Synthese (2007) 159:197–213
DOI 10.1007/s11229-007-9204-7

The illusion of conscious will

Peter Carruthers

Published online: 28 July 2007

Abstract Wegner (Wegner, D. (2002). *The illusion of conscious will*. MIT Press) argues that conscious will is an illusion, citing a wide range of empirical evidence. I shall begin by surveying some of his arguments. Many are unsuccessful. But one—an argument from the ubiquity of self-interpretation—is more promising. Yet is suffers from an obvious lacuna, offered by so-called 'dual process' theories of reasoning and decision making (Evans, J., & Over, D. (1996). *Rationality and reasoning*. Psychology Press; Stanovich, K. (1999). *Who is rational? Studies of individual differences in reasoning*. Lawrence Erlbaum; Frankish, K. (2004). *Mind and supermind*. Cambridge University Press). I shall argue that this lacuna can be filled by a plausible a priori claim about the causal role of anything deserving to be called 'a will.' The result is that there is no such thing as conscious willing: conscious will is, indeed, an illusion.

Keywords Confabulation · Conscious thought · Conscious will · Dual systems · Self interpretation · Wegner

1 Wegner on conscious will

Wegner (2002) argues for the illusoriness of conscious will across a broad front, and presents a wide array of evidence. But he has a number of different illusions in mind at different points in his discussion, not all of which, arguably, really constitute an attack on conscious willing. One is the traditional idea of a *metaphysically free* will, as an uncaused cause of action. But on this he has nothing to add to traditional philosophical critiques, I think (e.g. Dennett 1984). And it is very doubtful whether our ordinary

P. Carruthers (✉)
Department of Philosophy, University of Maryland,
1125 Skinner Building, College Park, MD 20742, USA
e-mail: pcarruth@umd.edu

idea of a conscious will requires that it should be metaphysically free in the intended sense. On the contrary, all that is required is that our conscious decisions should cause our actions, not that those decisions should themselves be uncaused causes.

Another of Wegner's targets is the idea that we have direct awareness of the *causality* of conscious will. The claim under attack here is that we have non-inferential and immediate awareness of the causal relationship between our conscious acts of deciding or intending and the actions that result. So even if our conscious decisions do cause our actions, on this account conscious will would still be an illusion if we weren't immediately aware of the *causing*-relation.

Wegner makes out a plausible case that there is no such immediate awareness. In part he tries to do this by drawing on Hume's (1739) claim that we never perceive causal relations themselves, not even when one billiard ball bounces into another, causing it to move. But this claim is eminently deniable for anyone who thinks that perception can be to some degree theory-laden. Given suitable background theories of the causal processes involved (some of which are, very likely, innate—see Baillargeon et al. 1995; Spelke et al. 1995), we surely *can* perceive the causal efficacy of some physical events. And if physical, why not mental too? Wegner is on stronger ground when he points out that the causal connections between our conscious thoughts and our actions are too complex in nature, and too variable from case to case, to count as perceivable.

Even if Wegner's argument on this point were made out, however, it is doubtful that the idea of conscious will would thereby be undermined. Granted, there would be no such thing as conscious will if our conscious thoughts were never, or only very rarely, the causes of our actions. For causal efficacy seems built into the very idea of what a *will* is. But provided that these conscious thoughts are often enough amongst the causes of our actions, then I think that the reality of conscious will needn't be undermined by our lack of awareness of the causing-relation. For these events would, by virtue of their frequently-causal status, constitute the operations of a *will*. And, by hypothesis, they are conscious. So there *would* be such a thing as conscious willing, even if we aren't conscious of the causal efficacy of our acts of willing.

Yet another idea that Wegner defends is that, very often, the conscious events that we take to be the operations of our conscious will, causing our actions, aren't really doing so. Rather, both those conscious events and the actions that we subsequently perform have a common cause in some set of *un*conscious mental events. These events cause both our conscious 'willings' *and* our actions. In which case that which we are aware of isn't what causes our actions. For example, an intention activated in our practical reasoning system might cause *both* the action intended *and* an appropriate verbalization of the intention ('I shall go *that* way'). The latter is mentally rehearsed and globally broadcast (in the sense of Baars 1988, 1997), thereby becoming conscious. Although it is easy for us to mistake this conscious event for the act of willing itself, in reality it doesn't play any causal role in the production of the action (or at least, not always—I shall return to this qualification at length in Sect. 4). Rather, there is a common underlying cause—the activated intention—which causes both the action and the conscious verbalization of the intention to act.

Unless the present argument collapses into the previous one, however, it would have to be claimed that our conscious 'willings' and our actions are *always* the effects

of a common underlying cause, and not just that they sometimes are. For otherwise, in those cases where our conscious thoughts *do* cause our actions, there *would* be such a thing as conscious willing. But Wegner doesn't really do anything to defend this stronger claim, I think, unless via the idea that I shall discuss next. And that strong claim is surely inconsistent with the reality of 'System 2' reasoning and decision-making, defended by a number of authors (Evans and Over 1996; Frankish 2004), in which conscious verbalizations of intention do play a causal role. I shall return to this point in due course (in Sect. 4).

The final claim that Wegner makes—which I propose to spend a fair bit time in Sect. 3 elaborating and defending—is that our access to our own will (that is, our own acts of deciding or intending, which cause our actions) is always *interpretative* in character. Our awareness of our own will results from turning our mind-reading capacities upon ourselves, and coming up with the best interpretation of the information that is available to it—where this information doesn't include our acts of deciding themselves, but only the causes and effects of those events. On this account, then, the events in question don't count as *conscious*, or at least not by the intuitive criterion that requires us to have non-inferential, non-interpretative, access to a mental event if the latter is to qualify as a conscious one. In the next section I shall present a brief defense of this latter claim, before turning to discuss the main point in Sect. 3.

