Chapter 5

An architecture for dual reasoning

Peter Carruthers

This chapter takes for granted the existence of some sort of real distinction between System 1 and System 2 reasoning processes, and asks how they are realized in the human mind–brain. In contrast with the usual view of the two systems as distinct from one another, it is argued that System 2 is partly realized in cycles of operation of System 1. But System 2 is also distinctive in being action-based. It is mental rehearsals of action that generate and sustain the cycles of operation of System 1 that constitute System 2. A number of advantages of the proposed account will be detailed, together with some implications for future research.

Introduction

Dual systems theories of human reasoning are now quite widely accepted, at least in outline (Evans and Over, 1996; Sloman, 1996, 2002; Stanovich, 1999; Kahneman, 2002; Frankish, 2004). I shall begin by delineating the contrasting sets of properties of the two systems that I propose to take as my explanatory target. (These are summarized in Table 5.1.)

Most researchers agree that System 1 is really a collection of different systems that are fast and unconscious, operating in parallel with one another. The principles according

Table 5.1 Distinctions between cognitive processes

S1	S2
Associative	Rule-based
Heuristic	Analytic
Parallel	Serial
Automatic	Controlled
Unconscious	Conscious
Low demands on cognitive capacity	High demands on cognitive capacity
Relatively fast	Relatively slow
Contextualized	Decontextualized
Evolutionarily old	Evolutionarily new
Conserved across species	Unique to humans

to which these systems function are, to a significant extent, universal to humans, and they aren't easily altered (e.g. by verbal instruction). (I should emphasize that in my view it is only the general principles of operation of System 1 systems that are hard to alter, rather than their contents. For many of these systems have been designed for *learning*, enabling us to extract new information from our environment in quick and reliable-enough ways. See Carruthers, 2006.) Moreover, the principles via which System 1 systems operate are, for the most part, heuristic in nature ('quick and dirty'), rather than deductively or inductively valid. It is also generally thought that most, if not all, of the mechanisms constituting System 1 are evolutionarily ancient and shared with other species of animal.

In addition, however, some researchers maintain that System 1 processes are associative in character (Sloman, 1996, 2002). I disagree. I believe that System 1 is a collection of semi-independent modules whose internal processes are, rather, computational in nature (Carruthers, 2006). I should stress, however, that those who are associationists about System 1 should face no obstacles in accepting the proposed architecture for the two systems that I shall outline and defend in the present chapter. I merely want to emphasize that by defending the reality of the System 1/System 2 distinction I am *not* intending to defend an account of the distinction that links it to the contrast between associative and rule-governed (nor between irrational and rational) forms of cognition.

System 2, on the other hand, is generally thought to be a single system which is slow, serial, and conscious. The principles according to which it operates are variable (both across cultures and between individuals within a culture), and can involve the application of valid norms of reasoning. (Indeed, some seem to believe that System 2 always and exclusively instantiates logical principles. This is not part of the view that I shall defend. I maintain that System 2, too, can involve the use of heuristics.) These System 2 principles are malleable and can be influenced by verbal instruction, and they often involve normative beliefs (that is, beliefs about how one *should* reason). Moreover, System 2 is generally thought to be uniquely human.

It seems likely that the mechanisms that constitute System 1 belong to at least three distinct kinds. There will be systems that issue in new beliefs, systems that issue in new goals, and systems concerned with swift decision making in the light of one's beliefs and goals. For there is good reason to think that System 1 will exemplify a belief/desire/decision making architecture (Carruthers, 2006, ch.2). This System 1 architecture is depicted in Figure 5.1 (which also incorporates the dual visual systems hypothesis of Milner and Goodale (1995), according to which the ventral/temporal-lobe system makes its outputs available for belief-formation and planning, while the dorsal/parietal-lobe system is concerned with the on-line guidance of movement).

One might then expect System 2 to carve up into three distinct sub-components also: one charged with conscious, reflective, belief-fixation; one subserving conscious, reflective, goal-adoption; and one of which takes conscious decisions (thereby forming new intentions) in the light of one's conscious beliefs and goals. For there certainly appear to be tasks of each of these three types (e.g. reasoning tasks versus decision making tasks) which would be categorized as involving the operations of System 2. I shall argue later, however, that there is actually just a single system constituting System 2, which can nevertheless operate in different 'modes' corresponding to belief, desire, and decision making.
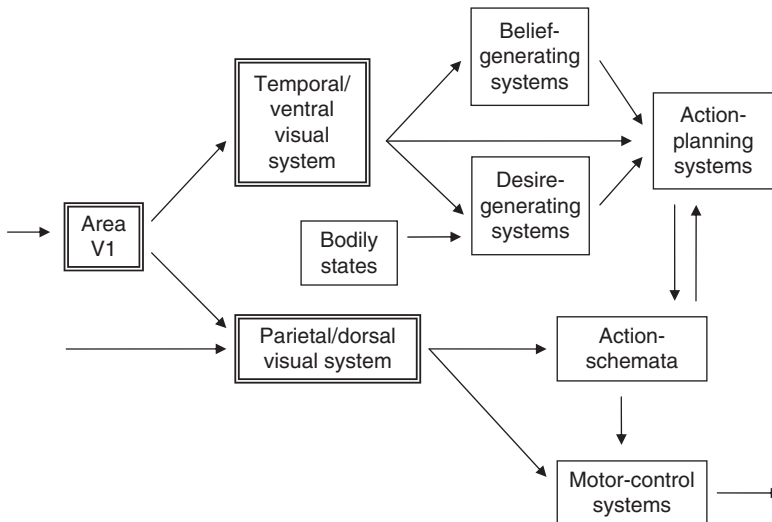
Fig. 5.1 The System 1 architecture.

It is generally assumed that System 1 and System 2 are (largely) distinct from one another. But then one immediate challenge for dual systems theory concerns the relationships between the two (sets of) systems. How, if at all, do they interact with one another? How is it possible for System 2 to override the operations of System 1 in controlling behavior? And how are we to imagine that System 2 could have evolved? If System 2 is charged with generating new beliefs, desires, and decisions, then how could it have evolved alongside a set of mechanisms (System 1) that already possessed precisely those functions? And what evolutionary pressures could have led to such a wholesale change in our cognitive architecture—effectively creating a new, species-specific, belief/desire/decision making system alongside of a more ancient set of belief/desire/decision making systems shared with the rest of the animal kingdom? The problem is further exacerbated if one allows—as we surely should—that there are a number of other species-unique cognitive adaptations that humans possess, including a language faculty, a sophisticated mind-reading faculty, a system for normative reasoning and motivation, and perhaps others besides (Carruthers, 2006, ch.3).

