Forthcoming in *The Journal of Consciousness Studies*

Peter Carruthers

Cartesian Epistemology

Is the theory of the self-transparent mind innate?

This paper argues that a Cartesian belief in the self-transparency of minds might actually be an innate aspect of our mind-reading faculty. But it acknowledges that some crucial evidence needed to establish this claim hasn't been looked for or collected. What we require is evidence that a belief in the self-transparency of mind is universal to the human species. The paper closes with a call to anthropologists (and perhaps also developmental psychologists), who are in a position to collect such evidence, encouraging them to do so.

1. Introduction

Bloom (2004) argues that humans have an innate tendency to believe in Cartesian dualism. He cites as evidence the fact that belief in an ontological separation between mind and body is early to emerge in development and is universal to all people in all cultures, with the exception of a few scientifically-minded university-educated people over the last century or so. (That a phenotypic property is innate needn't mean that it is absolutely unchangeable, of course — although one might predict that it would be *resistant* to change. But one should at least expect it to be *channeled* or *canalized* in normal development, reliably appearing — without learning — in a wide range of environments. See Ariew, 1999; Samuels, 2002, 2007; see also the discussion that follows a few paragraphs below.) Bloom also relies, in addition, on the general claim that our core beliefs about minds — or whatever implicit structures underlie our mind-reading capacities — are innate, citing an extensive body of developmental and neuropsychological evidence. Hence our belief in Cartesian dualism — or our tendency to form such a belief — should be seen as one component or consequence of these innate beliefs or structures.

More recently, Bloom has unearthed evidence that very young children seem not to

believe that persons (minded creatures) are bound by the same laws and principles that govern other physical systems (Kuhlmeier et al., 2004). Infants as young as five months of age show surprise if a physical object like a box or a toy disappears behind one screen and then re-emerges from behind another without having traversed the space in between; but they show *no* surprise if a person does the same. They seem to think that something about people enables them to get from one place to another without having to travel through the intervening space.[1] And in other recent data obtained from children of varying ages, Bering and Bjorklund (2004) found that while four-year-old children show a strong tendency to believe that mental attributes like thinking and feeling continue beyond death (whereas biological attributes don't), this tendency *decreases with age*. This is the opposite of what one would predict if such beliefs resulted from socialization rather than being innate.

While I shall not attempt to argue for this here, I think Bloom is correct that there is a strong case for claiming that Cartesian dualism is innate. Note, however, that the claim isn't that belief in dualism is an adaptation, having been directly selected for in evolution. Rather, the idea is that the innate physics system and the innate mind-reading system postulate states and events that appear to have incommensurable properties, making it hard for children (and adults) to integrate them into a single framework (hence the 'mind/body problem'). For example, all physical events are represented as occurring in some determinate place, whereas it forms no part of our ordinary conception of mental events like imagining a unicorn or thinking of one's mother that they should occur in some definite place within the body (McGinn, 1995). In contrast, the arguments that I shall give below for the innateness of a key aspect of Cartesian epistemology suggest that the latter *is* an adaptation.

My purpose in this paper is to explore the strength of the case that can be made in support of the innateness of the other main strand in Descartes' philosophy of mind, which is epistemological rather than ontological. This is his belief in the *self-transparency* of (key aspects

[1] There is a concern about these data that Bloom and colleagues hope to address in future work. For we know that infants expect agents to take variable paths to their goals, not always moving in a straight line. So the kids might just assume that the person got behind the second screen via some other route that wouldn't involve passing through the visible space between the first and the second screen. One crucial experiment, where this deflationary explanation wouldn't be available, would be to see whether or not infants expect people (like other physical objects) to fall under the force of gravity when not supported.

of) the mind.[2] Descartes' famously believed that our knowledge of our own mental events is more certain than any other knowledge, enabling it to serve as the premise in his notorious *cogito* argument ('I think, therefore I am'). And he also believed that in order to count as a *mental* event at all, the state in question should be immediately accessible to the subject. Hence the transparency thesis (as I intend it) should be understood as a conjunction of two distinct claims: *incorrigibility* ('If I believe that I am undergoing a given mental event, then so I am') and *self-intimation* ('If I am undergoing a given mental event, then I can immediately know that I am.') I propose to show that there are good reasons for thinking that a belief in the self-transparency of mind is innate — either forming an explicit component of our mind-reading faculty, or else being embedded implicitly in the structure of that faculty.

My argument, like Bloom's, will presuppose the innateness of at least some core aspects of our mind-reading abilities.[3] For only if mind-reading, in general, is innate will it be plausible to claim that the transparency thesis can be an innate part, or consequence, of that ability. This is, of course, a controversial assumption. It will be denied by those who endorse some or other variety of theorizing-theory, claiming that the beliefs that underwrite our mind-reading capacity are the product of hypothesis formation and testing, Bayesian statistical reasoning using directed causal graphs, or some other form of domain-general learning (Wellman, 1990; Gopnik and Melzoff, 1997; Gopnik and Schultz, 2004). While I think that the innateness of mind-reading is supported by general evolutionary considerations (Carruthers, 2006), by data from autistic individuals and other unusual cases (Baron-Cohen, 1995; Siegal and Surian, 2002), by the seemingly *very* early acquisition of key aspects of mind-reading (specifically, false-belief

[2] The self-transparency thesis doesn't extend to *all* aspects of the mind, I should stress. In particular, it doesn't apply to stored, inactive, states like memories and standing-state beliefs. For it is part of common sense that these can be hard to access. Nor does it apply to dispositional mental properties like irascibility and generosity. Rather, the claim is restricted to the set of mental *events* like seeing, hearing, thinking, judging, and deciding. (Note that this is the very same set of mental states that Descartes intended his use of the word 'cogito' to cover. See the translators' note, Descartes, 1970, pp. xlvii-xlviii.)

[3] What I actually think is that the mind-reading faculty consists of one or more innately structured *learning* mechanisms, which emerge in normal development under maturational constraints. For of course everyone should allow that learning takes place in development, both about mental states in general and about the minds of specific individuals. But it is learning that takes place within the parameters of an innate model of the mind, structured in terms of a set of innate core concepts.

understanding in the second year of life; Onishi and Baillargeon, 2005; Southgate et al., 2007; Surian et al., 2007), and by the good explanations that exist for why children below the age of about four should generally fail to display false-belief understanding in explicit tasks (specifically, the 'curse of knowledge'; Birch and Bloom, 2004), I shall not attempt to argue for this here. The assumption of an innate mind-reading faculty will be left as just that: an assumption.

It is important to note, however, that the claimed innateness of core aspects of mind-reading is consistent with the idea that *simulation* also plays an important role in the attribution of mental states to others. For many theorists have converged on the idea that the best account of our mind-reading capacities will be some or other form of simulation–theory mix (Botterill and Carruthers, 1999; Nichols and Stich, 2003; Goldman, 2006). It will be possible for us to claim, therefore, that mind-reading is subserved by an innately structured faculty interacting with other suppositional and reasoning systems in such a way as to generate and respond to simulations. What I do have to rule out, however, are simulationist theories that see mind-reading as grounded in introspective access to our own mental states, with the theoretical aspects of mind-reading being *learned* on that basis (Goldman, 2006). For this wouldn't leave any room for my assumed innateness claim.

