decision to get up and open a window once I have completed writing the current paragraph; and then a few minutes later I do just that. I have the powerful impression that I have immediate, introspective, knowledge of my decision. And even if imagery of some sort is involved, I am strongly inclined to think that I have knowledge of my decision without having to consider and draw inferences from that together with the context,

other recent items of imagery, and so forth.

Granted, these intuitions seem powerful, and are hard to eradicate. But that is just what we would predict if the mindreading faculty doesn't model its own interpretative activity in relation to oneself, but rather pictures the mind's access to itself as essentially Cartesian (see Carruthers, 2008b). And that we aren't *aware* of engaging in self-interpretation doesn't begin to show, of course, that we don't do it. For our hypothesis is that the interpretative processes in which the mindreading faculty engages are for the most part unconscious ones. Whether we are attributing mental states to others or to ourselves, the process of interpretation is generally swift and unconscious. In our daily interactions with other people we mostly just find ourselves with beliefs about their intentions and other attitudes, without awareness of how we got to them. Yet everyone now accepts that interpretation of one sort or another has taken place. It is only in cases of special difficulty that interpretation slows down and becomes explicit, and we become aware that we are doing it. So it is, too, I submit, in our own case.

A final puzzle remains. For how is it that our attributions of judgments and decisions to ourselves are so *reliable* in comparison to the attributions that we make to other people, if those attributions are equally interpretative? There are two parts to my answer. One is that we almost always have a great deal more evidence available to us when we interpret ourselves than when we interpret others. In addition to observations of overt behavior and circumstances, we also have access to patterns within our own perceptual attention, to visual and other imagery (including inner speech), to bodily sensations and emotional feelings, and to proprioceptive information.¹⁵ We also, of course, characteristically have access to a great deal more episodic and semantic information about ourselves (whether or not that information gets expressed consciously).

The second part of my answer has already been sketched in Section 4.1. If I interpret myself as having formed a judgment that *P*, or as having decided to do *Q*, then I thereby commit

¹⁵ What explains the mindreading system's capacity to make use of attentional and proprioceptive information, if that system evolved initially for reading the minds of other people? For surely such information wouldn't have been needed for that. Two points are apposite. The first is that by virtue of being globally broadcast, such information would be *available* to mindreading (as well as to all other conceptual systems); and then provided that the latter has the requisite conceptual resources, it should be able to make use of it. The second point is that the mindreading system might have come under secondary selection once it began to operate as an interpreter of the self, as a number of cognitive scientists have suggested (Gazzaniga, 2000; Wilson, 2002).

myself to thinking and acting in the future on the assumption that I believe that *P*, or that I intend to do *Q*. (And note this will be true even if the initial interpretations had been inaccurate in themselves.) My self-attributions thereby become self-fulfilling. We frequently mold our own behavior to conform to our own interpretation of ourselves, in a way that we cannot (of course) influence the behavior of other people (Velleman, 1989; McGeer, 1996). But the fact that we do this doesn't give us introspective access to our own judgments and decisions.

5 Conclusion

The argument of Section 3 survives, then, and the model presented in Section 2 is vindicated. We have good reason to think that there is no such thing as introspecting a (non-perceptual) judgment, or introspecting a decision. On the contrary, all access to our own judgments and decisions is a matter of interpretation, just as it is when we access the judgments and decisions of other people. And it follows, therefore (as we pointed out in Section 1), that if mental states have to be introspectable in order to count as conscious, then there are no such things as *conscious* judgments or *conscious* decisions, either. But this is for theorists of consciousness to adjudicate.¹⁶

References

- Aiello, L. and Wheeler, P. (1995). The expensive tissue hypothesis. *Current Anthropology*, 36, 199-221.
- Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.
- Baars, B. (1997). *In the Theatre of Consciousness*. Oxford University Press.
- Baars, B. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Science*, 6, 47-52.
- Baars, B. (2003). How brain reveals mind: neuroimaging supports the central role of conscious experience. *Journal of Consciousness Studies*, 10, 100-114.
- Baars, B., Ramsoy, T., and Laureys, S. (2003). Brain, consciousness, and the observing self. *Trends in Neurosciences*, 26, 671-675.

¹⁶ I am grateful to Keith Frankish, Michael Gazzaniga, Shaun Nichols, Georges Rey, Elizabeth Schechter, and Stephen Stich, as well as to an anonymous referee for this journal, for their insightful comments on earlier drafts of this paper.