Another set of challenges for dual systems theory concerns the ways in which System 2 operates. For it appears that the acquisition of new beliefs (especially normative beliefs) must somehow be capable of creating and/or rewriting the algorithms governing our System 2 reasoning. But it is hard to understand how this can happen. How is it possible for processes of belief-formation, for example, to be guided by acquired beliefs? (Put differently: how can you build a learning mechanism *out of beliefs*?) And how, moreover, can reasoning be controlled by beliefs about how one *ought* to reason? Likewise, how can reasoning be guided by verbal instruction? It looks as if System 2 reasoning must somehow be under our intentional control (Frankish, 2004, this volume). And that means, in turn, that it should somehow be action-based or action-involving. For it is actions *par excellence* that are under the control of beliefs and desires, and that can be directly guided by verbal instruction. But this is then

puzzling in its own right: how can a system of *reasoning* be constituted by *actions*?—for actions are the sorts of things that rather *issue from* prior reasoning (see Figure 5.1).

My goal in this chapter is to answer these challenges. I shall outline an architecture for System 2 that sees it as realized in cycles of operation of System 1 (rather than existing alongside of the latter). On the account that I shall provide, mental rehearsals of action lead to globally broadcast images of those actions, which are in turn received as input by all of the System 1 mechanisms for the formation of beliefs, desires, and emotions, leading to a changed cognitive and affective landscape which forms the context and partial cause for the next mental rehearsal. And because action-selection in general is under intentional control, so is System 2. In the second to fifth sections I shall explain the various components of the account, showing that they are independently motivated. I shall then, in the sixth section, describe the ways in which System 2 issues in new beliefs and intentions, before detailing the advantages of the account in the seventh section. In the eighth section I shall draw attention to a number of testable predictions, before outlining some implications for researchers in the field in the ninth.

The mental rehearsal of action

The uses of 'forward models' of action for fine-grained action control are now quite well understood (Wolpert and Ghahramani, 2000; Wolpert and Flanagan, 2001; Wolpert et al., 2003). At the same time as an activated motor schema issues motor commands to the muscles to initiate a movement, an 'efferent copy' of those commands is created and compared with the initial motor intention, thus allowing for swift self-correction before the movement itself has even begun. But the efferent copy is also transformed via one or more 'emulator systems' that model the kinematics of the body (Grush, 2004) so as to match the incoming proprioceptive and other (e.g. visual) perceptual representations of the action as it is executed, again allowing for fast on-line correction. (This is thought to be one of the main functions of the dorsal–parietal visual system, represented in the lower half of Figure 5.1, and its equivalent in other sense modalities.)

Efferent copies are not confined to the dorsal-parietal system, however, yielding bare representations of bodily movement. They are also used to drive visual imagery within the ventral-temporal system (and its equivalent for other forms of perception). These images can then interact with the inferential systems that normally operate on the basis of ventral input, so as to generate predictions of the likely immediate consequences of the movement. This allows for a comparison between predicted effects and perceived effects as the action unfolds, again allowing for correction of the original motor schema. (And note that in general it is the effects of our movements that interest us, rather than those movements themselves.) For example, someone lifting a jug to pour coffee into a cup will need to predict where, precisely, to position the spout of the jug in relation to the cup; but if the coffee is observed to fall too close to the edge of the cup as it begins to pour, the position of the hand and arm can be adjusted accordingly.

There is now robust evidence of the contribution of motor and pre-motor cortex to the generation and transformation of conscious visual images located in temporal cortex (Turnbull et al., 1997; Ganis et al., 2000; Richter et al., 2000; Kosslyn et al., 2001; Lamm et al., 2001; Schubotz, 2007). Although images are also frequently
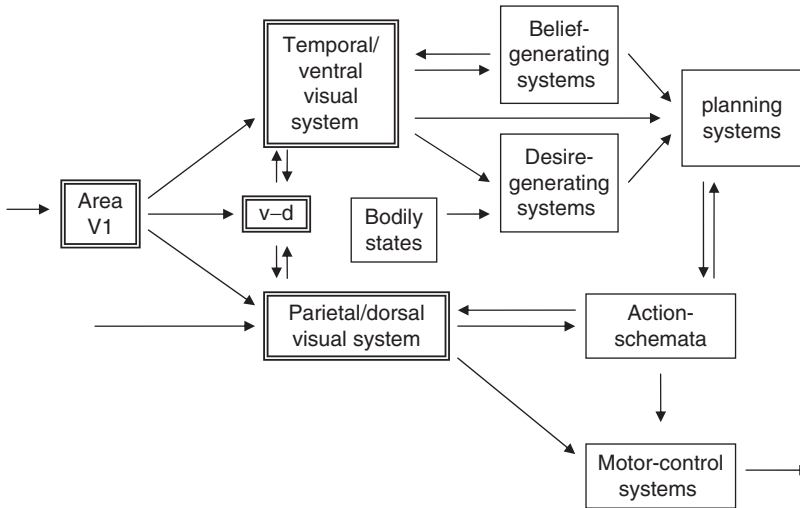
Fig. 5.2 Two visual systems with back-projecting pathways.

created in temporal cortex for purposes of object recognition (Kosslyn, 1994), it is motor cortex that initiates the generation of images in the absence of any relevant visual stimulus, and which moves and transforms the images created. This happens via an area of ventro-dorsal cortex that is probably best thought of as a common functional component of the two primary visual systems. The area in question is the superior temporal sulcus and area FP in the rostral part of the inferior parietal lobule. These are strongly interconnected with each other, and also with area F5 in the pre-motor cortex (Rizzolatti, 2005).[1] See Figure 5.2 (in which 'vd' stands for 'ventro-dorsal') for a depiction of this more complex version of the System 1 architecture.