Note that although my argument makes the same background assumptions as Bloom's, it can't take quite the same form. This is because no one has yet attempted to see whether young children conceive of the human mind as being transparent to itself. Nor has anyone done the necessary comparative research to see whether or not such a belief is universal to all peoples and cultures. (I shall make some remarks about how such research might be conducted in Section 7.) While there is suggestive evidence in these directions (see Section 5), no one has yet really been looking. On the upside, however, the innateness of the self-transparency assumption makes very good sense from an evolutionary perspective, as we will see in Section 4. So in this respect the argument can be significantly stronger than Bloom's.

Before we begin with the main discussion, however, something more needs to be said about the nature of innateness. Otherwise the overall thesis of this paper will be left unacceptably opaque. There have been a number of recent explications of the innateness concept. Ariew (1999) draws his inspiration from biology, arguing that a trait is innate when its development in ontogeny is strongly *canalized*, buffered against environmental variation. An innate trait is thus

one that appears at about the same point in the normal development of the organism across a wide range of variations in the environment. Samuels (2002, 2007), in contrast, restricts his attention to the role of the innateness concept in cognitive science. He argues that in this context, innate properties are those that are cognitively *basic* (admitting of no cognitive explanation), as well as emerging in the course of development that is normal for the genotype.

While there is much that can be — and has been — said about the respective merits of these and other approaches, the details need not concern us. The important points for our purposes are, first, that innate traits *aren't learned*. And second (since there isn't any reason to think that the self-transparency belief would be polymorphically distributed in the population), we should predict that a belief in self-transparency would, if innate, emerge in the course of development that is normal for the species (and not just for the individual genotype). So we should expect the trait to be universal.

It is worth emphasizing once more, however, that the innateness of a trait needn't imply that it is unchangeable. So the fact that I — and others — now deny the transparency thesis is perfectly consistent with the overall claim of this paper. And it will, in any case, always be possible for an innate (but unconscious) model of the mind to get overridden in its effects on behavior (without being changed or eradicated) by an explicit (conscious) culturally-acquired one. A good illustration from another domain might be the way in which our innate Aristotelian physics gets overridden, most of the time, by the results of our explicit Newtonian schooling, while being apt to reassert itself when we aren't paying attention (McCloskey, 1983).

2. Self-transparency and truth

Suppose that a belief in the self-transparency of mind turns out to be a human universal. Then this is just what the innateness thesis would predict. But how strong is the reverse argument, that the innateness thesis would therefore be warranted as *the best explanation* of such universality data? Some might claim that a lot will turn on whether or not the transparency thesis is *true* (or close enough to the truth), even given the assumption that the mind-reading faculty in general is innate. For it might be said that if the human mind (or some significant portion or aspect thereof) *is* transparent to itself, then we need no other explanation for why people should be so strongly inclined to *believe in* transparency — they believe in it because it is obviously true. This is the line of thought that I propose to evaluate in the present section, before I argue, in Section 3, that

the transparency thesis is, actually, deeply and radically false.

What I imagine, then, is an opponent who is inclined to be concessive about the discoveries of modern cognitive science. The opponent will concede, for example, that there are unconscious mental states (such as the states that occur in early vision, or during syntactic processing), and will thus allow that the self-intimation thesis is false for a significant class of mental states. The opponent might also concede that we can sometimes be in error when forming beliefs about even our conscious mental states, thus requiring us to give up on a strict incorrigibility thesis. But still, it might be claimed, the self-transparency thesis is *almost* true in respect of a large sub-set of our mental states (specifically, those occurrent events that belong to the familiar kinds postulated by folk psychology, such as percepts, judgments, and decisions); and this is all that is necessary to explain why a belief in transparency should come so naturally to us. For in respect of our experiences, judgments, imaginings, and decidings, it might be said that there are mechanisms in place that automatically and reliably (but not infallibly) make those states available for higher-order description and report.

Let me grant, for the moment, that there may exist mechanisms that give us reliable access to the occurrence of many types of mental state. How is it, however, that the approximate *truth* of the self-transparency thesis is supposed to explain the universality of *belief in* that thesis (as opposed to belief in the mental states concerning which the thesis is true)? For there are, of course, a great many truths about the world, about ourselves, and about our mode of access to the world that people don't end up believing. It might perhaps be suggested that the transparency of mind is itself one of those mental properties to which we would have transparent access. But this plainly won't work. It is mental *events* of the above familiar kinds that are supposed to be transparently accessible to us, not the causal processes through which those states are produced. Hence, that we occurrently *believe in* transparency of mind (under some description) might be allowed to be something that we have transparent access to, as would be the conscious mental states that we form our intuitive beliefs about. But that our access to these mental states is actually transparent surely wouldn't be.

It might be said in reply that we don't need *transparent* access to the processes that issue in our belief in the self-transparency of mind in order for the latter to stand in no special need of explanation. There just needs to be *some* significantly reliable process that would take us from the occurrence of a class of (approximately) transparently accessible mental states to a belief in

the self-transparent mind. Just as there is a reliable process that takes us from the fact that we are visually perceiving something to the knowledge *that* we are visually perceiving it, without us having much idea *how* we know that we are visually perceiving, so there might be a process that takes us from the occurrence of transparently accessible mental states to the knowledge that they are so accessible.

This analogy is a poor one, however, since there are many features of the *content* of visual percepts sufficient to cue us to the fact that they are visual rather than auditory, for example — such as that they contain representations of color and of a simultaneously-presented three dimensional layout. A better analogy would be the existence of some reliable method for giving us knowledge about the *processes* of visual perception, such as that they are, or aren't, inferential in character. But this is mysterious. What could such a method be like, without being tantamount to an innate belief in the inferential or non-inferential character of perception? And then likewise, I suggest, in the case of a universally-held belief in the self-transparent mind: even if that belief is true (or close enough to the truth), the best explanation for its existence (in the context of a broadly nativist account of mind-reading in general)[4] will be that it is either innate, or a direct consequence of innate features of the mind-reading faculty.

3. The mind is *not* transparent to itself

Carruthers (2006, forthcoming) defends an account of the location and connectivity of the mind-reading system within the overall architecture of the human mind which implies that the self-transparency thesis is radically erroneous. The key elements of the account are represented in Figure 1, which uses vision as its example of a perceptual system, and which incorporates the dual visual systems hypothesis of Milner and Goodale (1995). (There is good reason to believe that a similar bifurcation of function occurs within other sense modalities also.)

