7

Inner Sense Theories

The goal of this chapter is to explain and provide a preliminary evaluation of so-called "inner sense" accounts of self-knowledge, contrasting them with the interpretive sensory-access (ISA) theory.[1] There are a trio of such accounts to be considered. These are distinguished from one another by the varied relationships that they postulate between inner sense and other-directed mindreading. Some forms of evidence that are alleged to support one, or another, or all inner sense views will be discussed and evaluated here. Consideration of other relevant evidence will be deferred to later chapters.

1. Inner Sense and Mindreading: Three Theories

If one believes that there is a special faculty of inner sense for detecting our own mental states, then there are three possible accounts of the relationship between it and our mindreading capacity. First, it might be claimed that the two are realized in separate mechanisms that operate independently of one another, as Nichols and Stich (2003) maintain. Second, it might be claimed that the mindreading faculty has access to the outputs of inner sense, enabling the former to attribute mental states to the self in a transparent way (Frith and Happé, 1999; Happé, 2003). Or third, it might be said that the first-person knowledge obtained through the operations of inner sense is evolutionarily and developmentally basic, and that we are capable of attributing mental states to other people by utilizing this together with our imaginative and simulative abilities (Goldman, 2006). The present section will provide a preliminary discussion of each of these types of account in turn, abstracting as much as possible from the details of specific proposals.

Note that all three accounts (as well as the attitudinal working memory and mental-action theories discussed in Chapter 6) predict the occurrence of "unsymbolized" thinking. Since the channels of inner sense are supposed to give us transparent non-sensory access to our current thoughts, there should be many instances where people know themselves to be entertaining a specific thought in the absence of any relevant

[1] Recall that "inner sense" is actually a misnomer, inviting confusion with some form of interoception. In contrast with the latter, the outputs of inner sense are supposed to be intrinsically higher-order in content, representing the occurrence of our own mental states, as such.

sensory accompaniment (such as a sentence in inner speech). This prediction will be discussed in Section 4. It contrasts with one of the main predictions of the ISA theory laid out in Chapter 1.2.

1.1. Two Mechanisms

According to Nichols and Stich (2003), we possess two types of system for attributing mental states to ourselves and others. We have a complexly structured mindreading system for attributing mental states to other people (and also to ourselves in some circumstances). But we also possess a set of monitoring mechanisms for detecting and attributing mental states to ourselves. Nichols and Stich don't specify how many distinct mechanisms belong to this latter set, but they think that there must be at least one for detecting our own perceptual states and at least one for detecting our own propositional attitudes.

Given the structure of their account, Nichols and Stich must (and do) predict a two-way dissociation between capacities for self-knowledge and for other-knowledge. (Indeed, they also appear committed to a double dissociation between our capacities to detect our own experiences and our capacity to detect our own propositional attitudes.) Since the mechanisms involved are distinct, we should be able to find people who have lost their ability to attribute mental states to others while retaining their capacity to attribute mental states to themselves, as well as people who have lost the ability to attribute mental states to themselves while retaining a capacity to ascribe mental states to others. Nichols and Stich think that people with autism fall into the first category (intact self-attribution, damaged other-attribution), and that some forms of schizophrenia exemplify the second (intact other-attribution, damaged self-attribution). These and other claimed dissociations will be examined in Chapter 10. Recall from Chapter 1.2 that the ISA theory predicts, in contrast, that there should be no such dissociations.

In addition, Nichols and Stich (2003) maintain that capacities for self-knowledge should emerge significantly in advance of other-knowledge in infant development. Since the monitoring mechanisms are innate, while (they think) the development of mindreading depends importantly on learning, the former should be present quite early, whereas the latter (they think) emerges in stages over the first four years of life. The evidence that they provide in support of these claims will be examined in Section 2 of this chapter, and we will then return to the issue again in somewhat broader focus in Chapter 8. We noted in Chapter 1.2 that the ISA theory, in contrast, is committed to claiming that core competencies for self-knowledge and other-knowledge should emerge together in development.

It is important to note, however, that someone endorsing a two-mechanisms account is by no means forced to predict that self-knowledge will emerge in development in advance of other-knowledge. Nichols and Stich make this prediction because they happen to believe that competence in third-person mindreading requires significant amounts of learning (whereas the postulated monitoring mechanisms are innate).

But a two-mechanisms account could equally be combined with a more nativist perspective on the mindreading system. On this view, capacities for both mindreading and self-attribution would be early to emerge, and no prediction need be made that one would make its appearance in advance of the other.

A little reflection suggests that a separate *experience*-monitoring mechanism isn't necessary, moreover. For the mindreading faculty must be capable of receiving perceptual input. It will need to receive perceptual representations of the relations that obtain between target subjects and their environment, for example, on which it will need to effect various computations (such as tracking the subject's line of sight) to figure out what the subject is perceiving or wanting. But then if the mindreading system is already receiving the subject's own perceptual states as input, it should be capable of self-ascribing those experiences, as we saw in Chapter 3.5. The existence of a separate perception-monitoring mechanism is therefore unnecessary. This consideration doesn't rule out the existence of such a mechanism entirely, however. That will depend on the order of evolution (about which Nichols and Stich themselves remain silent). If the perception-monitoring mechanism evolved prior to the mindreading faculty, then it might have remained in place thereafter. In that case each of us would have two separate ways of monitoring and self-ascribing perceptual states. Since the existence of transparent access to our own perceptual states is not at stake in these discussions, however, I shall say nothing further about the alleged experience-monitoring mechanism in what follows.

One of the main arguments that Nichols and Stich (2003) offer in support of their mechanism for monitoring propositional attitudes is that it would be trivially easy to implement. They say that it just has to be capable of receiving as input any belief or any desire, and then of embedding the content of the state as a that-clause in a suitable self-ascription. For example, if the mechanism receives the representation, IT WILL RAIN SOON from the belief system, it just has to embed it to form the representation, I BELIEVE THAT IT WILL RAIN SOON. But this alleged simplicity is largely illusory, with the illusion stemming partly from the authors' failure to distinguish between standing-state propositional attitudes and occurrent, activated, ones, and partly from their apparent commitment to a form of functionalism in the philosophy of mind that is much too naive. Let me comment on the latter point first.