The systems described above are probably quite ancient, and evolved initially for purposes both of controlling movement and anticipating its effects so that one can monitor and prepare to respond to those effects in turn. But these systems are also ideally suited to subserve the mental *rehearsal* of action, independently of any overt movement. When operated in this mode one activates an action schema 'off line', with the normal set of instructions to the muscles suppressed. The efferent copy of this activated schema is then used to drive representations of the action and its likely effects for purposes of decision making. By 'trying out' actions in imagination

---

[1] These areas form part of the so-called 'mirror neuron' system in humans and monkeys (Gallese et al., 1996; Rizzolatti et al., 2000). There are neurons within this system that fire *both* when the monkey perceives someone performing a quite specific movement (such as grasping a piece of food with the whole hand), *and* when the monkey executes just such a movement for itself. Much of the discussion of the function of these neurons has concerned their possible role in enabling us to understand the actions and goals of others (Gallese and Goldman, 1998; Goldman, 2006). But they are just as well positioned to enable us to map our own intended movements onto visual representations of movement.

we can test whether they are likely to achieve our goals, and by creatively activating and rehearsing actions from our repertoire we can sometimes hit upon novel solutions to problems. There is some evidence of the use of mental rehearsal when problem solving by other apes (especially chimpanzees; see Carruthers, 2006, ch.2). And there is robust evidence of mental rehearsal of action amongst members of *Homo ergaster* from over a million years ago, which enabled them to create stone tools of fine symmetry. For the only known way of producing such tools, out of materials that always vary in detail from case to case, involves mentally rehearsing the likely effects of a given strike on the core, thus enabling the stone knapper to plan several blows ahead (Mithen, 1996; Wynn, 2000; Carruthers, 2006, ch.2).

The global broadcast of rehearsed actions

What becomes of the representations in ventral-temporal cortex that result from the mental rehearsal of an action? Under certain conditions (e.g. when attended to) those representations are 'globally broadcast' to a wide range of consumer systems for forming new beliefs and memories, for creating new desires and emotions, and for practical reasoning (see Figure 5.2). Initially proposed by Baars (1988), there is now robust and varied evidence of the global broadcasting of the outputs of the ventral-temporal visual system and of the coincidence of such broadcasts with conscious experience (Dehaene and Naccache, 2001; Dehaene et al., 2001, 2003, 2006; Baars, 2002, 2003; Baars et al., 2003; Kreiman et al., 2003).

It is worth stressing that global broadcasting theory is consistent with a variety of different accounts of the nature of consciousness itself. First-order theorists of consciousness like Dretske (1995) and Tye (1995, 2000), for example, can claim that perceptual and imagistic contents become conscious *in virtue of* being globally broadcast. For such theorists maintain that consciousness consists in the immediate availability of such contents to first-order processes of belief-formation and decision making. (These processes are described as 'first-order' because the resulting beliefs and decisions are about the worldly events represented in perception, rather than about those perceptual states themselves.) Higher-order theorists of consciousness like Lycan (1987, 1996), Carruthers (2000, 2005), and Rosenthal (2005), in contrast, will maintain that only a single aspect of global broadcasting is really relevant to the conscious status of the percepts and images in question, which is their availability (via global broadcast) to the mind-reading system. For such theorists maintain that conscious perceptions are those that we are aware of having at the time, in a higher-order way.

It is also worth emphasizing that it is only the perceptual *contents* that get globally broadcast that are conscious. The cognitive processes that underlie and sustain those broadcasts certainly aren't. And many of the processes that issue in a content being globally broadcast aren't conscious either, whether these are 'bottom-up' (such as the mechanism that identifies one's name in an unattended conversation at a cocktail party, causing it to pop into awareness) or rather involve top-down forms of attention. The only exception would be that *sometimes* a decision to direct attention at a stimulus or type of stimulus can be conscious (as when I remind myself when doing a jigsaw puzzle, 'I must pay attention to the shape'). But the intervening

processes that begin from a conscious thought of this sort and issue in a globally broadcast perception or image won't themselves be conscious ones.

It is likely that the mechanisms underlying the global broadcast of attended perceptual events are evolutionarily ancient. Indeed, they provide a core aspect of the System 1 cognitive architecture. (see Figure 5.1). Global broadcast enables perceptual contents to be made available as input simultaneously to the full range of System 1 systems. Some of these have been designed to draw inferences from the input, generating new beliefs. Others have been designed to create new desires and emotions. (Think how seeing a piece of chocolate, even when one is replete, can give rise to a desire to eat chocolate; and think how the sight of a snake rippling through the grass can give rise to a shiver of fear.) And yet others are charged with the creation of new plans for achieving one's goals in relation to the perceived environment. ('I'll go around *that* way and pick up *that* one.')

When the globally broadcast representations are, not percepts of external events, but rather images resulting from the mental rehearsal of an action, the System 1 systems process those representations as normal (at least initially). Mentally rehearsed actions, too, generate inferences and emotional reactions. But it is important that those inferences should *not* issue in beliefs as they usually would. When the *sight* of a hand pushing a vase gives rise to the prediction that the vase will fall, it is appropriate for the subject to believe (in advance of it doing so) that the vase will fall. (This belief might then lead the subject to leap to catch the vase to prevent it breaking.) But when the *image* of a hand pushing the vase is generated through mental rehearsal of a pushing action-schema directed at the vase, and the same prediction results, the subject should *not* believe that the vase will, actually, fall. For the vase will only fall if it is actually pushed. What should be believed, rather, is a conditional: it will fall if pushed like that. This might lead subjects to go ahead and execute the pushing schema, or to abandon it and begin a search for some alternative means of moving the vase, depending on their purposes.

Notice that mental rehearsal of an action is the functional equivalent of *supposing* that the act is performed. Inferences drawn from the rehearsal must therefore be tagged, somehow, as dependent upon a self-produced mental rehearsal, in such a way that the conclusions reached aren't believed outright. Although small, this was by no means a trivial alteration in the mode of operation of the set of System 1 systems. For the result was the creation of what Nichols and Stich (2003) call a 'possible worlds box' – a workspace where possibilities can be tested and explored in advance of action.