Insert Figure 1 about here

The existence of a distinct action-guiding visual system located dorsally in the parietal

[4] If mind-reading were a product of scientific theorizing, as Gopnik and Melzoff (1997) believe, then it might be possible to explain a universal belief in transparency of mind as resulting from an inference to the simplest explanation. See footnote 6 for further discussion.

lobes is now quite well established (Milner and Goodale, 1995; Jacob and Jeannerod, 2003; Glover, 2004). The outputs of this system are produced extremely swiftly for use in the on-line visual control of action, and are always inaccessible to consciousness. The visual system located ventrally in the temporal lobes, in contrast, is slower, and is used for object recognition as well as for belief formation and desire formation (think how the mere sight of a piece of chocolate cake can make one feel hungry), as well as being used for planning in relation to the perceived environment ('I'll go *that* way and pick up *that* one'). Its outputs (when attended to) are globally broadcast to a wide range of belief-forming systems (including the mind-reading system), desire-forming systems, and planning systems; and these globally broadcast outputs are always conscious (Baars, 1988, 1997, 2002; Dehaene and Naccache, 2001; Baars et al., 2003; Dehaene et al., 2003, 2006).

The dual-systems model provides us with a partial vindication of the transparency thesis. For the globally broadcast outputs of the temporal-lobe system will be accessible to the mind-reading faculty *inter alia*, and hence subjects will find it trivially easy to self-attribute those percepts (and also images, which utilize the same systems; Kosslyn, 1994; Kosslyn et al., 2006). A mind-reading system that possesses the appropriate concepts and understanding of perception, and which receives as input a globally broadcast visual percept of red, for example, will find it trivially easy to form the judgment, 'I am seeing something red.' (At least, this will be easy provided that the visual state in question has been partially conceptualized by other mental faculties, coming to the mind-reading system with the concept *red* already attached.)[5]

So in respect of this limited class of perceptual states (namely, the globally broadcast ones) something quite close to the transparency thesis is true: such states are immediately available for self-report, and those reports are likely to be highly reliable (if not outright

[5] As this example makes clear, the thesis that I shall discuss in a moment, that *judgments* aren't transparently accessible, requires important qualification. In particular, it should be restricted to non-perceptual judgments. According to Kosslyn (1994) and others, the initial outputs of the visual system interact with a variety of conceptual systems that deploy and manipulate perceptual templates, attempting to achieve a 'best match' with the incoming data. When this is achieved, the result is globally broadcast as part of the perceptual state itself. Hence we see an object *as* red or *as* a man or *as* bending over. Since this event can give rise immediately to a stored belief, it qualifies as a (perceptual) judgment. But since it will also be received as input by the mind-reading system (by virtue of being globally broadcast), it will also be introspectable. The thesis that judgments aren't transparently accessible should therefore be understood as being restricted to *non-perceptual* judgments.

incorrigible). But by the same token, however, the dual-systems model decisively *undermines* the self-transparency thesis in respect of a large class of perceptual states, namely those that are the unconscious outputs of the action-guiding perceptual systems (which in the case of vision are located in the parietal lobes). Hence the thesis that perceptual states *in general* are transparently available is radically false.

More importantly, the account represented in Figure 1 claims that the mind-reading system lacks direct access, not only to the *processes* that issue in novel beliefs, desires, and plans, but also to the ensuing events (especially judgments and decisions) themselves. (Note the absence of any arrows in Figure 1 back from the outputs of the conceptual systems to the mind-reading faculty.) Hence self-attributions of such states will *always* be a result of swift interpretative activity, utilizing perceptual input as data (including not only perceptions of the environment and the agent's own actions, but also patterns of attention, visual imagery, inner speech, and so forth). If this is correct, then there is no respect in which the self-transparency model is even approximately correct in respect of propositional-attitude mental events. On the contrary, such events are *never* immediately accessible to their subjects.

Carruthers (forthcoming) outlines a comprehensive argument in support of just this claim, drawing extensively on the work of Gazzaniga (1998), Wegner (2002), Wilson (2002), and other cognitive scientists, and reviewing a wide range of empirical data. These include such facts as the following. First, split-brain subjects who are induced to perform an action by information presented only to their right hemisphere will nevertheless confabulate an explanation (using their left hemisphere) with all of the seeming introspective obviousness as usual (Gazzaniga, 1995). Second, normal subjects who are induced to make a movement via magnetic stimulation of motor cortex (but who are ignorant of this fact) will claim to have been aware of *deciding* to make that movement (Brasil-Neto et al., 1992). Third, provided that they no longer recall having been hypnotized, subjects who follow instructions given to them while under hypnosis will also confabulate explanations, while seeming to themselves to be introspecting (Edwards, 1965; Sheehan and Orne, 1968). Fourth, subjects' sense that they had intended an outcome, which was in fact caused by another person, can be manipulated by the simple expedient of having a semantically-relevant stimulus presented to them shortly before the action itself (Wegner and Wheatley, 1999). And fifth, the social psychology literature on belief attribution is *rife* with studies demonstrating the effects of people's own behavior on the judgments that they will

mistakenly attribute to themselves (Eagly and Chaiken, 1993; Briñol and Petty 2003).

Carruthers (forthcoming) claims that the best explanation of these and other data is that people never have immediate access to their own (non-perceptual) judgments and decisions, but only ever know of such events via a swift process of self-interpretation (which remains unconscious, of course). Indeed, the resulting self-attributions are often confabulated and false. I shall not review that argument here. But let me just emphasize one point. This is that the empirical data are pretty decisive in showing that subjects themselves are unable to distinguish between confabulation and introspection. Hence subjects themselves can't tell whether or not mental states are transparently available to them (Gazzaniga, 1995).

I should also point out that the argument here doesn't turn on the discovery by cognitive science of mental states *in addition to* those postulated by our common-sense psychology, such as those involved in early visual processing, within motor planning systems, or in syntactic processing. On the contrary, the events that we turn out not to have transparent access to are (some of) those to which we pre-theoretically think we *should* have such access. Thus we normally think that the details of our physical movements are guided by the very same conscious visual percepts that inform our thoughts and planning. But this turns out not to be so. Rather, those movements are guided by the unconscious outputs of the parietal-lobe system. And likewise the judgments and decisions that we turn out to have merely interpretative (rather than transparent) access to are perfectly ordinary ones, of the sort that we think should be transparently accessible to us.

Suppose, then, that the case made by Carruthers (forthcoming) in support of the Figure 1 architecture is sound. And suppose, too, that it turns out that a belief in the self-transparency of mind is a human universal. This would then present us with the challenge of explaining *why* people should believe in the transparency of mind, given that this belief is so radically wrong. I submit that in such circumstances (and assuming the innateness of our mind-reading faculty), the best explanation would be that the belief is an innate part or consequence of the structure of our mind-reading system. And this in turn, I shall now argue, is something that we might have predicted in advance.