2 The immediacy of consciousness

There is, of course, a great deal of dispute about the nature of consciousness. By far the major part of this concerns the nature and best explanation of *phenomenal*, or experiential, consciousness. This is a topic on which I have well-developed views of my own (Carruthers 2000, 2005). But it isn't one that concerns us here. For our topic isn't conscious experience, but rather conscious thought: and in particular, conscious acts of *deciding* or *intending*. And here it is arguable that our common-sense understanding of what it is for a thought to be a conscious one is that it is a mental event *of which we are immediately and non-interpretationally aware* (Carruthers 2005, chapters 7 and 8).[1] It is this that gives us the contrast between our conscious thoughts and our unconscious ones. For the latter, if they are known to us, are only known via a process of self-interpretation, such as might occur during counseling or on the psychoanalyst's couch. And it is also what gives us the asymmetry between our own conscious thoughts and those of other people. The latter, too, are only known to us through a process of interpretation, whereas our own thoughts, when conscious, aren't, we believe.

This conclusion can be reached from a variety of different perspectives on the nature of conscious thought. Many philosophers—especially those influenced directly or indirectly by Wittgenstein—are apt to emphasize that we are *authoritative* about our own conscious thoughts, in a way that we cannot be authoritative about the thoughts of others (Malcolm 1984; Shoemaker 1990; Heal 1994; Moran 2001; see also Burge 1996). If I sincerely claim to be entertaining a particular thought then this provides sufficient

[1] Strictly speaking, all that the argument of this paper requires is that non-interpretational access is a *necessary condition* on conscious thoughts and intentions.

grounds for others to say of me—and to say with justification—that I *am* entertaining that thought, in the absence of direct evidence to the contrary. Put otherwise, a sincere avowal of what I am currently thinking is *self-licensing*—perhaps because such claims are thought to be somehow constitutive of the thoughts thereby ascribed—in a way that sincere claims about the thought processes of others aren't.

It is very hard indeed to see how we could possess this kind of epistemic authority in respect of our own occurrent thoughts, if those thoughts were only known to us on the basis of some sort of self-interpretation. For there is nothing privileged about my standpoint as an interpreter of myself. Others, arguably, have essentially the same kind of interpretative access to my mental states as I do. (Of course I shall normally have available a good deal more data to interpret in my own case, and I can also generate further data at will, in a way that I can't in connection with others—but this is a mere quantitative, rather than a qualitative difference.) So believers in first-person authority should also accept the immediacy of conscious thought, and maintain that our access to our own occurrent thoughts, when conscious, is of a non-inferential, non-interpretative, sort.

The immediacy of consciousness can also be motivated from a variety of more cognitivist perspectives. On the sort of approach that I favor, a mental state becomes conscious when it is made available to a 'mind-reading' faculty that has the power, not only to entertain thoughts about the content of that state (e.g. about an item in the world, perceptually or conceptually represented), but also to entertain thoughts about the *occurrence* of that state itself (Carruthers 2000, 2005). When I perceive a ripe tomato, for example, my perceptual state occurs in such a way as to make its content available to conceptual thought about the tomato, where some of those concepts may be deployed recognitionally (e.g. *red* or *tomato*). That state is then a conscious one, if it also occurs in such a way as to be available to thoughts about *itself* (e.g. 'It looks to me like there is a tomato there,' or 'I am now experiencing red')—where here, too, some of the concepts may be deployed recognitionally, so that I can judge, straight off, that I am experiencing red, say. On this sort of account, then, a conscious thought will be one that is available to thoughts about the occurrence of that thought (e.g. 'Why did I think *that*?'), where the sense of *availability* in question is supposed to be non-interpretative, but rather recognitional, or at least quasi-recognitional.

It is worth noting that this account is fully consistent with so-called 'theory-theory' approaches to our understanding of mental states and events (a version of which I endorse). On such a view, our various mental concepts (*perceive, judge, fear, feel*, *intend*, and so on) get their life and significance from their embedding in a substantive, more-or-less explicit, *theory* of the causal structure of the mind (Lewis 1966; Churchland 1981; Stich 1983; Fodor 1987). So to grasp the concept *percept of red*, for example, one has to know enough about the role of the corresponding state in our overall mental economy, such as that it tends to be caused by the presence of something red in one's line of vision, and tends to cause one to believe, in turn, that one is confronted with something red, and so on. It is perfectly consistent with such a view, that these theoretical concepts should also admit of recognitional applications, in certain circumstances. And then one way of endorsing the immediacy of consciousness is to say that a mental state counts as conscious only if it is available to a recognitional application of some corresponding mental concept.

The immediacy of consciousness can likewise be endorsed by those who believe that consciousness results from the operations of an internal self-scanning mechanism, or 'inner sense' (Armstrong 1968, 1984; Lycan 1987, 1996), provided that the mechanism in question is encapsulated from our beliefs about ourselves and our present circumstances. All perception is to some degree inferential in character, of course (Pylyshyn 2003), and presumably inner perception is no different. But this wouldn't show that our access to our own mental states isn't immediate, in the sense that matters for our purposes. For the inferences in question can depend only on information that is *general*, not accessing our beliefs about our other beliefs and goals, about our recent movements and physical actions, or about events taking place in our current environment. And provided that this is so, then we still have a principled distinction between the sort of access that we have to our own mental states, and the sort of interpretative access that we have to the mental states of others. For the latter does require us to access beliefs about the other person's actions, as well as beliefs about that person's other beliefs, goals, and so forth.