Soma-sensory monitoring of the effects of mental rehearsal

While the images resulting from mentally rehearsed actions don't give rise to real beliefs when they are received as input by the System 1 belief-generating systems (except perhaps conditional beliefs), they do appear to give rise to real emotions and motivations. As is familiar, imagined sex acts can make you sexually aroused, by imagining yourself eating a piece of chocolate cake you can make yourself hungry, and so forth. So the impact of imagery on System 1 motivational systems appears somewhat different from its impact on belief systems. But even here, something a bit
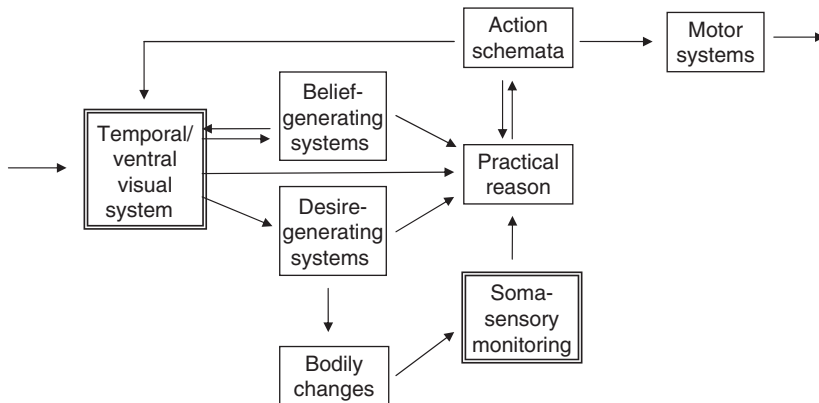
Fig. 5.3 Mental rehearsal and soma-sensory monitoring.

like tagging of conclusions to display their dependence on a supposition would appear to be involved, since the functional roles of the emotions caused via mental rehearsal are distinctively different from normal.

According to Damasio (1994, 2003) our emotional responses to mentally rehearsed actions and their predicted consequences give rise to a variety of physiological changes, such as altered heart rate, respiratory rate, and so on.[2] (See also Rolls, 1999. For a related but distinct account, see Schroeder, 2004.) These somatic changes are in turn perceived, and are used to ratchet up, or down, the desirability of the rehearsed action. If the over-all somatic effects of the rehearsal of the action are positively valenced then a desire to perform that action will generally result (unless the rehearsal in question is a mere fantasy, unrelated to any current project or possibility); whereas if the overall effects are negatively valenced, then we are motivated *not* to perform the action (see Figure 5.3).

Consider a particular example. Looking at my monthly credit card statement, I realize that I need more money. After reviewing some options, I hit upon the idea of going to ask my boss for a raise. I mentally rehearse the action of walking into his office and broaching the question of salary. The resulting images are globally broadcast, and are elaborated by System 1 inferential systems to include my boss' likely response (the glowering face, the harsh words). The result is that I feel fear and disappointment. And that leads me to abandon any thought of asking for a raise, and returns me to considering other options.

We spend much of our waking lives, as adults, in mental rehearsals of this sort, often to good effect. (And yet sometimes *not* to good effect—see Gilbert, 2005, for discussion of the ways in which our imaginings can lead us astray in our reasoning about the future.) Initially promising plans can turn out to be disastrous when rehearsed; and plans whose success at first seems implausible can turn out to be much more likely to succeed. Moreover, when the frontal-lobe systems concerned with soma-sensory monitoring are

---

[2] Damasio (1994) also argues that these systems can operate in swifter 'as if' mode, during which physiological changes are predicted and responded to without actually needing to occur.

damaged, our capacities for medium and long-term planning are severely interfered with (Damasio, 1994). The patients in question can *reason*, theoretically, about practical matters in perfectly sensible ways, but their decision making in light of that reasoning is very poor. In consequence their practical lives are often in a terrible mess.

Inner speech

As we noted above, a distinction somewhat similar to that between the ventral-temporal and dorsal–parietal visual systems has been confirmed for other sense modalities (Michel and Peronnet, 1980; Paillard et al., 1983; Rossetti et al., 1995). And we also know that mental rehearsal of actions can give rise to conscious imagery of other sorts besides vision. More specifically, rehearsal of speech actions gives rise to imagery, resulting in so-called 'inner speech'. (Most often these images are auditory, representing the sounds that would result if those speech actions were executed. But they can also be articulatory or—in the case of deaf signers—visual). And here, too, one function (and perhaps the original function) of the systems underlying inner speech is fast on-line repair of action itself (Levelt, 1989).

It is commonly accepted that the language faculty contains distinct production and comprehension sub-systems, each of which can draw on a common database of linguistic knowledge (Chomsky, 1995). In that case the systems underlying the phenomenon of inner speech can be seen depicted in Figure 5.4. (It should be noted that there is evidence that both the language production and the language comprehension areas of the cortex are active during inner speech. See Paulescu et al., 1993; Shergill et al., 2002.) Here is how it works. In light of the subject's beliefs and goals, a speech action-schema is formulated by the language production sub-system. While overt action is suppressed, an efference copy of the motor instructions is transformed via an emulator system into an auditory representation of the sounds that would have resulted had the action been carried out. This representation is globally broadcast in the manner of conscious images generally, and is received *inter alia* by the language comprehension sub-system. The latter constructs an interpretation of the utterance in the normal way, and presents that (attached to the sounds) to the various other System 1 inferential systems. Hence, just as is the case with external speech, we seem to hear the meaning of the imagined sounds of inner speech (the message expressed) as well as hearing those imagined sounds themselves.

Hurlburt (1990, 1993) has demonstrated the near ubiquity of inner speech in the waking lives of normal people. Subjects in his studies wore headphones during the course of the day, through which they heard, at various intervals, a randomly generated series of bleeps. When they heard a bleep, they were instructed to immediately 'freeze' what was passing through their consciousness at that exact moment and then make a note of it, before elaborating on it later in a follow-up interview. Although frequency varied widely, all normal (as opposed to schizophrenic) subjects reported experiencing inner speech on some occasions—with the minimum being 7% of occasions sampled, and the maximum being 80%. Most subjects reported inner speech on more than half of the occasions sampled. (The majority of subjects also reported the occurrence of visual images and emotional feelings—on between 0% and 50% of
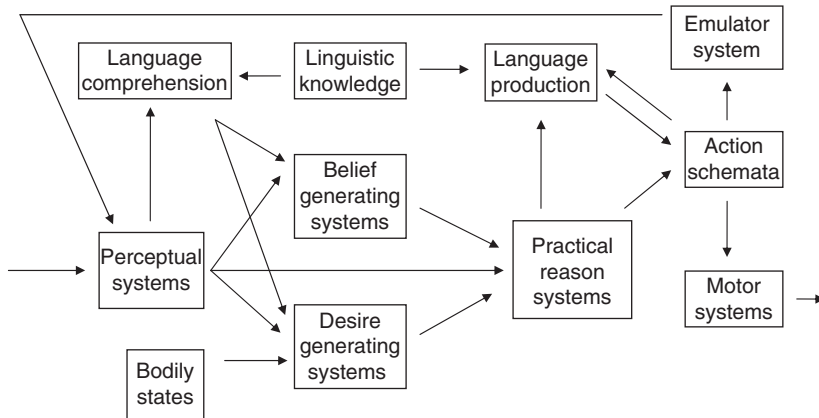
Fig. 5.4 The metnal rehearsal of speech.