4. Why believing in transparency of mind is useful

Everyone now agrees that mind-reading is computationally very expensive. According to people

who work within the framework of a 'Machiavellian intelligence' hypothesis (Byrne and Whiten, 1988, 1997), it took extremely powerful adaptive pressures resulting from an 'evolutionary arms race' in social cognition to build our mind-reading capacity; and this is said to be the main engine driving the evolution of distinctively-human intelligence. Likewise, Dunbar (2000) argues that the need for increased group sizes drove the demand for increasingly sophisticated social cognition, driving up the need for computational resources (issuing in much-increased brain size) in nearly exponential fashion. This is not only because when group size increases one has to keep track of more individuals and their mental states, but also because one has to compute and store their attitudes to one another.

In a group of three people one has to compute what A thinks about B and C, what B thinks about A and C, and what C thinks about A and B (six sets of computations). But in a group of four, one has to compute what A thinks about B, C, and D, what B thinks about A, C, and D, what C thinks about A, B, and D, and what D thinks about A, B, and C (12 sets of computations). And so it goes (a group of five will require 20 sets of computations, and so on). And even this only really begins to scratch the surface of complexity. For it will often be important to figure out, not just what A thinks about B, but also what A thinks that B thinks about A (or C), as well as what A thinks that B thinks about what A thinks about B (or C), and so on and so forth.

Consistent with these points, and thirty years after Premack and Woodruff (1978) first raised the question whether our nearest relatives, the chimpanzees, are capable of mind-reading, there is an emerging consensus that *sophisticated* forms of mind-reading, at least, are a uniquely human adaptation. In fact, Povinelli (2000) and Povinelli and Vonk (2003) have argued forcefully that chimpanzees are merely extremely clever behaviorists — they are adept at computing and tracking, and drawing inferences from, the statistical relationships between observed behaviors, but they lack any conception of the mental states that lie behind those behaviors. And even those who have been vigorous in defending the (limited) mind-reading abilities of chimpanzees have been forced to concede that the latter's understanding may be confined to some aspects of perception and desire (hence not including belief and the possibility of false belief; Tomasello et al., 2003a, 2003b).

In order to be effective, the mind-reading system needs to contain some sort of model of the way that minds, in general, work. It needs to know that perception, while generally reliable,

can also be partial and misleading; it also needs to know that perceptions tend to give rise to beliefs, and also to trigger desires; it needs to know that beliefs can be false, and that desires can vary in strength and will often differ between individuals; it needs to know that beliefs and desires interact with one another in the construction of plans of action, and that the latter are then guided in their execution by perceptual states so as to issue in behavior; and so on, and so forth. But what does the mind-reading system need to represent about its *own* operations, and about its own access to the mental states of the agent? Any attempt to model its own interpretative activity would vastly complicate its computations, but without any significant increase in reliability (and perhaps with some decrement) — or so I shall now argue. On the contrary, the mind's model of its own access to itself should be a form of transparency thesis. This provides the evolutionary rationale for the existence of an innate belief in the self-transparency of minds.

One of the important tasks that the mind-reading system needs to perform is to assist in the interpretation of speech about mental states, specifically the speaker's own mental states. Humans spend a lot of time, in interactions with others, in talking about their own mental states. People talk about what they think, what they want, what they feel, and what they plan to do (as well as, more rarely, what they can presently see or hear). Such reports play a crucial role in the formation and maintenance of cooperative social relationships of many kinds, as well as being used in competitive ones. Yet if the model represented in Figure 1 is correct, all such reports of the speaker's propositional attitude events are the results of unconscious forms of self-interpretation, undertaken by the speaker's mind-reading faculty. If the mind-reading system of the hearer attempted to model this interpretative relationship, then its own task would become a great deal more complicated.

Suppose that someone says to me, in a particular context, 'I want to help you.' And consider the tasks that my mind-reading system faces in consequence. First, it must assist in the interpretation of this speech act, working together with the language faculty to figure out what the speaker means. (Is the utterance literal, or is it spoken in jest or irony? And what is meant by 'help' in this context? Does the speaker mean, 'help in general', or 'help in some specific task', or what? And so on. See Sperber and Wilson, 2002.) Then second, the mind-reading system must try to figure out whether the offer of assistance, thus interpreted, is sincere or not. To this many, many, bodies of evidence are relevant — including the tone of voice and facial expression with which the words are uttered, and the body language of the speaker; the past history of the

speaker's interactions with me, and with others; whether the speaker has anything to gain, in the circumstances, from an insincere offer of assistance; and so forth. These are amongst the most complex matters of judgment that we ever face. Yet we confront them routinely every day, and in most cases we have to reach a decision swiftly, without much time for contemplation.

Now suppose that the mind-reading faculty contained an accurate representation of its own interpretative access to the mental states of the same subject. In that case, in addition to the above tasks, it would also have to judge whether or not the speaker had interpreted his own desires correctly. This would add another whole layer of computational complexity, requiring many different sorts of evidence to be taken into account. Far better, surely, that the mind-reading system should model its own access to the mind of which it forms a part as entirely transparent — at least, provided that it can do so without too much loss of reliability in the judgments that it forms. (I shall return to this point in a moment.) And that will create a pressure for an innate belief in the self-transparency of mind to be added to the mind-reading faculty itself, or for such an assumption to be built implicitly, somehow, into the latter's structure.

Would an innate self-transparency assumption lead to any decrease in the mind-reading system's reliability? Considered purely in the abstract, the answer must be, 'Yes.' For that assumption would cause the system to miss out on any cases where subjects have misinterpreted themselves, since the transparency assumption leaves no room for such a possibility. In practice, however, there are two distinct reasons why the transparency assumption wouldn't lead to any decrease in reliability (and might actually lead to an increase). The first is that any expansion in the computational complexity of a system will introduce additional sources of error (as well as imposing a cost in terms of speed of processing, of course), as will any increase in the types of evidence that need to be sought. It is now a familiar point in cognitive science, not only that simple (but invalid) heuristics can prove remarkably reliable in practice, but that they can often out-compete fancier computational processes once the costs imposed by computational errors, as well as missing or misleading information, are factored in (Gigerenzer et al., 1999).

The second reason why the transparency assumption is unlikely to lead to a significant decrease in reliability comes in two parts. The first is that people are remarkably good interpreters of themselves. This means that in normal circumstances instances of confabulation will be rare, and hence any errors introduced by the existence of a transparency assumption will be few. And the second point is that even confabulated attributions of mental states are apt to

become, in a sense, self-verifying. Once people have articulated a belief about one of their mental states, then there are pressures on them of various sorts to constrain their behavior in accordance with the state so attributed, even if the initial attribution was confabulated. In effect, even an initially false self-attribution, once made, can become self-fulfilling (Frankish, 2004; Carruthers, 2006). Once someone has said to me, 'I want to help you', then this is no longer just a report of a desire, but will also be interpreted (by others as well as the speaker) as a sort of *commitment* (not necessarily a commitment to *do* anything in particular, but a commitment to having a desire to help). And then other desires and beliefs (the desire to keep one's commitments, the belief that one ought to act in such a way as to honor one's commitments) can lead the person to behave just as if they *did* want to help me, even if the initial self-attribution resulted from a misinterpretation.