Nichols and Stich frame their accounts of mindreading and self-monitoring within a functionalist approach to the mind. I have no objection to that: my own commitments are thoroughly functionalist also. But when functionalism was first introduced into philosophy in the 1960s and 70s, it was widely assumed that functional organization might be quite independent of the physical organization of the brain. Although almost all functionalists were physicalists, most of them thought that there would be rampant multiple realization of mental systems in the brain. But the more scientists have learned about the relationship between mind and brain, the more cases of physical localization have been discovered, including a great many instances where micro-arrays of neurons possess quite specific functional and representational properties. There also exists

significant plasticity in brain development, of course, as well as variability between individuals in the precise physical realizations of cognitive systems. But for any particular cognitive function, the expectation is now that there is a single set of physical networks in any given brain that performs it. A belief-monitoring system would therefore need a physical channel of information from whatever brain systems realize belief to whatever mechanism subserves attributions of belief to oneself. A little reflection suggests that this informational channel would need to be complex in structure, and by no means trivially easy to implement.

It is widely accepted in cognitive science that memory fractionates into two distinct forms, with dissociable realizations in the brain: episodic and semantic. Furthermore, there is extensive evidence that semantic memory, in turn, fractionates into a number of different brain systems. Clinical evidence of dissociations suggests at least that memory systems for animate living things, inanimate living things (e.g. fruits and vegetables), artifacts/tools, and people/faces are all distinct from one another (Capitani et al., 2003; Mahon and Caramazza, 2003; Caramazza and Mahon, 2006). People can be impaired in their knowledge of any one of these kinds while being normal in their knowledge of the others. In addition, it is widely accepted as a general principle governing memory of all kinds that information tends to be stored where it is produced (Mayes and Roberts, 2002). Then since the evidence suggests that there are many, many, distinct information-producing systems in the human mind–brain (Gallistel, 1990; Carruthers, 2006a), we can expect that there will be many more distinct memory systems than we currently have clinical evidence of. If there is a monitoring mechanism latched onto the "belief system," therefore (as Nichols and Stich suggest), then its implementation will be anything but trivial. On the contrary, it will need to possess sophisticated search and retrieval mechanisms, since any normal human will have literally millions of stored beliefs at any one time, represented in a great many physically distinct systems.

A similar point holds if the monitoring mechanism is supposed to operate on activated beliefs, since there won't be any one system in which such events will occur. On the contrary, the principle that information is stored where it is produced suggests that activated forms of that information will initially be equally local, while also being sent to other inferential and executive systems when called for. So the monitoring mechanism in question would need to be built with access to the outputs of each of these distinct systems. And this is just in the domain of belief; something similar will be true in connection with desires, decisions, intentions, and so on. The result would seem to be anything but a simple mechanism. Rather, it will need to be designed with multiple connections, spreading its octopus-like tentacles into a great many distinct systems of the human mind–brain. Hence Nichols and Stich's simplicity-argument for the existence of propositional self-monitoring mechanisms carries little or no weight.[2]

[2] In fact there is an alternative way in which Nichols and Stich (2003) could have developed their "trivially easy" argument, but it is one that would have involved dropping their commitment to the existence of

In light of the points made above, it is plain that if one or more monitoring mechanisms exist, then they would need to have resulted from a powerful set of selection pressures, as we noted in Chapter 3.3. Brain tissue never comes for free; quite the contrary (Aiello and Wheeler, 1995). And a mechanism with any degree of complexity would need to be assembled and shaped over evolutionary time. In fact all inner sense accounts of self-knowledge make essentially the same evolutionary commitment. The extent to which they can make a plausible case for the evolutionary importance of transparent, non-interpretive, self-knowledge will be examined in Chapter 9. But to foreshadow, the idea deployed by all of these accounts is that inner sense evolved for purposes of cognitive self-monitoring and control. For, by monitoring our own ongoing learning, reasoning, and decision-making processes we can guide them and improve them (Shallice, 1988). So, all inner sense theories should predict that humans possess a robust form of metacognitive competence. Here, too, there is a conflict with one of the predictions of the ISA theory laid out in Chapter 1.2. The issue will be examined in Chapter 9.

1.2. Two Modes of Mindreading Access

Frith and Happé (1999) and Happé (2003) propose that there is just a single meta-representational faculty that subserves both other-directed mindreading and knowledge of one's own mind, but they suggest that this faculty has two distinct channels of access to the domain of mental states. It has both an outward-looking, perceptually grounded, channel of access (of the sort discussed at length in Chapter 3). But it also has an introspective channel used for self-monitoring. The view in question remains somewhat underdeveloped, however, and the authors' commitment to it appears tentative. (Indeed, a similar view can be discerned as a background assumption in Carruthers, 1996b.) Since no real arguments have been offered in its support, one suspects that it is tacitly motivated by intuitions of transparency of the sort discussed in Chapter 2. In consequence, our discussion will have to extrapolate somewhat and will be comparatively brief. Such views need to be addressed, however, because they are often mentioned by developmental scientists in conversation, especially by those whose view of the mindreading system is to some degree "modularist."[3]

self-monitoring mechanisms altogether. They could have bought into the idea of a purely attitudinal working memory system, of the sort discussed in Chapter 6.2. Since they already seem to believe in some such system embedded at the heart of their model of the mindreading faculty (see Chapter 8), this wouldn't come at much extra cost. However, they themselves think that only belief-like representations can figure in the working memory system in question, excluding desires, decisions, and intentions. Moreover, since the idea of propositional working memory faces multiple difficulties of its own (as we saw in Chapter 6.2), it is perhaps just as well that Nichols and Stich do *not* develop their transparent-access account of self-knowledge in this way.

[3] In the present context the term "module" needs to be taken somewhat more strongly than the weak notion introduced in Chapter 3.1. In particular, modular theories of mindreading maintain that the system in question is innately channeled and early developing.

Frith and Happé's account has one immediate powerful advantage over Nichols and Stich's (2003) model, however. This is that it postulates just a single set of concepts, and a single body of "core knowledge" of the mind, realized in just one metarepresentational faculty.[4] On Nichols and Stich's account, in contrast, it is puzzling how the two capacities are supposed to be related to one another. When infants monitor and identify their own attitudes, for example, at a time before mindreading has fully developed, what do they identify them *as*? Are we to imagine that the infant already has a concept of belief, but that this is encapsulated from the mindreading system? Does the infant then acquire a separate concept of belief as the mindreading faculty develops?