What really *is* inconsistent with the immediacy of consciousness is a view of our relation to our own mental states that makes the latter dependent upon our *particular* beliefs about our current environment or circumstances, or about our recently prior thoughts or other mental states. If my awareness that I am in some particular mental state depends, not just on recognitional deployment of theoretically-embedded concepts, but also on inferences that draw upon my beliefs about the current physical or cognitive environment, then introspection really *will* be inferential and interpretative in a manner that conflicts with the relevant sort of immediacy. But it is, I claim, a presupposition of our common-sense conception of consciousness that our access to our conscious mental states is *not* interpretative in this sense. For that would mean that there was no principled difference between our access to our own thoughts and our access to the thoughts of other people.

There is a strong case for saying that our access to our own acts of will must be immediate and non-interpretative, then, if those acts are to qualify as conscious ones. At any rate this is what I propose to assume in what follows. I shall take for granted that if it can be shown that the only form of access that we have to our own intentions and decisions to act is interpretative—in this respect like the access that we have to the intentions and decisions of other people—then there is no such thing as conscious willing or conscious deciding. Conscious will would have been shown to be illusory.

3 Is our access to our own acts of will interpretative?

Wegner (2002) argues on empirical grounds that our access to our own intentions and acts of intention-formation is, indeed, interpretative, drawing on the extensive 'cognitive dissonance' and 'self-perception' literatures that have been built up in social psychology over the last 50 years (Festinger 1957; Bem 1967, 1972; Wicklund and Brehm 1976; Nisbett and Wilson 1977; Eagly and Chaiken 1993; Wilson 2002). In a wide range of circumstances people can be induced to *confabulate* explanations of their own behavior, attributing to themselves attitudes and intentions that they don't actually have (and they do so in all honesty, without awareness that they are confabulating).

In particular, they will ascribe immediately prior intentions to themselves that are manifestly at odds with their real intentions.

Wegner interprets these data as showing that the mind-reading system, or mental state attribution system, lacks any immediate access to the thought processes of the individual, including the processes involved in forming and acting on intentions. Rather, the system only has access in the first instance to perceptual input. It is focused outwards on the world rather than inwards on the agent's own mental states, and can only attribute such states to the agent by observing and interpreting the agent's own behavior. Gazzaniga (1998) and Carruthers (2005) give these ideas an evolutionary spin, linking them to the suggestion that mind-reading evolved in the first instance for purposes of interpreting and manipulating other people. (This is a version of the 'Machiavellian intelligence' hypothesis; see Byrne and Whiten 1988, 1997.) Wilson (2002) also suggests, however, that the use of the mind-reading system to interpret one's own behavior—building up a self-conception, and confabulating where necessary—may subsequently have become adaptive.

The resulting architecture can be seen depicted in Fig. 1. (The latter has been constructed in such a way as to accommodate the 'two visual systems' hypothesis of Milner and Goodale 1995.) The mind-reading system receives input from the visual system (and also from the other sensory systems, which aren't depicted). But it receives no input from the systems involved in reasoning and decision-making. (Note the lack of any arrows back to the mind-reading system from the belief and desire forming systems, or from practical reason.) It can, however, access memory systems of various kinds; so it can base its interpretations not only on the contents of current perception, but also on what has happened in the past (e.g. accessing a memory of the instructions given by the experimenter).

It might be objected that the mind-reading system would very likely need access to more than just memories produced by other systems, in order to do its job. Surely it would also need access to the current outputs of those systems. In order to interpret

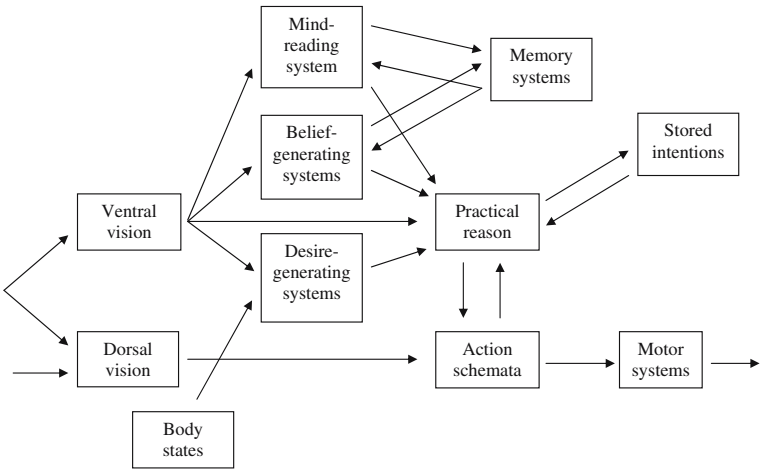

Fig. 1 A mind with an outwardly focused psychology faculty

the intention behind someone's observed physical movement, for example, it might be crucial to project forwards to the likely physical consequences of that movement—such as the trajectory that an item is likely to follow once thrown. This would require the mind-reading system to access the current output of the common-sense physics system.

A number of replies can be made to this objection, of varying strength. One would be to concede the point in respect of belief-generating systems. We could allow that the mind-reading system has routine access to current active beliefs and judgments (and perhaps also to current desires), while denying that it has such access to decision-making processes and current intentions. All that matters for our purposes is whether or not the mind-reading system has immediate (non-interpretative) access to the latter. And it is very hard to see why the mind-reading system would *need* any such access, in order to interpret the behavior of other people. It is easy to see why it might help, in reading the minds of others, if the mind-reading system could monitor the current outputs of all belief-generating systems, since some of these beliefs might be highly relevant to the interpretation of others' behavior. But it isn't at all obvious how it would be of any help to have access to the interpreter's own intentions.