occasions sampled in each case). This is an extremely high proportion of time to devote to mental rehearsal of a single type of activity.

Mental rehearsals of speech, like mental rehearsals of action generally, are often undertaken to test whether or not the utterance should actually be made, overtly. We often 'try out' utterances in imagination in advance of making them, allowing our various System 1 inferential systems to evaluate the likely effects on our audience, while reacting emotionally to those effects. But this is by no means the most common way for inner speech to be utilized. More often we are interested in the content of imagined utterances, rather than in (the effects of) those utterances themselves. And we take those imagined utterances to express, or to be constitutive of, our thought processes. This is a large part of the foundation on which System 2 thinking and reasoning is built.

Putting it all together

One set of advantages of this account of System 2 should be readily apparent. For there is no need to postulate a distinct belief/desire/decision making system alongside of System 1. Rather, System 2 is *realized in* cycles of operation of System 1. We have independent reasons to believe in the existence of many of the component mechanisms and processes utilized in such cycles, and it is likely that they evolved initially for other purposes. The basic capacity for mental rehearsal of action when problem solving is probably shared with other great apes. And when a language faculty, a more sophisticated mind-reading faculty, and some kind of system for normative reasoning and motivation were added to the System 1 architecture alongside its other components, it would have taken but small changes for the full range of System 2 processes to come into existence.

We do need to explain, however, how System 2 beliefs, goals, and decisions emerge out of the cyclical operations of System 1. The easy case is the mental rehearsal of action (whether bodily action or speech action) which eventuates in an evaluation of the consequences of that action. In such a case a schema for the action in question is activated and rehearsed, giving rise to a globally broadcast imagistic representation of the act. This is received as input by the System 1 inferential and motivational

modules, some of which may elaborate it to include some likely consequences. Our bodily/emotional reactions to the envisaged act and its consequences are monitored, with the desirability of the action being adjusted up or down. Such a newly created desire to perform the action may then be sufficient in the circumstances to issue in an intention to perform it (either immediately or when the envisaged circumstances arise), utilizing our regular System 1 decision making procedures.

More challenging, however, is to explain how mentally rehearsed speech acts can give rise to new beliefs in the propositions expressed by those acts, or to novel intentions to act as described. For here it isn't the consequences of *performing* the speech acts in question that need to be evaluated, but rather the *contents* of those acts. I shall defend a pluralist position. It seems to me that there are a variety of ways in which inner speech can issue in new beliefs, goals, and intentions. I shall begin, however, with the proposal made by Frankish (2004, this volume).

On Frankish's account the mind reading system has a crucial role to play. If it interprets a given utterance in inner speech, in the circumstances, as a *commitment*, or a 'making up of mind', then the functional equivalent of a new belief, desire, or intention will be the result (depending on the sort of commitment that gets attributed), provided that there exists a standing System 1 desire to execute one's commitments. Suppose, first, that an utterance is interpreted as a commitment to the truth of the proposition that it expresses. The subject therefore forms a System 1 belief that a commitment of that kind has been made. This will then in future interact with the System 1 desire to honor commitments, issuing in further overt or covert verbalizations. If asked whether he believes the proposition in question, for example, the subject will reply that he does. For one of the things one ought to be prepared to do, if one has committed oneself to the truth of a proposition, is assert it. And likewise, during one's System 2 practical or theoretical reasoning one will be prepared to rely upon the sentence in question as a premise. For if one has committed oneself to the truth of a proposition then one ought also to commit oneself to any other proposition that one believes follows from it. And so on.

Frankish provides a similar account of the formation of System 2 goals and intentions. Let me here just work through the case of intention. Suppose that a sequence of System 2 reasoning concludes with an inner verbalization of a sentence of the form, 'So, I shall do *P*.' This is globally broadcast and received as input by the mind-reading system *inter alia*. The latter interprets it as expressing a *commitment* to do *P*, and stores a record of this commitment in memory. Later, when the time to act arrives, this memory is activated, and it combines with the subject's standing (System 1) desire to execute his commitments in such a way as to issue in a (System 1) decision to do *P*. Notice that on this account, although the initial verbalization wasn't, itself, the formation of an intention, by being interpreted as committing the subject to act in a certain way it becomes the System 2 functional equivalent of an intention to act.

I have no doubt that System 2 belief formation and intention formation sometimes work in these ways. Sometimes, for example, I might quite naturally report what took place by saying, 'I realized that I had been thinking long enough, and that I had better commit myself to believing/doing *something*, so I made up my mind to believing/doing *P*.' But there is a problem about claiming that System 2 attitude formation is *always* a matter of commitment. For on other occasions I might report what took place by saying, 'After thinking about the matter for a while, I realized that *P*' or

by saying, 'After thinking about it for a while, I realized that I should do *P*.' Here there is no mention of commitment, or anything resembling commitment.

Frankish (2004, this volume) can (and does) respond that the beliefs about commitment that realize System 2 believing and intending are often unconscious ones, and so aren't reportable in speech. But this is problematic. For although System 1 *processes* aren't conscious, the beliefs and intentions in which those processes issue generally are. (While there are multiple memory systems within System 1, in general many of the beliefs that I express in speech are formed via the operations of System 1.) In particular, when my mind-reading system (operating unconsciously for the most part, at the System 1 level) issues in a belief about the beliefs or intentions of other people, I can generally express that belief in speech. So why should it be so different when the belief produced by the mind-reading system is a belief about my own commitments? Why should the latter beliefs so often be 'screened off' from consciousness and verbal report?