On a semantic (or referential) level, what Nichols and Stich should probably say is that there is a single type of representation (BELIEF, as it might be) that is initially produced only by the self-monitoring mechanism, but which is later also produced by the mindreading system. This single type of representation tracks the very same kind of state (namely beliefs) no matter how it is produced. So far so good. But it remains mysterious how two distinct innately channeled systems can access the same set of conceptual representations, and how elaborations produced by learning within one system should also be available within the other. What insures that the same representations are available to each, and what keeps them aligned? In contrast, these difficulties fall away on Frith and Happé's account.

Moreover, the two forms of theory make different predictions. We have seen that Nichols and Stich (2003) are committed to the emergence of self-knowledge in advance of other-knowledge in the course of child development (although a two-mechanisms account as such is not). Since Frith and Happé (1999) claim that there is a single metarepresentational competence underlying both self-knowledge and other-knowledge, in contrast, they should predict no significant differences in development. (In this respect their predictions are the same as those made by the ISA theory.) Either self-knowledge and other-knowledge will both emerge early (with heavy innate channeling, perhaps via the maturation of a modular system of some sort), or both will emerge late (perhaps resulting from a process of learning and theorizing). In fact Frith and Happé incline towards the former view; but the latter is equally a theoretical option.[5]

The present proposal also predicts a different pattern of dissociation from that predicted by a two-mechanisms account. It should be possible to find people in whom self-knowledge is damaged (because the introspective channel of information is disrupted or broken) but in whom other-knowledge is normal. But anyone capable

[4] On the idea of core knowledge, together with accounts of a number of different core systems, see Spelke and Kinzler (2007).

[5] Recall from Chapter 1.2, however, that the claim that there is a single metarepresentational *competence* underlying both forms of knowledge doesn't necessarily imply that subjects' *performance* should be equivalent on all types of task. For even the most ardent believer in innate faculties will allow that learning has an important role to play in development. And some learning might initially take place in the first-person, or in the third, issuing in a critical (albeit brief) period of time when children's performance is asymmetric.

of self-knowledge (whose metarepresentational faculty is therefore intact) will also be capable of other-knowledge, unless they suffer from completely debilitating multimodal perceptual failure.[6] (Multimodal perceptual failure, in contrast, will disrupt a great many other forms of cognition as well, and not just knowledge of others' mental states.) The discovery of a double dissociation, therefore, would support Nichols and Stich's model over that of Frith and Happé. These issues will be examined in Chapter 10.

Note that Frith and Happé, like Nichols and Stich, need make no specific predictions regarding the order of evolution of first-person and third-person capacities. It may be that metarepresentational capacities initially evolved in the first-person, with mindreading being added later; or it may be that mindreading was the first to emerge, with the mechanisms of inner sense evolving subsequently. But also like Nichols and Stich, Frith and Happé must claim that there was some strong or persistent selection pressure that resulted in inner sense—presumably involving the benefits of monitoring for control of our own cognitive processes. So they, too, should predict that we have robust forms of metacognitive competence.

1.3. Introspection-Based Mindreading

Goldman (2006) not only claims that humans possess a special channel of introspective access to their own attitude states, but that attributions of mental states to others are grounded in such first-person awareness. On this view, we know of the mental states of others through *simulation* of their perspective on the world and their thinking and reasoning abilities. The final step in each episode of mindreading is to identify the mental state in oneself with which the simulative process has concluded, and then to attribute that mental state to the other person.

Goldman thinks that a significant amount of learning needs to take place before children can become capable of more than the very simplest forms of simulation. This learning will be of two sorts, one pertaining to prediction and one to explanation. Consider the latter first. Simulation theorists maintain that when we observe an action of any sort we must entertain some suppositional beliefs and desires in our own minds, use our own reasoning and decision-making abilities with those materials, and see whether the upshot matches the target behavior. If it does, then we have our explanation; if it doesn't, then we need to try another set of supposed attitudes. This is what Goldman calls the "generate and test" procedure for behavior explanation. Since there are indefinitely many candidate beliefs and desires that one *could* adopt as possible explanations of the other person's behavior, it is plain that significant amounts

[6] Subjects might also suffer from more local mindreading-related perceptual failures, of course, such as an inability to recognize faces or facial expressions, or an inability to perceive biological motion. Such failures would impact one's knowledge of others' mental states in a limited way, while self-knowledge could remain fully intact.

of learning will need to take place first, to constrain the hypothesis space down to manageable proportions.

Now consider prediction. In order to predict what someone will do using a simulation strategy one needs to begin by attributing some mental states to them, which one can then adopt for oneself in suppositional reasoning mode to see what behavior might result. But these attributions can't generally themselves be a result of current simulation. Rather, one will need to have acquired, on the basis of previous simulations, some generalizations about what people in certain circumstances are likely to want or think.

In contrast with third-person mindreading, Goldman holds that our introspective abilities are innate. He therefore predicts that capacities for self-knowledge should emerge in development some considerable time before other-knowledge. Indeed, he cites in support essentially the same studies and sources as do Nichols and Stich (2003). These data will be examined in Section 2. Notice, however, that this prediction isn't optional for an introspection-based simulation theorist, in the way that it is for two-mechanism theorists. For each of the commitments in question—to innate introspective abilities, on the one hand, and to the need for significant amounts of learning to develop mindreading capacities, on the other—would seem to be mandatory for a simulation theorist.

Moreover, Goldman, too, must make a distinctive prediction about patterns of dissociation. On an introspection-based simulation account, a collapse in self-knowledge must lead to an equivalent collapse in other-knowledge. Hence there should be no cases where self-knowledge is damaged but mindreading is intact. But the converse dissociation is predicted to occur. There should be people in whom self-knowledge is intact but who are incapable of reading the minds of others because they lack the relevant simulative and imaginative capacities. This prediction will be evaluated in Chapter 10. Goldman (like Nichols and Stich) thinks that people with autistic-spectrum disorders fit this description.