A stronger reply would be to distinguish between the capacity to *query* belief-generating systems for information, and the routine monitoring and representing of the outputs of those systems to subserve immediate self-attributions of current belief and judgment. It is plausible that the mind-reading faculty should have the former capacity, for the reason presented above. But this gives us no reason to think that it should routinely monitor the outputs of the various belief-generating systems, irrespective of context. (AI is replete with systems that can query one another for information when executing their own tasks, for example, but which have no access to the reasoning processes that go on within those other systems, and which make no attempt to monitor and represent the outputs of those systems on a routine basis.) And we have even less reason to think that the mind-reading faculty should routinely monitor the subject's own practical decision-making and active intentions.

An even stronger reply to the objection above might be possible. It might be claimed that the mind-reading system can only have access to the outputs of other belief-generating systems and the practical reasoning system *indirectly*, insofar as those outputs are used to form visual or other images or mentally rehearsed sentences in 'inner speech'. The latter (being quasi-perceptual in character) are then 'globally broadcast' (in the sense of Baars 1988, 1997) and provided as input to a wide range of belief-generating and desire-generating systems, including the mind-reading system. This is a possibility that could be married quite nicely to both the 'two systems theory' of reasoning (see Sect. 4 below) and to accounts of mind-reading that find some role for *simulation* of the mental operations of other people.[2] But I shan't pursue this possibility any further here, since it is far stronger than I need, in order to make the point that a working mind-reading system is unlikely have access to current practical decisions and intentions.

[2] For a recent and well worked out development and defence of a limited version of this idea, see Nichols and Stich (2003).

Now, given that the mind-reading system has access to perceptual input (see Fig. 1), it will of course have access to any other kinds of mental event that utilize the same perceptual mechanisms, resulting in 'global broadcasting' of those events (when conscious). These will include visual and other forms of imagery as well as rehearsals of natural language sentences in 'inner speech'. So there will actually be a great deal more evidence available to the mind-reading system, when constructing an explanation of the agent's own behavior, than could be available to any outside observer. Indeed, in many cases the only 'behavior' forming the basis for a self-attributed thought or intention might be a sequence of conscious images, felt bodily reactions to emotion, and/or items of inner speech—none of which would be available to an outside observer. This seems sufficient to explain the much greater reliability of self-attribution over other-attribution, and can plausibly account for the fact that many ordinary folk regard people as authoritative in their attributions of thoughts to themselves.

It doesn't *follow from* the empirical data that the mind-reading system is embedded in the mind in the manner depicted in Fig. 1, of course. For access-relations amongst systems don't have to be infallible. So it might be consistent with the confabulation data that the mind-reading system *does* have routine access to many of the agent's own thought-processes, and to the agent's acts of intention-formation in particular. It might just be that the monitoring process is spotty and unreliable for some reason. The resulting 'self-monitoring' architecture is depicted in Fig. 2. (Note the arrows back to the mind-reading system from the belief-forming systems and from practical reason.) If this were correct, then there could be such a thing as conscious will after all, because decisions to act and activated intentions would at least *sometimes* be made available to the mind-reading system for self-attribution immediately, in a way that wouldn't require self-interpretation.

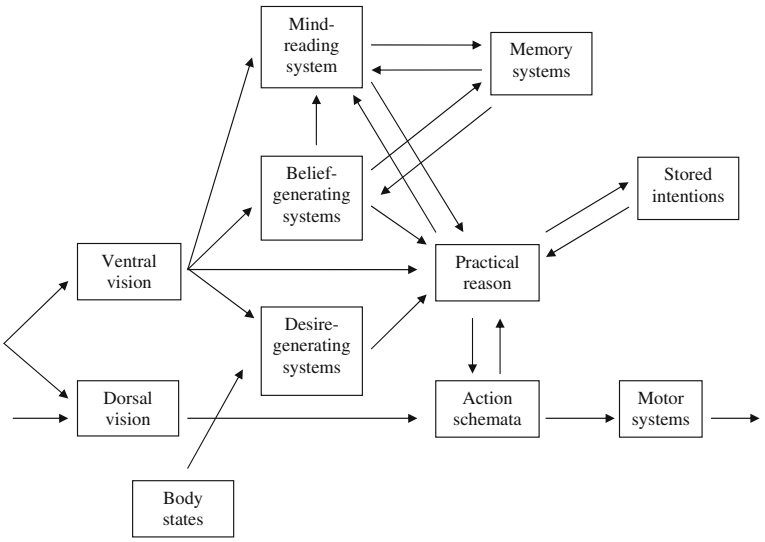

Fig. 2 A mind with a self-monitoring psychology faculty

There are two powerful objections to this alternative self-monitoring architecture, however. The first is that it doesn't seem capable of explaining, in anything other than an ad hoc way, the patterning of frequent failures of self-monitoring, as revealed in the extensive social-psychology data on confabulation. If the mind-reading faculty had the power to monitor the outputs of belief-generating systems and the practical reasoning system, then why is it that the self-attribution process should go awry in just the ways that it does? For what seems to hold true of the cases represented in the confabulation literature is either that the thoughts that really occur to people in the experimental circumstances are ones that would be surprising to common-sense psychology (hence some other explanation needs to be confabulated), or that there is a plausible but false common-sense explanation, which is then mis-attributed.[3]

This patterning in the confabulation data suggests that people are engaging in self-interpretation in all cases. And this suggestion is confirmed when we turn to look at perfectly ordinary cases where neither of the above conditions holds, but where the mind-reading system is denied access to some of the crucial data needed to construct an explanation. These occur in people who have undergone commissurotomy (separating the two halves of the brain), whenever data are presented to the right-brain alone (such as a card saying, 'Walk!') in such a way as to cause behavior (getting up and walking). In such cases the mind-reading system (which is located in the left-brain) will confabulate an explanation (such as, 'I was going to the fridge for a drink') with all the seeming-obviousness of any self-attributed thought or decision. (see Gazzaniga 1998, for discussion of many such examples.) At the very least it follows that the mind-reading system itself has no knowledge of when it does or doesn't have access to the agent's thoughts and intentions. And the most plausible explanation is that it actually has no such access.