In (some of) the cases where Frankish's commitment-based account fails to apply, it might be better to see System 2 as driven by beliefs about one's own beliefs or intentions, together with a desire for *consistency*. When I say to myself at the conclusion of a period of System 2 reasoning, 'So, P is the case', I will interpret myself as having formed the belief that *P*. (And I might quite naturally express that second-order belief in speech by saying that at that moment I realized/formed the belief that *P*.) Hence I shall thereafter believe that I believe that *P*. In my later reasoning and planning, this higher-order belief may become activated. But I also have the normative belief that, if I believe *P*, I should think and act in ways that are appropriate in the circumstances. Wanting to think and act as I should, or wanting my thinking and acting to be consistent with what I believe myself to believe, I am motivated to think and act just as if I believed that *P*. This account has the advantage of not requiring me to have any beliefs or goals that I couldn't articulate, or that I would be unwilling to attribute to myself.

System 2 thinking might also issue in beliefs and intentions more directly and simply, however. (See Carruthers, 2006, ch.6, for elaboration of the idea sketched here.) Entertaining in inner speech the sentence, 'P', the content of that utterance is extracted and evaluated much as if it were the testimony of another person. It is checked for consistency and coherence with existing beliefs, for example. If it meets the appropriate standards, and there are no reasons *not* to believe *P*, the content that *P* is accepted and stored in whatever manner is usual for a belief of that type.

Advantages of the account

In addition to removing any need to regard System 1 and System 2 as distinct, existing alongside one another, a further advantage of the account presented in this chapter is that it explains why System 2 processes should be comparatively slow and serial, and why (some of) its operations should be conscious. Since System 2 is realized in *cycles* of operation of System 1 it will be slow by comparison. And since only one action can be mentally rehearsed and globally broadcast at a time, System 2 will be serial in its operation (but utilizing the parallel-process functioning of System 1). And since the images that result from mental rehearsal of each action in the cycle are globally broadcast, and we know that such broadcasts correlate closely with consciousness, we can explain why each such stage in each cycle should be conscious. (The other stages, by contrast, will be *un*conscious, including the processes that select a given

action-schema for rehearsal, and those that draw inferences from, or generate emotional reactions to, the broadcast image.)

In addition, this account can explain some of the individual variation that exists in people's capacity to solve System 2 tasks, specifically the variation that correlates with 'cognitive style' (Stanovich, 1999; see also Stanovich, Evans, Buchtel and Norenzayan, this volume). For this can result from individual differences in the likelihood of utilizing System 2 at all, or of activating it when problem solving. A disposition to be reflective, for example, which correlates with success in System 2 tasks, will consist in a capacity to suppress one's initial response generated by System 1, mentally rehearsing it and other alternatives that come to mind, hence allowing the knowledge that is stored across the various System 1 systems to influence the eventual outcome.

Moreover, since on this account System 2 is action-based, it predicts that System 2 processes should be malleable and subject to learning in any of the ways that action itself is. We acquire behavioral skills and abilities by imitation of others, by receiving verbal instruction, and by forming normative beliefs about the ways in which one should behave. So we can predict that System 2 thinking skills should be acquirable by imitation and by instruction, and that sequences of System 2 reasoning should be shaped by beliefs about the ways in which one *should* reason. Each of these predictions is borne out, I believe.

In support of the first of these predictions (that System 2 skills should be acquirable by imitation) one can cite the common belief—held by many teachers—that one way of teaching intellectual skills is by *exhibiting* them. Many university professors (especially in philosophy) hope that by *thinking through* a problem in the presence of the students, some of the ways in which one should think will thereby be imitated and acquired. Likewise, this is a plausible construal of what happens in scientific lab meetings. It is by talking through problems and potential solutions in the presence of undergraduate and graduate students that the more senior researchers impart many of the intellectual skills necessary for science. And this is consistent with the widely recognized fact that much of what younger scientists have to acquire is *know how* of various sorts, in addition to specific facts and theories.

Another skill acquisition mechanism is explicit instruction. People can *tell* me what actions and sequences of action to perform in the service of a given goal. Thus a logic teacher might tell me what steps to take in order to evaluate the truth of a conditional, just as an experienced kayak maker might tell a novice how to prepare and shape the materials for use in the frame. And in both cases the beliefs thus acquired can be recalled at the point of need, and used to guide novel instances of the activity in question (evaluating a conditional, building a kayak).

People also learn normative facts, and come to believe that there are certain things that they *should* or *should not* do. And in many cultures these normative beliefs concern the processes of (System 2) reasoning itself. Since it is a general fact about normative beliefs that they are apt to give rise to an intrinsic motivation to perform (or refrain from) the action required (or forbidden; Sripada and Stich, 2006), this will mean that people will be intrinsically motivated to rehearse sequences of action that will constitute certain abstract patterns of System 2 thinking.

Consistent with this prediction, people do seem to be intrinsically motivated to entertain or to avoid certain types of thought or sequence of thought. Thus if

someone finds himself thinking that *P* while thinking that $P \supset Q$, then he will feel *compelled*, in consequence, to think that *Q*. And if someone finds herself thinking that *P* while also thinking that *~P*, then she will feel herself *obligated* to eliminate one or other of those two thoughts. The explanation is that the system for normative reasoning and belief has acquired a rule requiring sequences of thought/action that take the form of *modus ponens*, as well as a rule requiring the avoidance of contraction, and is generating intrinsic motivations accordingly.

As should be clear from the above, the present account also predicts that there should be wide variations in the patterning of System 2 reasoning across cultures, just as there is variation in both skilled behavior and normative belief and behavior. For of course cultures will vary in their beliefs about which sequences of System 2 reasoning are likely to be successful; and they will differ in their beliefs about what sequences are normatively required or forbidden. A significant part of what has changed in the history of science, for example, consists in culturally shared beliefs about the ways in which one *should* reason scientifically, and the ways in which one *should* evaluate hypotheses. Just think, for instance, of the impact that statistical methods have increasingly had in the social sciences over the last century.