In addition, Goldman, like other inner sense theorists, must claim that some significant selection pressure operated on our ancestors to enable the mechanisms of inner sense to evolve. And he, too, needs to suggest that the benefits of inner sense derive from a capacity to monitor, intervene in, and partly control our own cognitive processes. But unlike Nichols and Stich (2003) and Frith and Happé (1999), he is committed to a particular evolutionary *order* in which our capacities for self-knowledge and for other-knowledge should have emerged. Since the latter depends upon the former, inner sense must have evolved first. There must, therefore, have once existed creatures which were able to know their own mental states but were incapable of knowing the mental states of others. Indeed, such creatures might still exist, perhaps among monkeys and apes. Goldman isn't committed to claiming that they do, however. For it may be that the entire evolutionary process took place within the hominin line, and the predicted self-knowing but otherwise mind-blind creatures might all belong to extinct species of *Homo*. But if it *were* to be demonstrated that

there are animals who can monitor and entertain thoughts about their own mental states while being incapable of thinking about the mental states of others (either in general of for some specific class of mental states like beliefs), then this would lend significant support to Goldman's introspection-based view of mindreading. This issue will be addressed in Chapter 9.

Most of Goldman's (2006) arguments in support of his simulationist account are negative. He argues *against* pure forms of the alternative theories of mindreading (in particular, theorizing-theory of the sort proposed by Gopnik and Meltzoff, 1997, and modularity approaches of the kind endorsed by Scholl and Leslie, 1999). But such arguments are now moot. For, almost everyone now accepts the important role that simulation can play in certain forms of mindreading. Certainly the account that I shall endorse in Chapter 8 accepts it. But this is simulation that need presuppose no special faculty of introspection, but only globally broadcast sensory representations of one sort or another (in whose existence we already have reason to believe, of course). It is therefore fully consistent with the ISA theory of self-knowledge. Indeed, Goldman's own account of what he calls "high-level simulation" requires that the initiating representations should be imagistic in character. So this aspect of his theory is consistent with the ISA account. (In contrast, the "Possible Worlds Box" proposed by Nichols and Stich appears to be purely propositional in nature, as we will see in Chapter 8.)

Goldman does, however, present one positive argument in support of his introspection-based simulation theory from the existence of what he calls "low-level simulation." He cites extensive data on emotional mirroring in people. The sight of someone in pain causes one to feel pain, the sight of someone disgusted makes one feel disgusted, the sight of someone afraid makes one feel afraid, and so on. Moreover, the data seem to suggest that the mirroring emotions play an important role in recognizing those same emotions in others. For people who are damaged in their capacity to *feel* fear have difficulty in *recognizing* fear in the face of another person, for example. This suggests that emotional forms of mindreading, at least, might be grounded in introspective access to one's own emotional attitudes. This argument will be discussed and evaluated in Section 3.

1.4. Inner Sense Theories and Dual-Method Theories

Recall from Chapter 1.2 that the ISA theory predicts that people should make numerous errors when attributing attitudes to themselves. Since self-attribution is done by the same mental faculty that does other-attribution, and since it relies on the same interpretive principles and many of the same forms of evidence, there should be similar sorts of error in each case. As we will see in Chapter 11, this prediction is richly confirmed.

Inner sense theories, as such, make no specific predictions regarding errors of self-attribution, beyond saying that mistakes may result whenever the inner sense mechanism breaks down or malfunctions. And on the contrary, if inner sense is like our other senses, then errors should be comparatively few. Inner sense theorists recognize that

this is a problem (as does at least one of the defenders of attitudinal working memory; see Rey, forthcoming). This is because evidence of confabulation with respect to one's own propositional attitudes is now quite robust. In consequence, all embrace what Goldman (2006) calls a "dual-method" theory. On this account, we attribute attitudes to ourselves in two quite different ways. Sometimes we rely on inner sense, and our attitudes are thereby reliably and transparently accessible. But sometimes we rely on interpretation of sensorily-accessible data, just as the ISA theory maintains. In such cases we often make mistakes. But since the process of interpretation is swift and unconscious, we aren't generally *aware* that we employ two different methods. On the contrary, confabulated self-attributions will seem to their subjects to be just as transparent as those that rely on inner sense.

It is important to realize that the dialectical landscape of the debate between the ISA theory and dual-method accounts is heavily tilted towards the former. This is because dual-method theories have to buy into everything that ISA theorists believe, with the exception only of the scope of application of ISA. In particular, they have to accept that our mindreading abilities are often turned on ourselves, and that in such cases the mindreading faculty arrives at its interpretations smoothly and unconsciously, so that subjects nevertheless have the impression that they are merely introspecting. The onus is therefore on dual-method theorists to demonstrate that these resources aren't sufficient to explain *all* cases in which we unhesitatingly attribute propositional attitudes to ourselves. For the additional complexity introduced by postulating mechanisms of inner sense (or a non-sensory working memory system) needs to be justified.

Moreover, dual-method theories face an urgent challenge. This is to specify the circumstances in which the two methods are used. Until this is done, no predictions can be made, and consequently none of the data can be explained. If a dual-method theory says no more than that we sometimes rely on inner sense and sometimes on self-directed mindreading, then there is no way to predict the circumstances in which errors are likely to be found. And in respect of any particular set of experimental results showing confabulation effects, the best that one would be able to say is that these are among those instances where people rely on mindreading-based self-interpretation. This is not, of course, an *explanation* of the data. For that, we would need a theory that, when combined with auxiliary assumptions, is capable of *predicting* the data. And that plainly isn't the case here. In Chapter 11 we will consider various ways in which dual-method theories might be elaborated in an attempt to overcome this problem. For now it can be left as an open question whether inner sense theories (as well as attitudinal working memory theories) can be adapted in such a way as to accommodate the confabulation data successfully.

1.5. Summary

The predictions of the three forms of inner sense theory can be seen displayed in Table 7.1, where they are contrasted with the predictions of the ISA account. These predictions will be put to the test in our subsequent discussions.

Table 7.1. Comparing Inner Sense and ISA

Prediction / Theory	Introspection: † Awareness of unsymbolized thinking	Dissociations: Self-k damaged, other-k intact	Dissociations: Other-k damaged, self-k intact	Development: Self-k before other-k	Evolution: Self-k before other-k	Metacognition: †† Strong native competence	Confabulation: ††† Systematic self-attribution errors
Two Mechanisms	Yes	Yes	Yes	—⋆	—	Yes	? ⋆⋆
One Mechanism, Two Channels	Yes	Yes	No	No	—	Yes	? ⋆⋆
Introspection-Based	Yes	No	Yes	Yes	Yes	Yes	? ⋆⋆
ISA Theory	No	No	No	No	No	No	Yes

† Note that in addition to all three inner sense accounts, attitudinal working memory and mental-action theories, too, predict the existence of unsymbolized thinking.