The patterning in the data suggests that all active intentions and acts of intention-formation are self-attributed via a process of self-interpretation, then—which of course supports the Fig. 1 architecture against the self-monitoring architecture represented in Fig. 2. This is the first objection to the latter. The second objection is evolutionary. The resources necessary to monitor and represent the outputs of all belief-generating and desire-generating systems, as well as the operations and output of practical reason, would surely be considerable. In which case there would need to be some significant adaptive payoff in order to warrant the investment. What could this be? What would be the evolutionary benefit of building the architecture and links needed to monitor and represent our own thought-processes and reasoning-processes on a regular basis?

The only remotely plausible proposal that I know of is that the benefits of self-monitoring derive from the resulting improvements in the processes that are monitored (Shallice 1988; Smith et al. 2003). By monitoring our own reasoning we can trouble-shoot in cases of failure, watch out for mistakes and errors, and intervene to improve the quality of our own reasoning on an incremental basis. There are sound reasons for rejecting this proposal, however (unless it is married together with a version of two

[3] For example, that people choosing amongst what are actually identical items should have a tendency to select from the right hand side of the array falls into the first category; and that people should enjoy a game less if they are paid to participate in it falls into the second. See Wegner (2002) and Wilson (2002) for discussion and references.

systems theory—see Sect. 4). For there are good reasons to doubt whether we have any natural competence in the domain of improvement of reasoning.

In the case of mind-reading, we have a highly developed competence with obvious adaptive significance. Human beings are quite remarkably good at attributing mental states to others (and derivatively to themselves), and hence at engaging in the sorts of manipulative and/or co-operative behaviors whose success depends upon those attributions being accurate. And everyone agrees that this competence is part of the natural endowment of any normally-developing human being. In contrast, it is far from clear that we possess any very useful competence in the domain of self-monitoring. Of course we are good at *attributing* mental states to ourselves (absent the confabulation cases), but for a mind-reading theorist this is just a matter of our turning our mind-reading abilities upon ourselves. What is in question is whether we have the sort of highly-developed capacity to monitor, trouble-shoot, intervene in, and improve upon our own reasoning processes on-line, in the way that the self-monitoring model requires.

There is little reason to think that we possess any such natural competence. Indeed, naïve subjects are quite remarkably poor at distinguishing good sequences of reasoning from bad ones, or at fixing up the latter to make them better (Kahneman et al. 1982; Piattelli-Palmarini 1994). And insofar as people do have any such capacity, it only emerges late in development: not until late childhood or early adolescence as a result of formal schooling (Pillow 2002; Moshman 2004) This isn't to say that naïve subjects are bad *at reasoning*, of course. For each of our first-order information-generating systems will have been under selection pressure for speed and reliability. And moreover, some of the heuristic reasoning processes that people employ turn out to be quite remarkably successful (Gigerenzer et al. 1999). Rather, it is to say that naïve subjects are bad at reasoning *about reasoning*—at identifying mistakes in reasoning, at theorizing about standards of good reasoning, and at improving their own and others' reasoning. Yet this is precisely the competence that the self-monitoring model predicts we should have.

One of the defining features of human civilizations, in fact, is that they contain socially-transmitted bodies of belief about the ways in which one *should* reason. And these bodies of belief have to be laboriously acquired through processes of formal education. (They include the canons of good scientific method, developed piecemeal over the last five centuries or so, as well as principles of validity, codified in systems of logic.) Insofar as we have any competence in evaluating and improving reasoning, therefore, this isn't a *natural* competence, but a socially transmitted one. Hence we have no reason to think that the architecture of our cognition is as the self-monitoring model claims, or as Fig. 2 depicts. And hence we have no reason to think that there is any such thing as conscious (immediately accessible) practical reasoning or intention-formation.

It might be objected that self-monitoring can be undertaken for purposes other than the imposition of correct normative standards of reasoning. Indeed, aren't cognitive science and AI rife with postulated mechanisms that monitor whether tasks have been completed, whether the output of some process matches what was expected, and so forth? This is quite true, but irrelevant. For the sort of monitoring that is in question in the present section is monitoring that issues in higher-order thoughts—that is to say,

thoughts that are explicitly *about* our first-order thoughts, beliefs, and intentions, as such, as generated by our inferential and decision-making systems. The sorts of self-monitoring processes just described, in contrast, can all be realized without subjects having the capacity to attribute mental states to themselves at all.

Monitoring whether some cognitive task has been completed can just mean having some mechanism that is sensitive to whether or not the system has generated an output. This needn't require that the mechanism in question should have the conceptual resources to self-attribute that output. Likewise for a process that matches the output of a system against some previously generated template or expectation. And monitoring of bodily reactions (such as often occurs during decision making, if Damasio 1994, is correct) is a first-order perceptual process, rather than a higher-order one.

I think that we can set the self-monitoring model of self-knowledge of our occurrent thoughts and intentions a dilemma, then. Insofar as we have any reason to believe that something like self-monitoring occurs, then this is either monitoring that doesn't really require any capacity for higher-order thought, or it does require such a capacity, but operates on globally broadcast perceptual or quasi-perceptual items (visual images, rehearsals of natural language sentences in 'inner speech,' and so on). And yet the latter sort of monitoring is fully explicable from the perspective of the mind-reading model, according to which the mind-reading faculty is focused 'outwards,' via its access to the outputs of the perceptual systems. Hence we have no reason to think that the architecture of our cognition is as the self-monitoring model claims, or as Fig. 2 depicts.