- Baron-Cohen, S. (1995). *Mindblindness*. MIT Press.
- Barrett, L., Dunbar, R., and Lycett, J. (2002). *Human Evolutionary Psychology*. Princeton University Press.
- Bem, D. (1967). Self-perception: an alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 74, 183-200.
- Bem, D. (1972). Self-perception theory. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, Volume 6, Academic Press.
- Block, N. (2005). Review of Alva Noë, *Action in Perception*. *The Journal of Philosophy*, 102, 259-272.
- Brasil-Neto, J., Pascual-Leone, A., Valls-Solé, J., Cohen, L., and Hallett, M. (1992). Focal transcranial magnetic stimulation and response bias in a forced choice task. *Journal of Neurology, Neurosurgery, and Psychiatry*, 55, 964-966.
- Bratman, M. (1987). *Intentions, Plans, and Practical Reason*. Harvard University Press.
- Bratman, M. (1999). *Faces of Intention: selected essays on intention and agency*. Cambridge University Press.
- Briñol, P. and Petty, R. (2003). Overt head movements and persuasion: a self-validation analysis. *Journal of Personality and Social Psychology*, 84, 1123-1139.
- Byrne, R. and Whiten, A., eds. (1988). *Machiavellian Intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press.
- Byrne, R. and Whiten, A., eds. (1997). *Machiavellian Intelligence II: extensions and evaluations*. Cambridge University Press.
- Carruthers, P. (2000). *Phenomenal Consciousness: a naturalistic theory*. Cambridge University Press.
- Carruthers, P. (2005). *Consciousness: essays from a higher-order perspective*. Oxford University Press.
- Carruthers, P. (2006). *The Architecture of the Mind: massive modularity and the flexibility of thought*. Oxford University Press.
- Carruthers, P. (2008a). Meta-cognition in animals: a skeptical look. *Mind and Language*, 23, 58-89.
- Carruthers, P. (2008b). Cartesian epistemology: is the theory of the self-transparent mind innate? *Journal of Consciousness Studies*, 15, 4 (2008), 28-53.

- Cohen, A. (1962). An experiment on small rewards for discrepant compliance and attitude change. In J. Brehm and A. Cohen (eds.), *Explorations in Cognitive Dissonance*, Wiley.
- Damasio, A. (1994). *Descartes' Error: emotion, reason and the human brain*. Papermac.
- Dehaene, S. and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79, 1-37.
- Dehaene, S., Naccache, L., Cohen, L., Bihan, D., Mangin, J., Poline, J., and Riviere, D. (2001). Cerebral mechanisms of word priming and unconscious repetition masking. *Nature Neuroscience*, 4, 752-758.
- Dehaene, S., Sergent, C., and Changeux, J. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Science*, 100, 8520-8525.
- Dutton, D. and Aron, A. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology*, 30, 510-517.
- Eagly, A. and Chaiken, S. (1993). *The Psychology of Attitudes*. Harcourt Brace Jovanovich.
- Edwards, G. (1965). Post-hypnotic amnesia and post-hypnotic effect. *British Journal of Psychiatry*, 111, 316-325.
- Evans, J. and Over, D. (1996). *Rationality and Reasoning*. Psychology Press.
- Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.
- Frankish, K. (2004). *Mind and Supermind*. Cambridge University Press.
- Frith, U. and Frith, C. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London, B*, 358, 459-473.
- Gazzaniga, M. (1995). Consciousness and the cerebral hemispheres. In M. Gazzaniga (ed.), *The Cognitive Neurosciences*, MIT Press.
- Gazzaniga, M. (2000). Cerebral specialization and inter-hemispheric communication: does the corpus callosum enable the human condition? *Brain*, 123, 1293-1326.
- Gertler, B. ed. (2003). *Privileged Knowledge: philosophical accounts of self-knowledge*. Ashgate Press.
- Goldman, A. (2006). *Simulating Minds: the philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.
- Gopnik, A. (1993). The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16, 1-14.