†† Inner sense theories as such don't make a prediction of strong native metacognitive competence. Rather, the prediction derives from a broader (adaptationist) construal of the monitoring-for-control functions of inner sense.

††† Note that all transparent-access theories (with the exception of constitutive-authority accounts) may have difficulties accounting for the confabulation data. The problem isn't confined to inner sense theories.

⋆ Nichols and Stich (2003) predict self-knowledge before other-knowledge; but two-mechanisms accounts as such make no commitment on the issue.

⋆⋆ Inner sense theories as such don't predict confabulation, but expanded into dual-method theories they do. It is an open question at this point whether dual-method theories can predict and explain the *patterning* of people's self-attribution errors.

2. Developmental Evidence

We noted in Section 1 that two of the three forms of inner sense theory predict that children's competence in attributing mental states to themselves should emerge significantly in advance of their capacities to attribute mental states to other people. However, only one of the three theories *must* make such a prediction. This is Goldman's (2006) introspection-based simulation account. In contrast, Frith and Happé's (1999) theory postulates just a single core competence underlying both self-knowledge and other-knowledge. And a two-mechanisms theorist who takes a modular view of mindreading could likewise claim that both sets of capacities will be early-emerging ones. In consequence, if parallelism for self-knowledge and other-knowledge in development were to be demonstrated, this would only rule out one variety of inner sense theory. The issue is nonetheless important for us because the ISA theory, too, is committed to parallelism in development. So if it were to emerge, in contrast, that competence for self-knowledge is developmentally prior to competence for other-knowledge, then this would be an important strike against the ISA theory.

The present section will focus on the arguments of Nichols and Stich (2003). This is because Goldman (2006) merely reiterates some of their arguments and defers to their discussion, describing it as resulting from "a comprehensive analysis of the literature" (p. 236). As we will see, this evaluation is exaggerated.

2.1. Evidence of Self-Knowledge before Other-Knowledge

Nichols and Stich (2003) cite just three pieces of evidence in support of their view that self-knowledge emerges in advance of other-knowledge. The first is a study by Wimmer et al. (1988), which explored children's understanding of their own and other people's knowledge-states. The results seem to show that young children have awareness of their own knowledge before they have awareness of the knowledge of other people. But in fact the study in question admits of an alternative explanation, as I shall show.

In the "self" condition, the children were first allowed to look, or not look, into a box, and were then asked whether they know what is in the box. In the "other" condition, in contrast, they observed another subject either looking, or not looking, into the box before being asked whether the subject knows what is in the box. Answering the question in the "other" condition requires children to reason appropriately from the generalization that seeing leads to knowing (or something similar). But answering the question in the "self" condition requires no such thing. The children can answer merely by accessing, or by failing to access, their knowledge of what is in the box. They can substitute a first-order question in place of the second-order question asked—namely, "What *is* in the box?"—and answer "Yes," that they do know what is in the box, if an answer comes to mind, otherwise answering "No." [7] Since the two

[7] Note the parallelism between this explanation and the account offered in Chapter 4.2 of how even adults will characteristically generate sentences of the form, "I believe that P." People first address a first-order

conditions used in this experiment aren't genuinely similar, therefore, it provides no support for the view that children's competence in self-attribution emerges in advance of their competence in other-attribution.

The second item of evidence cited by Nichols and Stich (2003) involves a contrast between a pair of studies of children's understanding of pretense. In one of these, Gopnik and Slaughter (1991) found that three-year-old children were easily able to say what they had recently pretended to be the case. The children were shown an empty glass and were asked to pretend that it had orange juice in it. Then the glass was turned over (to tip out the juice), and they were asked to pretend that it now contained hot chocolate. Thereafter they were asked, "When I first asked you . . . What did you pretend was in the glass then?" Children performed near ceiling in this task. In contrast, Rosen et al. (1997) had three-year-old children watch a television show in which the characters were sitting on a bench but pretending to be on an airplane. The experimenters then asked, "Are they thinking about being on an airplane or about sitting on a bench outside their school?" Around 90 per cent of the children answered that the characters were thinking about sitting on a bench. The moral, Nichols and Stich believe, is that three-year-olds have no difficulty in attributing pretense to themselves while having considerable difficulty in attributing pretense to other people.

One problem with this argument is that these are, of course, two quite different tasks, undertaken with two distinct sets of children. This should already prevent us from drawing any firm conclusions. (It is quite common in the developmental literature for two groups of children to show differences in their average level of performance.) Moreover, in the study by Rosen and colleagues the children were asked a rather odd question. Instead of being asked what the characters were pretending, they were asked what they were thinking about. If we suppose that the children were fully competent with the concept of pretense, then they might naturally have wondered why they had been asked about thinking when the most salient fact was that the characters were pretending. This might have suggested to them that a contrast was implied, and that they were really being asked what the characters were thinking *as opposed to pretending*. In addition, that *something* is wrong with Nichols and Stich's argument is demonstrated by the finding that even two-year-old children or younger can understand when someone else is pretending, at about the same age that they begin pretending for the first time for themselves (Bosco et al., 2006; Rakoczy and Tomasello, 2006; Onishi et al., 2007).

The third item of developmental evidence cited by Nichols and Stich (2003) involves complex forms of perspective taking. Gopnik and Slaughter (1991) showed children pictures that would look very different when viewed from different perspectives. For example, they might be shown a picture of a turtle that looks as if it is lying

question ("*P*?") to their memory systems, linguistically encoding the answer, "*P*," if a suitable entry is found, before attaching, "I think that . . . " or, "I believe that . . . ," often as a mere stylistic convenience. There seems no reason to expect that children would not follow the same procedure.

on its back from one perspective, but which looks as if it is standing on its feet when seen from the opposite perspective. Three-year-old children mostly failed in the "other" version of this task. When presented with the picture and asked how it would look to the person sitting opposite them, they mostly gave answers from their own perspective. In contrast, most three-year-olds had little difficulty with the "self" version of the task. In this condition, they were first shown the picture of the turtle from one perspective (on its back, say), and were then moved around the table to view it from the other perspective. When they were then asked how the turtle had looked to them previously, before they had traded seats, they mostly answered correctly.