4 Two systems theory to the rescue?

At this point it appears that one of Wegner's (2002) arguments for the illusoriness of conscious will is quite compelling. For in order to count as conscious, active intentions and formations of intention would have to be accessible to subjects in a way that isn't interpretative, but rather immediate and recognitional. Yet the extensive social-psychology literature on confabulation suggests that our attributions of intention to ourselves *does* result from our mind-reading system constructing the best interpretation that it can on the basis of the data available to it (perceived actions, proprioceptive feedback from movement, visual imagery, inner speech, and so on). There is, however, an obvious lacuna in the argument as developed thus far. This is provided by so-called 'two systems theories' of reasoning and decision making (Evans and Over 1996; Stanovich 1999; Kahneman 2002; Frankish 2004). I shall first sketch the outlines of such theories, before showing how they might be deployed to block Wegner's argument.

Almost everyone now working on human reasoning and rationality, and our pervasive lapses of rationality, has converged on some or other version of two systems theory. System 1 consists of a large number of unconscious reasoning systems, many of which are evolutionarily quite ancient. These operate extremely swiftly in comparison with conscious thought, and many of them deploy 'quick and dirty' but reliable enough heuristics for reaching their conclusions. (Some but not all researchers believe that they are realized in association-based connectionist networks of various sorts;

Evans and Over 1996.) Moreover their processing is either unalterable, or at least impenetrable by conscious learning or control. They certainly aren't influenced by verbal instruction.

It is these System 1 processes that are responsible for the 'intuitions' that assail us (and often lead us astray) when confronted with reasoning or decision-making problems. Thus even those well-versed in probability theory will still feel some 'tug' towards saying that it is more likely that Linda is both a bank teller and a feminist than it is that she is just a bank teller, when presented with a scenario in which we are told that in her youth she went to college and was active in various feminist groups on campus—and this despite the fact that it is impossible for a conjunction to be more probable than one of its conjuncts.

System 2, in contrast, is conscious, and its operations are comparatively slow. It operates on learned rules and procedures, many of which are consciously applied. These procedures are all of them acquired via learning, and often they are acquired via verbal instruction from other people. Successful use of System 2 correlates highly with IQ, and also with various measures of 'cognitive style', such as a disposition to be reflective, and a capacity to distance oneself from one's own views (Stanovich 1999). Moreover, all theorists seem agreed that natural language sentences—either overtly uttered, or covertly rehearsed in 'inner speech'—play an important role in the operations of System 2, although not everyone thinks that System 2 is exclusively language based. (Some think that other sorts of imagery could play the mediating role, too.) But everyone agrees that these conscious events that constitute the operations of System 2 play a genuine causal role in the production of our behavior, either displacing or overriding the workings of System 1 on those occasions when it operates.

Thus we often 'talk our way through' difficult problems, explicitly applying learned rules in the course of our verbalizations. For example, someone faced with a version of the Wason selection task—who has been asked to determine which of four cards he should turn over in order to establish the truth of a conditional statement of the form 'P $\supset$ Q'—might pause to reflect, and repeat to himself (recalling something learned in a course of logic, perhaps), 'In order to evaluate a conditional I need to look for cases where the antecedent is true and the consequent is false; so I shall turn over the P-card and the not-Q-card.' And this concluding verbalization causes him to act in the manner described (hence answering the question correctly).

But *how* does an inner verbalization (or some other mentally rehearsed image) cause action? How can it supplant the operations of System 1, and lead to the behavior described or otherwise envisaged? Some theorists don't really address this question. But Frankish (2004), developing and extending the work of Cohen (1992), argues at length that what we think of as conscious decisions (such as saying to myself at the conclusion of an episode of reasoning, 'So I shall do Q') are actually a form of *commitment*, realized in the operations of System 1 processes. It is because I believe that I have decided to do Q, and have a standing desire to do the things that I decide to do, that my action is caused. So although the conscious episode of saying to myself, 'So I shall do Q' does cause my action of doing Q, it does so through the mediation of other beliefs and desires of mine. (This point will be important later.)

Now here is how the distinction between System 1 and System 2 might be deployed to challenge Wegner's argument. We can concede that System 1 decisions and intentions

are never conscious. Hence we can concede that *often* when we attribute intentions to ourselves it is on the basis of a swift piece of (unconscious) self-interpretation. But we can still maintain the reality of conscious will by aligning it with System 2. For System 2 events are undeniably conscious, and they often cause our subsequent behavior, with the content of these mental events matching the behavior executed. (I say to myself, 'I shall do Q', and I do Q.) So we can claim that these events *are* conscious willings—saying to myself, 'So I shall do Q' *is* a conscious making-up-of-mind to do Q; it is a conscious formation of an intention. So conscious will is a reality, after all.

5 Wegner's argument vindicated

Although promising, this defense of the reality of conscious will is flawed. For surely our conception of a decision to act, or our idea of what it is to form an intention to act, is the idea of an event that causes action either immediately, or through the operations of further reasoning processes that are purely first-order in nature (such as figuring out a sufficient means to the execution of that act). But the event of saying to myself, 'I shall do Q' doesn't have these properties. On the contrary, it only leads to action via processes of reasoning that are higher-order in character, including such events as *believing that I have decided to do Q*, and *wanting to do what I have decided*. In which case, while the act of saying to myself, 'I shall do Q' is conscious, and does play a causal role in the production of the behavior of doing Q, it doesn't have the causal role characteristic of a genuine decision to do Q. And so it turns out, after all, that there is still no such thing as conscious deciding.