Finally, it should also be stressed that the present account is fully consistent with the fact that 'think aloud' protocols are reliable indicators of the ways in which subjects actually solve System 2 tasks, as demonstrated by task analysis combined with timing and error patterns (Ericsson and Simon, 1993). As Ericsson and Simon emphasize, it is crucial that experimenters should *not* ask subjects to report *on* their thoughts while working on the problem. For this meta-reflection will (and demonstrably does) interfere with the conduct of the first-order reasoning in question. Rather, they should be asked to 'think their thoughts aloud', articulating in an unreflective way the reasoning that they go through. This should be predicted not to interfere with the task if, but only if, that task is normally conducted in inner speech. For otherwise the cognitive resources necessary to formulate into speech the underlying reasoning should have an effect on task performance.

Some predictions

What is distinctive of the theory of System 2 reasoning presented here is that it is *action based*. According to this account, it is mentally rehearsed actions that initiate and sustain System 2 reasoning, thereby recruiting and utilizing the mechanisms that also subserve System 1 reasoning, as well as activating normative beliefs about proper reasoning, together with the stored schemata for skilled action sequences, and so forth. This leads to a number of clear predictions.

The first prediction is that patients with Huntington's disease, which attacks especially motor and pre-motor cortex, should be much weaker at System 2 than System 1 tasks in comparison with other brain-damaged populations. Forms of brain damage that interfere with System 1 mechanisms will tend also to interfere with System 2 processes, on the account provided here. For example, damage to the 'association areas' of temporal cortex that underlie many forms of intuitive inference will also interfere with System 2 tasks that recruit those inferences. Moreover, damage to the areas of the frontal lobes concerned with 'executive function' will also interfere with

System 2 tasks that recruit those same decision making mechanisms. (All System 2 tasks will implicate System 1 decision making, of course, since these are the mechanisms that select amongst available action schemata for activation and rehearsal.)

Damage to motor and pre-motor cortex, in contrast, should leave System 1 reasoning mechanisms intact. Patients should still be capable of forming swift and intuitive inferences in the light of perceptual data, they should still display swift emotional reactions in response to that data, and they should still be capable of making swift and unreflective decisions. But because they will have difficulty activating and rehearsing action schemata, they should have problems with System 2 tasks, which crucially involve such rehearsals, if the account that I have provided is correct.

Essentially the same prediction could also be tested more directly by temporarily 'freezing' motor and pre-motor cortex by trans-cranial magnetic stimulation (provided that subjects can be given some means for indicating their responses, of course). What we should find is that performance on System 2 tasks should suffer dramatic collapse, while System 1 tasks should be left relatively untouched. In contrast, 'freezing' areas of cortex implicated in System 1 should have an effect on both systems equally; and 'freezing' of other areas (e.g. primary visual cortex) should have no effects on either system.

We can also predict that subjects who are required to 'shadow' speech while conducting many forms of System 2 task (even 'nonsense speech' that doesn't require comprehension) should perform much more poorly than subjects who are required to 'shadow' a complex rhythm that places equivalent demands on working memory. (See Hermer-Vazquez et al., 1999, for an example of this paradigm used for another purpose.) For speech shadowing will tie up the resources of the language production sub-system, thus making it difficult for subjects to engage in inner speech, whereas rhythm shadowing, while utilizing the resources of motor cortex broadly construed, should have no effect on speech production.

Implications for System 1/System 2 research

Many researchers in the field think that it is distinctive of System 1 processes that they should be triggered automatically by perceptual cues (in 'bottom–up' fashion), and that they should operate outside of intentional control. DeSteno et al. (2002), for example, argue that jealousy isn't a System 1 system on the grounds that it is sensitive to cognitive load manipulation. Their view is that System 1 systems, because automatic and non-intentional, should continue to operate in ways that are unaffected by whatever may be occurring elsewhere in the cognitive system. (See Barrett et al. (2006) for an extended critique of DeSteno et al.'s methodology from an evolutionary psychology perspective.)

The account defended in this chapter shows quite clearly what is wrong with this way of thinking of the relationship between System 1 and System 2, however. If System 1 systems are organized around the global broadcast of perceptual output (as depicted in Figure 5.1), then whether any given System 1 system gets 'turned on' in a given context will depend crucially on what the agent is attending to. For we know that top–down attention is one of the main determinants of whether or not a given perceptual content becomes globally broadcast (Dehaene et al., 2006). So to the extent

that the agent's attentional resources are occupied elsewhere, to that extent we should expect that the processing of stimuli giving rise to jealousy—and hence the emotion of jealousy itself—will be negatively affected.

Moreover, on the account presented here, System 1 systems can get turned on by imagistic representations as well as by external cues. And since the former are characteristically under our intentional control, System 1 processing of this sort will likewise be under intentional control. By actively generating and maintaining images that give rise to the emotion of jealousy, the causing and sustaining of such feelings can be under the agent's intentional control, even if (as I believe) the system that issues in jealousy belongs to System 1. Once again, therefore, we should predict that feelings of jealousy will be highly sensitive to cognitive load manipulations of various sorts. For to the extent that agents are occupied with other cognitive tasks, to that extent they will be unlikely to maintain the imagery necessary (in the absence of attention-grabbing external cues) for a strong emotional response.

One general point that this brings out, I think, is that some care needs to be taken in how we characterize System 2 tasks. In particular, we shouldn't classify as 'System 2' all tasks whose solution requires processing that is under the intentional control of the agent. For although this will, indeed, implicate System 2 (because involving the mental rehearsal of action), it shouldn't be counted as a System 2 *task* if System 1 processing, activated from the task instructions via mental rehearsal and global broadcast, is sufficient for a solution. Rather, System 2 tasks should either require the recall and rehearsal of some appropriate culturally-acquired item of information (e.g. a normative belief), or they should require the controlled activation of sequences of mental rehearsal in accordance with learned rules, or they should implicate practices of self-interrogation (e.g. asking oneself, 'What should I do next?') Merely being prompted to imagine one's spouse being unfaithful, for example, and creating and sustaining the appropriate images, shouldn't count.

Conclusion

I suspect many people have been puzzled about how there could be two distinct systems in the human mind–brain for reasoning and decision making, each of which must replicate the functionality of the other to a significant degree. In this chapter I have aimed to remove that puzzlement, thereby providing a defense of one sort of dual systems theory. I have shown that there are good reasons for thinking that System 2 is realized in cycles of operation of System 1, utilizing mechanisms and processes that we have independent reason to believe in. The resulting account of System 2 is action-based, since it is activations and mental rehearsals of action schemata that initiate and sustain its operations.[3]

References

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge 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**, 100–14.