There is an important difference between these two versions of the task, however, which Nichols and Stich don't notice, but which should prevent us from drawing the conclusion they want. For in the "self" version of the task, the child just has to *recall* how the turtle had looked a few moments previously. The "other" version of the task, in contrast, can only be solved by generating and mentally rotating an image. The child has to create an image of the object as it is currently presented and then mentally rotate it through 180 degrees to see how it would appear from the other perspective. This is plainly a *much* harder task. Children's failures in the third-person version of the task, therefore, might have reflected this comparatively greater difficulty, rather than differences in their competence for mental state attribution.

I conclude, therefore, that none of the evidence that Nichols and Stich (2003) cite genuinely supports the view that self-knowledge emerges in development in advance of other-knowledge. The relevant forms of inner sense theory are thus unsupported. And by the same token, no problems are raised here for the ISA account.

2.2. Critiquing Evidence of Parallel Development

In responding to evidence that counts against their model, Nichols and Stich (2003) discuss at some length a single study suggesting that children's competence with the notion of belief emerges at the same age for self and for other. This is work by Gopnik and Astington (1988), who used a version of the now-famous "Smarties task." Children were shown a familiar type of candy box and asked what they thought it contained. Naturally enough, they answered, "Candy." The box was then opened to reveal that it actually contained pencils. In the "other" version of the task, the children were then asked what a friend would think was in the box when she was first shown it. In the "self" version of the task, the children are asked what they themselves had *thought* was in the box when *they* were first shown it. Responses to both versions of the task were strongly correlated, with the three-year-old children for the most part failing each, and with four-year-olds generally passing both.

Nichols and Stich say that on their view the three-year-old children should have been able to remember their previously-monitored thoughts, and should thus have been able to solve the task. For the occurrence of those thoughts would have been detected and represented by the self-monitoring mechanism. And since the children were asked the question just a minute or two later, the thoughts would be unlikely to

have been forgotten. (In Section 2.3 I shall suggest reasons why an inner sense theorist should reject this claim.) So the fact that children seem unable to answer on the basis of their memory presents something of a puzzle. Nichols and Stich feel obliged to conclude that in this case the children's answers were driven by the output of their mindreading faculty (the same faculty that delivers the answer in the "other" condition), rather than by the output of inner sense. Their immediate challenge is to explain why. Why should children choose to mindread themselves and answer on that basis, rather than simply remembering?

The explanation that Nichols and Stich (2003) offer is that the memory trace produced by the self-monitoring mechanism may have been comparatively weak. This appears quite arbitrary, however. No reason is given for *why* the introspective memory trace should be weaker than others, shortly after the fact, nor for why the child should choose to ignore it even if it were. (Given the special reliability that is supposed to attach to inner sense, one might think that children would utilize even a weak memory resulting from the latter in preference to the output of the mindreading system.) However, in support of this interpretation Nichols and Stich cite some unpublished data collected by German and Leslie. This involved both a "self" and an "other" version of a false belief task, in which the children either watched a video of another child looking in the wrong place for an object that had been moved, or in which they watched a video of their own earlier attempt, when they themselves had searched while entertaining a false belief. Children were significantly better at attributing a false belief to themselves than to the other child. Nichols and Stich interpret this as showing that once suitable memory supports are provided to children, then they are able to access and rely on their memory of their own introspected false belief.

Although the data in question are cited as "forthcoming" in Nichols and Stich (2003), they have not been published in the years that have elapsed since then. This is because when the experiment was replicated using a different task the effect was much weaker, and the investigators didn't pursue the matter (Tamsin German, personal communication). Even if we set this aside, however, and suppose that the data are robust, it is quite unclear that they are best explained by the claim that introspection develops in advance of mindreading. For one would expect that watching oneself conduct the task would evoke a good many first-order memories formed at the time, which would make interpreting one's own behavior easier than interpreting the behavior of the other child. These might be memories of the content of one's perceptual experience while one searched, for example, or of the content of the memory image that one had in mind of the target object placed at its original location. Although these memories would be purely first-order in character, they would provide significant support for the mindreading faculty to interpret one's behavior when one searches in the wrong location.

Much more significant than the weakness of Nichols and Stich's treatment of this one set of experiments, however, is that an extensive number of studies have failed to find any evidence of asymmetry in development between self-knowledge and other-

knowledge. Indeed, Wellman et al. (2001) conducted a meta-analysis of well over 100 pairs of experiments in which children had been asked, both to ascribe a false belief to another person, and (in otherwise parallel circumstances) to attribute a previous false belief to themselves. They were able to find no significant difference in performance, even at the youngest ages tested. This would appear to count heavily against the claim that self-knowledge emerges in children substantially in advance of other-knowledge. However, there are reasons why inner sense theorists should doubt the relevance of much of this data (as well as doubting the relevance of the studies on which Nichols and Stich themselves rely). Or so Section 2.3 will argue.

2.3. The Window of Introspective Memory

Recall from Chapter 3.3 that there is only one serious suggestion regarding the evolution of inner sense. This is that it emerged for purposes of cognitive monitoring and control. By monitoring our own learning and reasoning processes we can trouble-shoot in cases of mistake or difficulty, and we can exercise some degree of executive control over the course of our own mental lives. Plainly this requires that some sort of short-term record of the relevant set of mental events should be kept. One can't, for example, locate what has gone wrong in a piece of reasoning unless one can recall the steps that one has taken. We can therefore expect that inner sense should be able to identify events for at least a few seconds after they have occurred. This means that tests of introspective capacities don't have to confine themselves to the present. On the contrary, we should expect that representations of monitored mental events should still be accessible for a few seconds thereafter.

We should not, however, expect that mental events should be introspectable beyond the bounds of such a few-second window. Nor is there any reason to think that long-term memories of mental events should routinely be kept (as opposed to, or in addition to, memories of the worldly events that our thoughts and experiences mostly concern). For recall the proposed monitoring function of inner sense: if everything goes smoothly, and if our cognitive processes are successful, then there will have been no need to intervene, and there is likewise no need for a longer-term record to be kept. What would be the benefit of cluttering up our minds with memories of what we have previously felt and thought, unless those feelings and thoughts are tied to important and memorable public events? We can therefore predict that the results of introspective monitoring should fade quite rapidly, in something like the way that memories of our dreams fade rapidly on waking (unless attended to and rehearsed).