Let me develop this argument in a little more detail. Consider, first, an intention or decision to do something that is to be done here-and-now, such as deciding to open a window for the breeze. We surely think that, in such a case, a genuine decision must be the last *deliberative* mental event in the causal chain that leads to the action. A genuine decision will be something that causes a motor schema to be activated, which is then guided and up-dated in the light of ongoing perceptual input. Hence a genuine decision to do something here-and-now needn't be the last *mental* state in the causation of the action. But once the decision is made, we think, there is no further role for the interaction of beliefs with desires in any sort of process of practical reasoning. Rather, a genuine decision, in these sorts of circumstances, should *settle* the matter. (It only settles it subject, of course, to there being no problems arising in the execution of the action—e.g. I find that my legs have 'gone to sleep', and that I can't walk—and subject to there being no unforeseen circumstances leading me to revise the original decision—e.g. I find that the window is swarming with biting ants.) But saying to myself, 'I shall do Q' *doesn't* settle the matter. It only results in an act of doing Q via further (unconscious) deliberation, given that I have further beliefs and desires of the right kind.

Now consider a decision that is taken for the more distant future. Often such intentions are 'incomplete', in the sense that they don't yet contain a full specification of the means to be taken in executing the decision; so some further reasoning needs to take place (Bratman 1987, 1999). For example, I decide to purchase a particular book after reading its description in the press' catalog. But this doesn't yet fix *how* I should

make the purchase—should I place an order on-line through Amazon, phone my local bookstore, or complete and post the order-slip in the catalog itself? So in such a case a decision *isn't* the last deliberative step in the causal chain that leads to action.

All the same, we still think that a decision in this sort of case should settle *what* I do (subject, of course, to the usual qualifications about unforeseen difficulties and changes of mind). It just doesn't settle *how* I do it. Put differently, while we think that a decision, if it is to genuinely count as such, can be followed by further deliberation, this should only be deliberation about the *means* to execute the action, not about the action itself. So if the act of buying a book is Q, the deliberation that follows a decision to do Q shouldn't be about whether or not to do Q (that should already have been settled), but merely about *how* to do Q in the circumstances.

In a case of System 2 decision-making, in contrast, the conscious event of saying to myself in inner speech, 'I shall do Q' *doesn't* settle that I do Q, and the further (unconscious) practical reasoning that takes place prior to action *is* about whether or not to do Q. For on the account of System 2 practical reasoning sketched above, the sentence, 'I shall do Q' first needs to be interpreted *as* a decision to do Q by the mind-reading faculty, and it only leads to the act of doing Q via my desire to do what I have decided, in combination with my belief that I have decided to do Q. So this is surely sufficient to disqualify the conscious event, here, from counting as a genuine decision, even though it does play a causal role in the production of the action. For the role in question isn't the right *sort* of role appropriate for a decision.

I can imagine two replies to the argument presented here. One is that the event of saying to myself, 'I shall do Q' might be the last mental event that occurs *at the System 2 level* (albeit with further practical reasoning taking place thereafter in System 1). So it *does* directly cause my action at that level of description, unmediated by any System 2 practical reasoning. But this isn't enough to qualify it as counting as a genuine decision, in my view. Our common-sense notion of a decision doesn't make any allowance for the System 1/System 2 distinction. It is the idea of an event that causes action without the mediation of any further reasoning about whether or not to act. And given this idea, saying to myself, 'I shall do Q' plainly doesn't qualify.

It might also (and relatedly) be replied that my development of Wegner's argument overlooks the fact that System 2 processes are supposed to be *realized in* the operations of System 1. My decision to do Q is a System 2 event that is realized in a set of System 1 events, included in which is the belief that I have made a decision and the desire to do what I have decided. So again, the System 2 event directly causes my action *at that level of description*. But this reply, too, is off the mark. It is true enough to say that System 2 processes in general are realized in those of System 1. But surely the realizing conditions for an event cannot occur subsequent to that event itself. And it is only once the conscious event of saying to myself, 'I shall do Q' is completed that the System 1 reasoning leading to the action kicks in. Moreover, if we opt to say that the decision to do Q isn't *that* event, but rather the more extended event that also includes the subsequent System 1 practical reasoning, then *that* event isn't a conscious one. So either way, there is no one event, here, that is both conscious and a decision.

I conclude, then, that Wegner (2002) is correct: conscious will is an illusion. Given that the mind-reading system has no direct access to events that take place in the practical reasoning system (as we argued in Sect. 4), but only to their globally broadcast

effects, then our access to those events is always interpretative. Hence those events within practical reason don't qualify as conscious ones. And even where a conscious event (such as saying to myself in inner speech, 'I shall do Q') does play a causal role in the production of the action, through System 2 processes, that role isn't of the sort that should be characteristic of a 'willing' event of deciding or intending. For it doesn't settle anything by itself. Only if I want to do what I have decided and *believe* that by saying to myself, 'I shall do Q' I *have* decided, does the action get settled upon.

6 Whence the illusion?

How does the illusion of conscious will arise in us, however? And why is it so persistent? This isn't so difficult to explain. For the mind-reading system only contains a limited, highly simplified, model of its own operations. The mind-reading system's model of the architecture of the mind as a whole is highly simplified, but nevertheless accurate enough for most everyday purposes of prediction and explanation (Carruthers 2006). Likewise, it is plausible that the mind-reading system's model of the sort of access that the mind has to itself and its own operations is highly simplified. That model appears to be quasi-Cartesian in character: it pictures the operations of the mind as *transparent* to itself, in such a way that the events within a mind are immediately known by the mind. This model can be modified as a result of experience and subsequent System 2 reasoning, no doubt, and many people nowadays end up allowing for the existence of unconscious mental processes, to some degree. But the default is transparency, and that default might actually be adaptive in various ways (Gazzaniga 1998; Wilson 2002).