Given these facts about the ways in which self-attributions of mental states are frequently self-fulfilling, a mind-reading system that allowed for mistaken self-attributions (i.e. which *didn't* operate with a transparency assumption), but which didn't allow for the self-fulfilling character of self-attribution, would probably be significantly less reliable than a simpler mind-reading system embodying a self-transparency assumption. But any attempt to take account of these new facts would introduce yet a fourth layer of complexity. In addition to assisting in the interpretation of speech, and judging the speaker's sincerity, the mind-reading system would also have to consider how likely it is, in the circumstances, that the speaker has misinterpreted his own mental states, as well as attempting to judge whether this is one of those cases where an attribution of a mental state to the self is likely to be self-fulfilling. Computational complexity indeed!

Let me stress, however, that my claim isn't that if the mind-reading system's model of the mind were enriched to include the interpretative character of its own access to propositional attitude events in the same subject, then that would render the mind-reading system's operations computationally *intractable*. Rather, it is that an enrichment of this sort would cause the mind-reading system to become slower and more computationally demanding, but without any significant gain in reliability (and probably with significant loss). Yet the sort of access that the mind-reading system has to the rest of the mind that houses it could hardly be something that it remained silent about — the question is too obvious, and too important for purposes of explaining and predicting behavior. (And note that in connection with all other types of belief we

have beliefs about the relationships that typically obtain between those beliefs and the facts that they concern, via perception, testimony, and so forth.) So a strong pressure is created for the self-transparency assumption to be built into the mind-reading system's model of the mind.

The upshot of these considerations is that, given the assumption of an innate mind-reading faculty, an innate belief in the self-transparency of minds is exactly what we should predict. In which case, the discovery that some such belief is a universal feature of human minds would serve to confirm that the belief is an innate one.[6] And already, even in advance of any such discovery, we have some reason for accepting the innateness of a self-transparency belief, based on considerations of reverse engineering. Some further reasons will be offered in the section that follows.

5. Some explanatory benefits of the innateness thesis

One positive virtue of the innateness hypothesis is that it can explain the near-ubiquity of belief in the self-transparency of mind in the Western philosophical tradition up to the ground-breaking (and science-inspired) writings of Sigmund Freud early in the twentieth century. Although I have labeled the self-transparency thesis 'Cartesian', in fact Descartes was by no means its only proponent. On the contrary, it was taken for granted by every philosophical writer that I know of. To give an example taken from an otherwise entirely different (Empiricist) philosophical tradition, Locke (1690) could write, without thinking that he needed to provide any supporting argument, and as if he were merely stating the obvious, 'There can be nothing within the mind that the mind itself is unaware of.' Likewise, Kant (1781) could write, 'It must be possible for the "I think" to accompany all my representations.'

The innateness hypothesis also provides us with an account of the pressures that shape

[6] We do still need the assumption of an innate mind-reading faculty at this point, in order to warrant the claimed innateness of the self-transparency belief. For a theorizing theorist could reach the same conclusion (predicting that a self-transparency belief will be universal), by arguing that a theory of mind that contains the self-transparency assumption will be a great deal simpler than one that tries to take account of self-interpretation and confabulation. For an increase in simplicity of this sort is likely to lead to the transparency thesis being accepted. This is especially true given the point that adding the extra complexities to the theory, necessary to accommodate self-interpretation, would be unlikely to lead to any significant gains in predictive or explanatory power, for the reasons given in the text. (Additional complexity is an additional source of error, and so forth.) For then anyone who postulates such a (more accurate but more complex) theory is unlikely to see it confirmed.

ordinary people's reactions to scientific theorizing about the mind. It is striking that many readers of Freud, for example — once they become convinced of the existence of unconscious mental states — tend to interpret him as proposing a sort of 'two minds' theory, rather than as describing the ways in which the two sets of states operate within a single mind. Each of these minds (the conscious mind and the unconscious mind) has its own principles of operation and characteristic goals. This enables people to preserve the self-transparency of the conscious mind (and perhaps also the unconscious mind), while giving up on it with respect to the mind as a whole.

Furthermore, many writers, from Popper (1959) onwards, have noted the close analogies between Freudian psychoanalytic theory and religious belief. And on the account of religious belief provided by Boyer (2001) this is explained, provided that we are allowed to assume the innateness of the self-transparency assumption. For according to Boyer, all religious beliefs have in common that they violate one of the core assumptions of one of our innate faculties while enabling us, nevertheless, to access the rich inferential potential of that or another faculty. (Consider, for example, a statue that listens to and responds to prayers. This violates one of the core assumptions of the *artifact* faculty while allowing us to deploy all of the inferential resources of the mind-reading system.) This is just what Freudian psychology does: it violates the self-transparency assumption for the mind as a whole, while allowing us to deploy our mind-reading faculty in reasoning about each of its components.

It is also noteworthy that many philosophers, when confronted with the evidence from cognitive science for the existence of unconscious mental states, are apt to couch the difference in terms of a distinction between 'personal' and 'sub-personal' mentality. Given an innate belief in the self-transparency of minds, this is explicable. For it enables these philosophers to hold on to that thesis in respect of *the person's* mental states. But in the absence of such a belief, the tendency is puzzling. For why, otherwise, shouldn't the unconscious experiences that guide the details of my movements, the unconscious judgments that inform and shape my planning, and the unconscious decisions that issue in many of my actions, count as *mine*? Granted, there is a perfectly respectable notion of 'sub-personal' that applies to information that gets deployed unconsciously *within* a belief-forming system, or *within* the visual system, for example. This might well deserve to be called a 'sub-personal belief' if it isn't available for use outside of that system. But what needs to be explained is a near-ubiquitous tendency amongst philosophers to

assimilate the conscious–unconscious distinction to the personal–sub-personal one. And this *can* be explained if we see it as an attempt to preserve the self-transparency assumption.

In similar vein, consider how almost everyone reacts when they hear of Libet's (1985) data (if they don't deny or otherwise attempt to undermine those data, that is). What Libet claims to have found is that the brain events that initiate action take place shortly *before* the agent makes a conscious decision to act. And almost everybody reacts by saying that the data show, if accepted, that *I* (the agent) don't really have control over my own actions. This assumes, of course, that *my* mental life consists in the set of states that are transparently accessible to me (hence the brain events that cause those movements — and the mental events that those brain events might realize — aren't really *mine*).

Finally, in this catalog of additional virtues of the innateness hypothesis, it can explain the *extremely* vigorous scientific resistance that greeted the very idea of unconscious perception. For example, although the phenomenon of 'blind-sight' was first described over thirty years ago (Sanders et al., 1974), it took more than two decades of careful and painstaking research before the majority of scientists could be brought to recognize the reality of the phenomenon. Weiskrantz (1997) describes how the initial discoveries were met with incredulity, and how the majority of researchers would seemingly prefer *any* hypothesis, no matter how implausible, rather than be forced to accept the existence of unconscious visual perceptions. This is easily and smoothly explained if the latter violated one of their 'core knowledge' assumptions.