If the suggestion made here is correct, then it predicts that people should have very little awareness of the long-term patterns in their conscious mental lives. Since records of previous thoughts and thought processes aren't routinely kept (unless a decision is made to rehearse and remember those events), there will be no database that people can consult when constructing generalizations about their own minds. This prediction is strikingly borne out. For one of the robust findings in the introspection-sampling

literature built up by Hurlburt and colleagues is that it is very common indeed for subjects to make discoveries about the long-term patterns in their thinking and imagining that they had previously never suspected (Hurlburt, 1990, 1993; Hurlburt and Schwitzgebel, 2007). The methodology of these studies will be described in greater detail in Section 4. But the salient point for our purposes is that subjects are induced to jot down some notes immediately following introspected experiences at randomly generated intervals, thereby building up a record of conscious mental events that would otherwise (I suggest) have been forgotten.

The upshot of these considerations is that most of the studies that have been conducted by developmental psychologists on the question of self-knowledge versus other-knowledge don't really bear on the debate that concerns us here. For in most of these studies the children were only asked about their earlier mental states some significant time after the fact. They might, for example, have first been asked control questions to insure that they recall salient facts about the experimental conditions. But if so, then on the account suggested here no representation of the earlier mental state is likely to remain. Hence subjects will have no option but to engage in mindreading: trying to figure out what they are likely to have been thinking, in much the same way that they would try to figure out what another person in those circumstances would have been thinking. Small wonder, then, that there should be such high correlations between the results of self-knowledge tasks and other-knowledge tasks. This is because most of the former are, actually, third-person tasks with the self as subject. (As we will see in Chapter 10, however, not all first-person tasks suffer from the same criticism.)

2.4. Conclusion

I conclude that Nichols and Stich (2003) have failed to provide any evidence confirming one of the main predictions of their model (which is equally a prediction of Goldman, 2006). This is that a capacity to attribute mental states to oneself should emerge in development significantly in advance of a capacity to ascribe such states to others. Indeed the evidence, if taken at face value, would seem to push strongly in the other direction, and in support of parallelism in development. If true, this would support either the ISA account of self-knowledge or the model suggested by Frith and Happé (1999), both of which postulate just a single metarepresentational capacity, and so both of which predict such parallelism.

I have suggested in Section 2.3, however, that many of the tests of first-person awareness in the developmental literature are really third-person mindreading tasks in disguise. If so, then much of this literature has no bearing on our debate. As we will see in Chapter 8, however, there is a rapidly expanding body of work with infants using non-verbal paradigms. This suggests that basic mindreading competence is present *much* earlier in development than has traditionally been found using verbal methods. If this work is accepted, then it will provide significant support for parallelism after all. What we can conclude at this stage, however, is that there is no evidence *against* the

ISA theory, or in support of inner sense theories, from the main body of work in developmental psychology.

3. Emotional Mirroring

Recall from Section 1.3 that one of the main arguments presented by Goldman (2006) in support of an introspection-based simulation account of mindreading concerns emotional mirroring, together with the apparent causal role that one's own emotions play in recognizing the same emotions in other people. The present section will evaluate this argument.

3.1. Mirroring in General

There is now extensive evidence of what Goldman calls "low-level simulation" of emotion. Indeed, it has long been known that some emotions are contagious. When one baby in a nursery cries, perception of its distress is apt to cause the other babies to cry also. And likewise when one person in a group laughs, others are apt to be caused to laugh as well. But in fact the phenomenon seems to be a universal one, at least among basic emotions like fear, happiness, sadness, anger, and disgust (Wild et al., 2001). Seeing someone else afraid, for example, causes one to feel fear; and seeing someone experiencing disgust causes one to feel disgusted. Moreover the effect is both swift, on the one hand, and unconscious in the manner of its causation, on the other.

The causal mechanism underlying the phenomenon of emotional mirroring remains unclear. One suggestion is that it operates via behavioral mirroring, together with a "backwards" causal route from emotional behavior to emotional experience. We know that both of the steps needed for this account to work are real. Viewing someone else's facial expression causes minute muscle movements in one's own face corresponding to the emotion observed, which can be detected by electrical sensors (Jaencke, 1994; Hess and Blairy, 2001). Moreover, these movements occur very rapidly (within 300 milliseconds of stimulus onset; Dimberg and Thunberg, 1998), and they even occur in response to subliminally presented stimuli that are never consciously perceived (Dimberg et al., 2000). This gives us the first step in the postulated causal chain. But it is also known that prompting people to assume the facial expression characteristic of a particular emotion will give rise to the appropriate feelings and bodily responses (Adelman and Zajonc, 1989; Levenson et al., 1990). So the proposed explanation is perfectly possible.

An alternative suggestion is that perception of emotional situations or emotional behavior causes a mirroring emotion in oneself directly, which then in turn causes one to assume the appropriate facial expression. This is Goldman's preferred account. He adopts it, in part, because of the case of emotional mirroring for pain. A number of studies have shown that when people observe another person in a painful situation, the same affective areas of the brain that are involved when one feels pain oneself become active, especially the anterior insula and anterior cingulate cortex (Jackson et al., 2004;

Singer et al., 2004). But in these studies subjects did not view the face of the other person at all, they just saw someone's limb (a hand or a foot) in a painful situation. So their mirroring emotional state cannot result from facial mimicry.

It might be tempting to think that emotional mirroring and motor mirroring (of the sort discussed in Chapter 6.3) should be assimilated to one another, leading us to expect that each should be realized in similar mechanisms serving the same general functions. On this account, if we were correct to argue in Chapter 6.3 that motor mirroring depends upon prior mindreading, then we would need to say the same here: emotional mirroring, too, would depend upon prior mindreading of the emotions of others, contrary to Goldman's claims. And conversely, if Goldman is right that emotional mirroring plays a foundational role in enabling us to recognize the emotions of others, then this should lead us to re-evaluate our earlier assessment of the mirror-neuron literature. In fact, however, there seems no good reason why the different forms of mirroring should be regarded as similar, either in mechanism or function. Each might have taken very different evolutionary trajectories, and emotional mirroring might play an important role in mindreading even if motor mirroring doesn't.