- Gopnik, A. and Meltzoff, A. (1997). *Words, Thoughts, and Theories*. MIT Press.
- Gregory, R. (1978). *Eye and Brain: the psychology of seeing*. McGraw-Hill.
- Hirstein, W. (2005). *Brain Fiction: self-deception and the riddle of confabulation*. MIT Press.
- Hurlburt, R. (1990). *Sampling Normal and Schizophrenic Inner Experience*. Plenum Press.
- Hurlburt, R. (1993). *Sampling Inner Experience with Disturbed Affect*. Plenum Press.
- Hurley, S. (1998). *Consciousness in Action*. Harvard University Press.
- Kahneman, D. (2002). Maps of bounded rationality: a perspective on intuitive judgment and choice. Nobel laureate acceptance speech. Available at:
<http://nobelprize.org/economics/laureates/2002/kahneman-lecture.html>
- Kosslyn, S. (1994). *Image and Brain*. MIT Press.
- Kosslyn, S., Thompson, W., and Ganis, G. (2006). *The Case for Mental Imagery*. Oxford University Press.
- Kreiman, G., Fried, I., and Koch, C. (2003). Single neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Science*, 99, 8378-8383.
- Kruglanski, A., Alon, W., and Lewis, T. (1972). Retrospective misattribution and task enjoyment. *Journal of Experimental Social Psychology*, 8, 493-501.
- Linder, D., Cooper, J., and Jones, E. (1967). Decision freedom as a determinant of the role of incentive magnitude in attitude change. *Journal of Personality and Social Psychology*, 6, 245-254.
- Locke, J. (1690). *An Essay Concerning Human Understanding*. Many editions now available.
- Lycan, W. (1987). *Consciousness*. MIT Press.
- Lycan, W. (1996). *Consciousness and Experience*. MIT Press.
- McDowell, J. (1994). *Mind and World*. Harvard University Press.
- McGeer, V. (1996). Is “self-knowledge” an empirical problem? Renegotiating the space of philosophical explanation. *The Journal of Philosophy*, 93, 483-515.
- Nichols, S. and Stich, S. (2003). *Mindreading: an integrated account of pretence, self-awareness, and understanding other minds*. Oxford University Press.
- Nisbett, R. and Wilson, T. (1977). Telling more than we can know. *Psychological Review*, 84, 231-295.
- Noë, A. (2004). *Action in Perception*. MIT Press.

- Palmer, S. (1999). *Vision Science: from photons to phenomenology*. MIT Press.
- Paulescu, E., Frith, D., and Frackowiak, R. (1993). The neural correlates of the verbal component of working memory. *Nature*, 362, 342-345.
- Pitt, D. (2004). The phenomenology of cognition, or, what is it like to think that p? *Philosophy and Phenomenological Research*, 64, 1-36.
- Rosenthal, D. (2005). *Consciousness and Mind*. Oxford University Press.
- Scholl, B. and Leslie, A. (1999). Modularity, development, and “theory of mind”. *Mind and Language*, 14, 131-55.
- Sheehan, P. and Orne, M. (1968). Some comments on the nature of post-hypnotic behavior. *Journal of Nervous and Mental Disease*, 146, 209-220.
- Shergill, S., Brammer, M., Fukuda, R., Bullmore, E., Amaro, E., Murray, R., and McGuire, P. (2002). Modulation of activity in temporal cortex during generation of inner speech. *Human Brain Mapping*, 16, 219-27.
- Siewert, C. (1998). *The Significance of Consciousness*. Princeton University Press.
- Snowdon, P. (1990). The Objects of Perceptual Experience. *Proceedings of the Aristotelian Society Supplementary Volume*, 64, 121-150.
- Stanovich, K. (1999). *Who is Rational? Studies of individual differences in reasoning*. Lawrence Erlbaum.
- Tye, M. (1995). *Ten Problems of Consciousness*. MIT Press.
- Tye, M. (2000). *Consciousness, Color, and Content*. MIT Press.
- Velleman, D. (1989). *Practical Reflection*. Princeton University Press.
- Wegner, D. (2002). *The Illusion of Conscious Will*. MIT Press.
- Wegner, D. and Wheatley, T. (1999). Apparent mental causation: sources of the experience of the will. *American Psychologist*, 54, 480-491.
- Wellman, H. (1990). *The Child's Theory of Mind*. MIT Press.
- Wells, G. and Petty, R. (1980). The effects of overt head movements on persuasion: compatibility and incompatibility of responses. *Basic and Applied Social Psychology*, 1, 219-230.
- Wicklund, R. and Brehm, J. (1976). *Perspectives on Cognitive Dissonance*. Lawrence Erlbaum.
- Wilson, T. (2002). *Strangers to Ourselves*. Harvard University Press.
- Wright, C. (2000). Self-knowledge: the Wittgensteinian legacy. In Wright et al. (2000).

Wright, C., Smith, B., and Macdonald, C. eds. (2000). *Knowing Our Own Minds*. Oxford University Press.

Figure 1: The place of mindreading in the mind