There are a variety of reasons for taking seriously the hypothesis that we have an innate belief in the self-transparency of mind, then. Certainly the idea seems to be well enough motivated to justify moving on to the next stage: seeking evidence for the universality of the belief. (This will be discussed in Section 7.) For if found, this would provide decisive confirmation of the innateness hypothesis (given the background assumption of an innate mind-reading faculty, of course). Or so I have argued.

6. Learning theory revisited

Although for the most part I have taken the innateness of our mind-reading faculty for granted, it is worth asking whether the *intransigence* of the self-transparency belief (that is, its tendency to re-assert itself and to shape the thinking even of people who take themselves to have explicitly rejected it) counts against the competing theorizing theory. For while theorizing theory can

perhaps explain why the self-transparency assumption should be universal (assuming that it is), by appealing to the relative *simplicity* of theories, it would seem to have much more difficulty in explaining why that assumption should be so hard to modify. The best that it can do, here, is to appeal to depth of theoretical embedding, arguing that deeply embedded — as it were, 'over-learned' — beliefs are hard to change. But to make this work (since deeply embedded beliefs aren't *always* hard to change — think of religious conversions and de-conversions), theorizing-theorists would need to postulate a set of beliefs that are somehow immune to revision by normal processes of reflection, even though they are formed via a process very much like general-purpose theorizing. This suggestion lacks any independent motivation.

More importantly, someone might challenge the inference from the (supposed) universality of the self-transparency model of the mind to the innateness of that model, even given the innateness of mind-reading in general. For couldn't an innate mind-reading faculty *learn* (in the sense of acquiring justified beliefs, which is consistent with the falsity of those beliefs) that minds are transparent to themselves? The learning might proceed like this. Given that people are pretty good interpreters of themselves, let us suppose that, in general, when someone undergoes a propositional attitude event *E*, they thereby come to believe that *E* is occurring. If such beliefs are often enough verbalized and globally broadcast, then they would in turn become available to the mind-reading system, leading the system to believe that it believes that *E* is occurring. The system might then gradually acquire the generalization, 'Whenever I undergo an event *E*, I believe that I am undergoing *E*.' And since people in daily life are rarely if ever confronted with evidence that they have misinterpreted themselves, one might expect this to be further elaborated and generalized into a full self-transparency model.

One obvious difficulty with this suggestion is that it seems to require that most, if not all, of our higher-order beliefs about our own current judgments and decisions should get verbally expressed and globally broadcast in inner speech. For if only some of them are, generalization will only get us to, '*Sometimes* when I undergo an event *E*, I believe that I am undergoing *E*', which is far too weak to issue in a belief in self-transparency. But it seems very unlikely that this should be so. Relying on my own introspection, at any rate, I rarely find myself entertaining such sentences. And when people engage in 'think aloud' protocols (Ericsson and Simon, 1993), they mostly articulate only first-order thoughts about the problems in hand, not thoughts about their own occurrent thoughts.

The main difficulty with this and related proposals, however (such as that the mind-reading system routinely moves by semantic ascent from beliefs of the form, 'I am undergoing mental event *E*', to beliefs of the form, 'I believe that I am undergoing *E*') is that they beg the question at issue. For notice that an exactly parallel form of argument would be available in the case of other people. If we suppose that whenever my mind-reading faculty attributes to John a mental event *E* it arrives (by whatever route) at the belief that I believe that John is undergoing *E*, then I could similarly generalize to reach, 'Whenever John undergoes a mental event *E*, I believe that he undergoes *E*.' But it is obvious that the belief that John's mind is transparent to me wouldn't be warranted. For we have every reason to think that John will undergo many mental events that I never get to attribute to him. So why not, then, in my own case? I could only think that it would be appropriate to generalize along the lines sketched above, reaching a belief in the transparency of my own mind to myself, if I *already* believed that all my mental events are self-intimating.

Another way of emphasizing this point is to note that the critical move made by the learning theorist occurred at the outset. We were asked to suppose that whenever someone undergoes a propositional attitude event *E*, they thereby come to believe that *E* is occurring. But why should we suppose this? There are probably multiple belief-forming and decision-forming systems in the human mind that operate in parallel (Carruthers, 2006), issuing in judgments or decisions that I never have occasion to attribute to myself. And even if there aren't, by what right does the mind-reading system get to assume that there aren't, unless it already believes in the self-transparency of mind? So our conclusion stands: if the self-transparency belief is universal, then it is likely to be innate.

7. How to collect evidence of a universal belief in self-transparency

One source of evidence could be tapped by specialists in the world's different religious and philosophical traditions. If most of them, like Western philosophy until recently, were to endorse something like the self-transparency thesis, then that would count in favor of the universality, and hence innateness, of belief in the latter. (Remember, the fact that a belief is innate needn't mean that people's explicit theorizing can't lead to conflicts with that belief; so the existence of *some* philosophical traditions that deny self-transparency would nevertheless be consistent with the innateness of an intuitive, pre-theoretical, belief in the latter. What we should predict is just

that an innate belief in self-transparency would create a strong *pressure* towards explicit, reflective, theories of the same sort.) What I do want to stress here, however (as I did at the outset of the paper), is that *dispositional* mental states aren't covered by the self-transparency thesis. So the existence of intellectual traditions (such as Buddhism) that stress the difficulty of achieving self-knowledge won't count against the universality of belief in self-transparency, provided that the knowledge that is said to be difficult concerns such things as one's memories, character traits, and long-term needs.

Why *should* the transparency thesis be restricted to current mental events (experiences, feelings, and acts of thinking, judging, and deciding), however? For after all, it can be just as important to the success or failure of a cooperative endeavor to know what someone's character traits are as to know what they are currently thinking. A number of points are relevant here. One is that knowledge of mental events is logically prior to knowledge of traits. You can only know that people are generous, in the first instance (not by report), by induction from past instances in which they have, or haven't, helped others intentionally with no expectation of reward. So there couldn't be a mind-reading system that just identified character traits without first identifying mental events. Moreover, it seems that one doesn't need any special-purpose learning device to identify character traits, once you have mind-reading for mental events. Regular old general-purpose induction from past instances will do the trick. In addition (and closely related to the argument provided in Section 4), people rarely have occasion to report on their own character traits. And when they do, it is generally manifest that they *are* interpreting themselves — they may say things like, 'Looking back over the past couple of years, I think I can claim to be a generous sort of a person.' There is therefore no pressure to extend the self-transparency thesis to cover such cases.

The main evidence in support of the universality of the self-transparency thesis, however, should come from anthropology and developmental psychology. I propose to focus, for the most part, on anthropology — suggesting some simple experiments that could be conducted with adults cross-culturally — commenting on possible extensions to infants and young children in passing.