Baars, B., Ramsoy, T., and Laureys, S. (2003). Brain, consciousness, and the observing self. *Trends in Neurosciences*, **26**, 671–5.

Barrett, H., Frederick, D., Haselton, M., and Kurzban, R. (2006). Can manipulations of cognitive load be used to test evolutionary hypotheses? *Journal of Personality and Social Psychology*, **91**, 513–18.

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.

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

Damasio, A. (2003). *Looking for Spinoza: joy, sorrow, and the feeling brain*. Harcourt.

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**, 752–8.

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**, 8520–5.

Dehaene, S., Changeux, J-P., Naccache, L., Sackur, J., and Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, **10**, 204–11.

DeSteno, D., Bartlett, M., Braverman, J., and Salovey, P. (2002). Sex differences in jealousy: evolutionary mechanisms or artifact of measurement? *Journal of Personality and Social Psychology*, **83**, 1103–16.

Dretske, F. (1995). *Naturalizing the Mind*. MIT Press.

Ericsson, A. and Simon, H. (1993). *Protocol Analysis: verbal reports as data*. (Revised edition.) MIT Press.

Evans, J. St. B.T. and Over, D., E. (1996). *Rationality and Reasoning*. Psychology Press.

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

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, **12**, 493–501.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the pre-motor cortex. *Brain*, **119**, 593–609.

Ganis, G., Keenan, J., Kosslyn, S., and Pascual-Leone, A. (2000). Transcranian magnetic stimulation of primary motor cortex affects mental rotation. *Cerebral Cortex*, **10**, 175–80.

Gilbert, D. (2005). *Stumbling on Happiness*. Vintage Books.

Goldman, A. (2006). *Simulating Minds: the philosophy, psychology, and neuroscience of mind-reading*. Oxford University Press.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, **27**, 377–442.

Hermer-Vazquez, L., Spelke, E., and Katsnelson, A. (1999). Sources of flexibility in human cognition: dual-task studies of space and language. *Cognitive Psychology*, **39**, 3–36.

Hurlburt, R. (1990). *Sampling Normal and Schizophrenic Inner Experience*. Plenum Press.

Hurlburt, R. (1993). *Sampling Inner Experience with Disturbed Affect*. Plenum 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., Wraga, M., and Alpert, N. (2001). Imagining rotation by endogenous versus exogenous forces: distinct neural mechanisms. *NeuroReport*, **12**, 2519–25.

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**, 8378–83.

Lamm, C., Windtschberger, C., Leodolter, U., Moser, E., and Bauer, H. (2001). Evidence for premotor cortex activity during dynamic visuospatial imagery from single trial functional magnetic resonance imaging and event-related slow cortical potentials. *Neuroimage*, **14**, 268–83.

Levelt, W. (1989). *Speaking: from intention to articulation*. MIT Press.

Lycan, W. (1987). *Consciousness*. MIT Press.

Lycan, W. (1996). *Consciousness and Experience*. MIT Press.

Michel, F. and Peronnet, F. (1980). A case of cortical deafness: clinical and electro-physiological data. *Brain and Language*, **10**, 367–77.

Milner, D. and Goodale, M. (1995). *The Visual Brain in Action*. Oxford University Press.

Mithen, S. (1996). *The Pre-History of the Mind*. Thames and Hudson.

Nichols, S. and Stich, S. (2003). *Mindreading: an integrated account of pretence, self-awareness, and understanding other minds*. Oxford University Press.

Paillard, J., Michel, F., and Stelmach, G. (1983). Localization without content: a tactile analogue of 'blind-sight'. *Archives of Neurology*, 40, 548–51.

Paulescu, E., Frith, D., and Frackowiak, R. (1993). The neural correlates of the verbal component of working memory. *Nature*, **362**, 342–5

Richter, W., Somorjat, R., Summers, R., Jarnasz, N., Menon, R., Gati, J., Georgopoulos, A., Tegeler, C., Ugerbil, K., and Kim, S. (2000). Motor area activity during mental rotation studied by time-resolved single-trial fMRI. *Journal of Cognitive Neuroscience*, **12**, 310–20.

Rizzolatti, G. (2005). The mirror neuron system and imitation. In S. Hurley and N. Chater (eds.), *Perspectives on Imitation: from neuroscience to social science*, Volume 1, MIT Press.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2000). Cortical mechanisms subserving object grasping and action recognition: a new view on the cortical motor functions. In M. Gazzaniga (ed.). *The New Cognitive Neurosciences*, Second Edition, MIT Press.

Rolls, E. (1999). *The Brain and Emotion*. Oxford University Press.

Rosenthal, D. (2005). *Consciousness and Mind*. Oxford University Press.

Rossetti, Y., Rode, G., and Boissson, D. (1995). Implicit processing of somaesthetic information. *Neurological Reports*, **6**, 506–10.

Schroeder, T. (2004). *Three Faces of Desire*. Oxford University Press.

Schubotz, R. (2007). Prediction of external events with our motor system: towards a new framework. *Trends in Cognitive Sciences*, **11**, 211–18.

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.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, **119**, 3–22.

Sloman, S. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin, and D. Kahneman (eds.), *Heuristics and Biases: the psychology of intuitive judgment*. Cambridge University Press.

Sripada, C. and Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, and S. Stich (eds.), *The Innate Mind: culture and cognition*. Oxford University Press.

Stanovich, K. (1999). *Who is Rational? Studies of individual differences in reasoning*. Lawrence Erlbaum.

Turnbull, O., Carey, D., and McCarthy, R. (1997). The neuropsychology of object constancy. *Journal of the International Neuropsychology Society*, **3**, 288–98.

Tye, M. (1995). *Ten Problems of Consciousness*. MIT Press.

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

Wolpert, D. and Flanagan, R. (2001). Motor prediction. *Current Biology*, **11**, 729–32.

Wolpert, D. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, **3**, 1212–17.

Wolpert, D., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London*, B **358,** 593–602.

Wynn, T. (2000). Symmetry and the evolution of the modular linguistic mind. In P. Carruthers and A. Chamberlain, (eds.), *The Evolution of the Human Mind*. Cambridge University Press.