So when the mind-reading system sets to work interpreting the inner verbalization of, 'I shall do Q,' classifying it as the formation of an intention to do Q, it doesn't represent the fact that it engages in such a process of interpretation. Nor does it have knowledge of the subsequent reasoning that takes place. Rather, its results are couched in terms of the simple transparency model. So its output—if required to report on its own operations—would be something along the lines of, 'I (the person) am immediately aware that I have just formed the intention to do Q.' But in fact there is no such immediate awareness, if the arguments outlined above are correct. There is immediate awareness of the inner verbalization, admittedly. But the inner verbalization isn't itself an intention to do Q, although it may play a causal role in the production of that action somewhat similar to that of an intention.

7 Conclusion

I have been arguing in defense of Wegner (2002) that conscious will is an illusion. There are three main premises of the argument. The first is that an act of intention-formation, to be conscious, must be available to the agent in a way that doesn't call for self-interpretation. The second is that the best explanation of the extensive confabulation literature is that the mind-reading faculty either doesn't have access to the outputs of practical reason at all, or at any rate doesn't have the capability to monitor and represent those outputs 'on-line' on a regular basis. On the contrary, those intentions

and decisions can only be attributed via self-interpretation. The third premise is that while some System 2 events are conscious (such as mental rehearsal of the sentence, 'I shall do Q'), and do play a causal role in the production of appropriate behavior, it isn't the right *sort* of causal role to constitute the event in question as an intention.

Some parts of this argument are plausibly a priori, including the first premise, and the claims about the causal roles of genuine intentions underlying the third. But other aspects are undeniably empirical, grounded in an inference to the best explanation of the available data. This is obviously true of the second premise, but is also true of the proposed account of how System 2 events characteristically achieve their effects. Hence the argument as a whole is non-demonstrative. But it can be none the less convincing for that.[4]

References

Armstrong, D. (1968). *A materialist theory of the mind*. Routledge.

Armstrong, D. (1984). Consciousness and causality. In D. Armstrong & N. Malcolm (Eds.), *Consciousness and causality*. Blackwell.

Baars, B. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Baars, B. (1997). *In the theatre of consciousness*. Oxford University Press.

Baillargeon, R., Kotovsky, L., & Needham, A. (1995). The acquisition of physical knowledge in infancy. In D. Sperber, D. Premack, & A. Premack (Eds.), *Causal cognition*. Oxford 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* (Vol. 6). Academic Press.

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.

Burge, T. (1996). Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society, 96*, 91–116.

Byrne, R., & Whiten, A. (Eds.). (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press.

Byrne, R., & 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.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*, 67–90.

Cohen, L. J. (1992). *An essay on belief and acceptance*. Oxford University Press.

Damasio, A. (1994). *Descartes' error: Emotion, reason and the human brain*. Papermac.

Dennett, D. (1984). *Elbow room: The varieties of free will worth wanting*. MIT Press.

Eagly, A., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace.

Evans, J., & Over, D. (1996). *Rationality and reasoning*. Psychology Press.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fodor, J. (1987). *Psychosemantics*. MIT Press.

Frankish, K. (2004). *Mind and supermind*. Cambridge University Press.

Gazzaniga, M. (1998). *The mind's past*. University of California Press.

[4] A version of this paper was delivered to a philosophy colloquium in Cuidad University, Mexico City. I am grateful to all those who participated in the ensuing discussion. I am also grateful to Keith Frankish for his comments upon an earlier draft.

Gigerenzer, G., Todd, P., and the ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford University Press.
Heal, J. (1994). Moore's paradox: A Wittgensteinian approach. *Mind, 103*, 5–24.
Hume, D. (1739). *A treatise of human nature*. Many editions now available.
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
Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.
Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy, 63*, 17–25.
Lycan, W. (1987). *Consciousness*. MIT Press.
Lycan, W. (1996). *Consciousness and experience*. MIT Press.
Malcolm, N. (1984). Consciousness and causality. In D. Armstrong & N. Malcolm (Eds.), *Consciousness and causality*. Blackwell.
Milner, D., & Goodale, M. (1995). *The visual brain in action*. Oxford University Press.
Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton University Press.
Moshman, D. (2004). From inference to reasoning: The construction of rationality. *Thinking and Reasoning, 10*, 221–239.
Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding of other minds*. Oxford University Press.
Nisbett, R., & Wilson, T. (1977). Telling more than we can know. *Psychological Review, 84*, 231–295.
Piattelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds*. John Wiley.
Pillow, B. (2002). Children's and adult's evaluation of certainty of deductive inference, inductive inference, and guesses. *Child Development, 73*, 779–792.
Pylyshyn, Z. (2003). *Seeing and visualizing*. MIT Press.
Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.
Shoemaker, S. (1988). On knowing one's own mind. *Philosophical Perspectives, 2*, 183–209.
Shoemaker, S. (1990). First-person access. *Philosophical Perspectives, 4*, 187–214.
Smith, J., Shields, W., & Washburn, D. (2003). The comparative psychology of uncertainty monitoring and meta-cognition. *Behavioral and Brain Sciences, 26*, 317–373.
Spelke, E., Phillips, A., & Woodward, A. (1995). Infants' knowledge of object motion and human action. In D. Sperber, D. Premack, & A. Premack (Eds.), *Causal cognition*. Oxford University Press.
Stanovich, K. (1999). *Who is rational? Studies of individual differences in reasoning*. Lawrence Erlbaum.
Stich, S. (1983). *From folk psychology to cognitive science*. MIT Press.
Wegner, D. (2002). *The illusion of conscious will*. MIT Press.
Wicklund, R., & Brehm, J. (1976). *Perspectives on cognitive dissonance*. Lawrence Erlbaum.
Wilson, T. (2002). *Strangers to ourselves*. Harvard University Press.