One obvious thing that one might do, is to probe people's intuitions about whether it is possible for someone to make a judgment or decision without knowing that they have done so. If they think that such a thing isn't possible, then this will be evidence that they conceive of such

events as being transparently available to the agent. (One could likewise devise probes to test whether people think that it is possible for someone to be mistaken when they do form a belief about what they are currently judging or deciding. But I shall leave this as an exercise to the reader.) One might, for example, present people with little vignettes containing a probe question, somewhat like this: 'Suppose that Mary is sitting in the next room. She is just now deciding to go to the well for water, but she doesn't know that she is deciding to go to the well for water. Is that possible?' For the control question, one could substitute another agent as the one who is ignorant of the subject's decision, as follows: 'Suppose that Mary is sitting in the next room. She is just now deciding to go to the well for water, but John doesn't know that she is deciding to go to the well for water. Is that possible?' If belief in the self-transparency thesis is universal, then we should predict large differences in the answers to these two sorts of question.[7]

There is, apparently, a significant problem with this proposal, however. This is that possibility-judgments are notoriously context-sensitive and subject to multiple interpretations. This is a difficulty that confronts comparative linguists on a regular basis. For in order to test hypotheses about the grammatical rules governing a language, field linguists will often want to ask informants questions of the form, 'Is it possible to say, "….", in your language?' And I gather that whenever a comparative researcher presents possibility-data from a new language, most of the ensuing discussion focuses on the methodology — exactly what was the question asked? How were subjects primed for the question? And so forth.

I am not convinced that this problem is insuperable, in part because possibility in language ('what it is possible to say') may be a somewhat special case. For ordinary people in the course of their daily lives rarely, if ever, make judgments concerning what it is possible to say. Since in the case of minds the intended sense of 'possible' is natural, or causal, possibility, in contrast (and since these judgments *do* play an important role in our lives, in practical reasoning), one might disambiguate the task by embedding it in a series of natural-possibility judgments. So one might first ask, for example, 'While Mary is looking at a rock, it starts to rise

[7] In an informal pilot study conducted with a handful of test subjects amongst the Shuar of Ecuadorian Amazonia, Clark Barrett (personal communication) asked just these questions, and found the predicted large differences. His subjects had no difficulty with the idea that one person might be in ignorance of another's decision, but regarded the suggestion that one might be in ignorance of one's own decisions as well-nigh unintelligible, or conceptually incoherent, just as Western subjects do.

up into the air although nothing has pushed it, nothing is tied to it, and so on. Is that possible?' Subjects who answer, 'Yes', to the probe question could either be excluded from the subsequent data, or the opportunity might be taken to disambiguate the intended meaning of 'possibility' — by saying, for example, 'No, I mean, is it possible *in the absence of magic*?'

If such experiments can be made to work reliably, then it is worth asking whether simple forms of them could be employed with young children. One crucial question is whether children have an adequate understanding of modal terms like 'possible'. Now, there is plenty of evidence that a capacity to reason about rules and obligations is an early emerging one, cross-culturally, issuing in judgments about what one *must* do and what one *can* do (Cummins, 1996; Harris and Núñez, 1996; Núñez and Harris, 1998; Harris *et al*., 2001). Three year-old and four year-old children are highly reliable in identifying cases where someone has broken a rule; and they are also very good at distinguishing between intentional and accidental non-compliance (categorizing only the former as 'naughty'). Moreover, they do this not only in connection with rules imposed by an authority (e.g. a parent or teacher), but also when reasoning about agreements negotiated between children themselves. And as one might expect, deontic *concepts* are acquired even earlier still. Psycholinguistic research shows that children as young as two years of age can make appropriate uses of modal terms like 'have to' and 'must' (Shatz and Wilcox, 1991). Unfortunately, however, when children are probed directly for their understanding of physical (causal) possibility and necessity, they seem incapable of distinguishing between something that is *impossible* and something that is merely *unlikely* (Shtulman and Carey, 2007). So I am not very optimistic about the prospects for making our test of self-transparency beliefs work with young children.

Another sort of strategy to test the universality of the self-transparency assumption would be to probe subjects for a belief in a strong self–other *asymmetry*, rather than probing directly for belief in the self-transparency of mind. What we would be asking, in effect, is whether subjects think that people's ways of knowing about their own minds are different in *kind* from their ways of knowing about the minds of others. One might present subjects with two different sorts of otherwise parallel vignette, a *self* version and an *other* version. The *self* version would take the form, 'John does A and B in situation C. John now knows that he himself believes / intends …' Then the *other* version would take the form, 'Paul does A and B in situation C. John now knows that Paul believes / intends …' And in each case the probe question would be, 'How do you

think John might know that?'

One would expect that in response to the *other* question subjects will say that John might have *seen* what Paul did, and *inferred* from that; or something of the sort. In response to the *self* question, on the other hand, subjects might say, 'He's aware of it', 'He just knows', 'It's *his* decision, surely', or (perhaps more likely) they might just give the questioner an incredulous stare. My prediction would be that subjects would *not* say, 'He perceives his own situation and/or behavior and interprets himself.' Reliable differences of this sort, across cultures, would be evidence that the self-transparency model is a human universal, I think. But I rather doubt that one could try the same sort of experiment with young children, who are notoriously bad at providing explanations.

8. Conclusion

I have argued that there are good reasons to take seriously the idea that a key component of Descartes' epistemological views is innately believed (specifically, his idea that minds are transparently accessible to themselves). None of my arguments has been demonstrative, of course. Each is, at best, a sound inference to the best explanation, and is therefore defeasible. But I hope that they strike readers as plausible enough to warrant further investigation. Such inquiry should focus especially on the question whether a belief in the self-transparency of mind is a human universal.[8]

References

Ariew, A. (1999). Innateness is canalization: in defense of a developmental account of innateness. In V. Hardcastle (ed.), *Where Biology Meets Psychology*, MIT Press.

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

Baars, B. (1997). *In the Theatre of Consciousness*. Oxford University Press.

Baars, B. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in*

[8] An early version of this paper was presented at a workshop of the *Culture and the Mind* project, supported by the Arts and Humanities Research Board of Great Britain, which was held in Lisbon in January 2007. I am grateful to the Director of the project, Stephen Laurence, for inviting me, and to all those who participated in the ensuing discussion. I am also grateful to Clark Barrett, Paul Bloom, and two anonymous referees for this journal for their helpful comments on an earlier draft.

Cognitive Science, 6, 47-52.

Baars, B., Ramsoy, T., and Laureys, S. (2003). Brain, consciousness, and the observing self. *Trends in Neurosciences*, 26, 671-675.

Baron-Cohen, S. (1995). *Mindblindness*. MIT Press.

Bering, J. and Bjorklund, D. (2004). The natural emergence of reasoning about the afterlife as a developmental regularity. *Developmental Psychology*, 40, 217-233.

Birch, S. and Bloom, P. (2004). Understanding children's and adult's limitations in mental state reasoning. *Trends in Cognitive Sciences*, 8, 255-260.

Bloom, P. (2004). *Descartes' Baby: how the science of child development explains what makes us human*. Basic Books.

Botterill, G. and Carruthers, P. (1999). *The Philosophy of Psychology*. Cambridge University Press.

Boyer, P. (2001). *Religion Explained: the evolutionary origins of religious thought*. Basic Books.

Brasil-Neto, J., Pascual-Leone, A., Valls-Solé, J., Cohen, L., and Hallett, M. (1992). Focal transcranial magnetic stimulation and response bias in a forced choice task. *Journal of Neurology, Neurosurgery, and Psychiatry*, 55, 964-966.

Briñol, P. and Petty, R. (2003). Overt head movements and persuasion: a self-validation analysis. *Journal of Personality and Social Psychology*, 84, 1123-1139.

Byrne, R. and Whiten, A., eds. (1988). *Machiavellian Intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press.

Byrne, R. and Whiten, A., eds. (1997). *Machiavellian Intelligence II: extensions and evaluations*. Cambridge University Press.

Carruthers, P. (2006). *The Architecture of the Mind: massive modularity and the flexibility of thought*. Oxford University Press.

Carruthers, P. (forthcoming). Introspection: divided and partly eliminated. *Philosophy and Phenomenological Research*.

Cummins, D. (1996). Evidence of deontic reasoning in 3- and 4-year-old children. *Memory and Cognition*, 24, 823-829.

Dehaene, S. and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79, 1-37.

Dehaene, S., Sergent, C., and Changeux, J. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Science*, 100, 8520-8525.

Dehaene, S., Changeux, J., Naccache, L., Sackur, J., and Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, 10, 204-211.

Descartes, R. (1970). *Philosophical Writings*. Edited and translated by E. Anscombe and P. Geach. Open University Press.

Dunbar, R. (2000). On the origin of the human mind. In P. Carruthers and A. Chamberlain (eds.), *Evolution and the Human Mind*, Cambridge University Press.

Eagly, A. and Chaiken, S. (1993). *The Psychology of Attitudes*. Harcourt Brace Jovanovich.

Edwards, G. (1965). Post-hypnotic amnesia and post-hypnotic effect. *British Journal of Psychiatry*, 111, 316-325.

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

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

Gazzaniga, M. (1995). Consciousness and the cerebral hemispheres. In M. Gazzaniga (ed.), *The Cognitive Neurosciences*, MIT Press.

Gazzaniga, M. (1998). *The Mind's Past*. California University Press.

Gigerenzer, G., Todd, P., and the ABC Research Group. (1999). *Simple Heuristics that Make Us Smart*. Oxford University Press.

Glover, S. (2004). Separate visual representations in the planning and control of action. *Behavioral and Brain Sciences*, 27, 3-24.

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

Gopnik, A. and Meltzoff, A. (1997). *Words, Thoughts, and Theories*. MIT Press.

Gopnik, A. and Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, 8, 371-377.

Harris, P. and Núñez, M. (1996). Understanding of permission rules by pre-school children. *Child Development*, 67, 1572-1591.

Harris, P., Núñez, M., and Brett, C. (2001). Let's swap: early understanding of social exchange

by British and Nepali children. *Memory and Cognition*, 29, 757-764.

Jacob, P. and Jeannerod, M. (2003). *Ways of Seeing*. Oxford University Press.

Kant, I. (1781). *The Critique of Pure Reason*. Many translations and editions now available.

Kosslyn, S. (1994). *Image and Brain*. MIT Press.

Kosslyn, S., Thompson, W., and Ganis, G. (2006). *The Case for Mental Imagery*. Oxford University Press.

Kuhlmeier, V., Bloom, P., and Wynn, K. (2004). Do 5-month-old infants see humans as material objects? *Cognition*, 94, 95-103.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8, 529-566.

Locke, J. (1690). *An Essay Concerning Human Understanding*. Many editions now available.

McCloskey, M. (1983). Naïve theories of motion. In D. Gentner and A. Stevens (eds.), *Mental Models*, Lawrence Erlbaum.

McGinn, C. (1995). Consciousness and space. *Journal of Consciousness Studies*, 2, 220-230.

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

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

Núñez, M. and Harris, P. (1998). Psychological and deontic concepts: separate domains or intimate connection? *Mind and Language*, 13, 153-170

Onishi, K. and Baillargeon, R. (2005). Do 15-month-olds understand false beliefs? *Science*, 5719, 255-258.

Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge.

Povinelli, D. (2000). *Folk Physics for Apes: The Chimpanzee's Theory of How the World Works*. Oxford University Press.

Povinelli, D. and Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, 7, 157-160.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 4, 515-526.

Samuels, R. (2002). Nativism in cognitive science. *Mind and Language*, 17, 233-265.

Samuels, R. (2007). Is innateness a confused concept? In P. Carruthers, S. Laurence, and S. Stich (eds.), *The Innate Mind: volume 3: foundations and the future*, Oxford University Press.

Sanders, M., Warrington, E., Marshall, J., and Weiskrantz, L. (1974). 'Blindsight': vision in a field defect. *Lancet*, April 1974, 707-708.

Shatz, M. and Wilcox, S. (1991). Constraints on the acquisition of English modals. In S. Gelman and J. Byrnes (eds.), *Perspectives on Thought and Language*, Cambridge University Press.

Sheehan, P. and Orne, M. (1968). Some comments on the nature of post-hypnotic behavior. *Journal of Nervous and Mental Disease*, 146, 209-220.

Shtulman, A. and Carey, S. (2007). Improbable or impossible? How children reason about the possibility of extraordinary events. *Child Development*, 78, 1015-1032.

Siegal, M. and Surian, L. (2006). Modularity in language and theory of mind: what is the evidence? In P. Carruthers, S. Laurence, and S. Stich (eds.), *The Innate Mind: volume 2: culture and cognition*, Oxford University Press.

Sperber, D. and Wilson, D. (2002). Pragmatics, modularity, and mind-reading. *Mind and Language*, 17, 3-23.

Southgate, V., Senju, A., and Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18, 587-592.

Surian, L., Caldi, S., and Sperber, D. (2007). Attribution of beliefs by 13-month old infants. *Psychological Science*, 18, 580-586.

Tomasello, M., Call, J., and Hare, B. (2003a). Chimpanzees understand psychological states – the question is which ones and to what extent. *Trends in Cognitive Sciences*, 7, 153-156.

Tomasello, M., Call, J., and Hare, B. (2003b). Chimpanzees versus humans: it's not that simple. *Trends in Cognitive Sciences*, 7, 239-210.

Weiskrantz, L. (1997). *Consciousness Lost and Found*. Oxford University Press.

Wegner, D. (2002). *The Illusion of Conscious Will*. MIT Press.

Wegner, D. and Wheatley, T. (1999). Apparent mental causation: sources of the experience of the will. *American Psychologist*, 54, 480-491.

Wellman, H. (1990). *The Child's Theory of Mind*. MIT Press.

Wilson, T. (2002). *Strangers to Ourselves*. Harvard University Press.

Figure 1: the place of mind-reading in the mind