3.2. Goldman's Argument

Emotional contagion is only the first step in Goldman's (2006) argument. The second is that one's own emotional experience plays a causal role in recognizing the emotions of others. Much of the evidence derives from lesion studies. Two patients with bilateral amygdala damage have been studied in some depth (Adolphs et al., 1994; Sprengelmeyer et al., 1999). In each case these patients are profoundly impaired in their capacity to feel fear (and only fear—other forms of emotional experience remain intact), but they are likewise deficient at recognizing fear in other people. Similarly, two patients with damage to the anterior insula have also been studied, demonstrating a parallel pattern of deficits for the case of disgust (Calder et al., 2000; Adolphs et al., 2003). In each case these subjects are impaired in their capacity to feel disgust (and only disgust). And they are likewise deficient in recognizing only disgust, whether exhibited in people's facial expressions, or via non-verbal sounds (such as retching), or in verbal prosody.

An initial worry about this argument is that recognition of another's emotion must have *already* taken place, at some level, in order for emotional mirroring to occur. If the process isn't to be a magical one, then the appropriate perceptual cues for the particular emotion displayed in the other person will need to have been identified and integrated somehow, in order to set in motion the creation of a similar emotion in the perceiver. But Goldman might reply, and with some justice, that the "recognition" in question can be quite low-level. In particular, the information generated might be localized to a particular encapsulated system, and needn't involve any representation of the other's emotional state as such. So the experience of the corresponding emotion in oneself could still be necessary for one to reach a conceptual judgment about the emotional state of the other person.

Goldman's suggestion, then, is that recognition of one's own emotions via introspection is basic, and that one recognizes emotional states in other people by being caused, first, to mirror those emotions in oneself, and by then introspectively identifying the emotions mirrored. Hence identifying the emotions of other people is grounded in first-person awareness of one's own emotions. We will shortly examine whether Goldman's account is correct. But first I want to emphasize that it isn't enough for Goldman to show that emotional mirroring *helps with*, or plays *some* role in, third-person emotion recognition. For as we will see in Section 3.3, an ISA theorist can accept this. Rather, since the view to be supported is that mindreading is grounded in first-person awareness, it needs to be shown that recognition of one's own mirroring emotions plays a *foundational* role in recognizing the emotions of others.[8]

3.3. Is Experience of One's Own Emotion Basic?

In fact the data do nothing to support the stronger of the two views just identified. For they are consistent with an account of mindreading that postulates an innately channeled information-rich mechanism of some sort, rather than introspection-based simulation. This system would have access to globally broadcast perceptual and bodily information (as outlined in Chapter 3), and included in these broadcasts would be the affective and behavioral components of one's own emotional states, as we saw in Chapter 5. This affective and proprioceptive information might be used by the mindreading system in an ancillary way, without having the sort of foundational role that Goldman attributes to it (and consistently with the truth of the ISA theory). For there is good reason to think that emotional contagion is an ancient phenomenon, which almost certainly anteceded the emergence of mindreading. But when the latter evolved it might have co-opted this additional source of information. Let me take these points in turn.

When one animal shows signs of fear, those around it will likewise become anxious. This makes good evolutionary sense. For if one animal has identified a source of danger to itself, then nearby conspecifics are likely to be in equal danger. A similar evolutionary rationale applies to mirroring of disgust. For if one person is disgusted at something, then it will both aid evaluative learning and help you to avoid a possible source of contaminants if you immediately feel disgust at that thing yourself. Note that in neither case need mindreading of any sort be involved. For the cues that trigger a mirroring emotion are quite low-level, and need involve no conceptual recognition of the emotion as such.

[8] In fact there is some reason to think that a pluralist position may be preferable. For Oberman et al. (2007) found that blocking facial mimicry (by requiring subjects to bite down on a pen) interfered with recognition of happy faces, but not with subjects' recognition of disgust, fear, or sadness. This suggests either that the presence of mirroring emotions may play a role in recognizing others' emotions in some cases but not others (if blocking facial mimicry blocks the corresponding emotion), or else that facial mimicry makes an independent contribution to one's recognition of happiness, but not to one's recognition of other emotions.

Suppose, then, that emotional contagion pre-existed the evolution of any form of mindreading capacity. And suppose, for argument's sake, that hominins at this point lacked any abilities for introspective self-knowledge. Indeed, suppose they lacked the concept of emotion altogether. But then social pressures of various sorts led to the evolution of a basic mindreading faculty (or to an innate disposition to construct one via learning), which conferred on hominins for the first time the capacity to think about emotions as such. Recognition of other people's emotions would nevertheless have been a noisy and error-prone process. It would therefore have been adaptive to be able to utilize any reliable source of information that came to hand. This is where awareness of the affective and behavioral components of one's own emotional states would help. Given that emotional contagion is a reliably occurring process, one is likely to be more reliable in identifying other people's emotions if one can integrate external perceptual cues of emotion, relating to the other person's facial expression and bodily posture, with internal signals that more or less reliably indicate one's own corresponding emotion.[9]

If this account is on the right lines, then it would explain the pairing of deficits that Goldman (2006) appeals to in his argument. For if someone's capacity to feel emotion is impaired, then their capacity to recognize that emotion in others will likewise be impaired to some degree, since this provides one of the sources of information that they rely on. Note, moreover, that the collapse in emotion recognition in others is never total, even for those who seem to be wholly incapable of experiencing the relevant emotion. This suggests that recognition normally relies on other cues as well, and that emotional recognition isn't purely introspection-based. On this account, therefore, introspection would no longer be basic. In addition, the account wouldn't presuppose introspection for emotional *propositional* attitudes as such at all. Rather, awareness of the affective component of the emotion would be sufficient. (This is just as well, in light of our discussion of the limits of emotional awareness in Chapter 5.4.)

An additional problem for the claim that awareness of one's own emotional states is basic concerns the case of pain. Danziger et al. (2006) compared twelve patients with congenital insensitivity to pain with normal controls in a variety of pain recognition tasks. The patients didn't differ from the controls in their estimates of the painfulness to other people of various verbally-described events. Nor did they differ from controls in their estimates of someone's degree of pain judged on the basis of facial expression. However, they did display considerably more variance than controls in their estimates of the painfulness of various videos of painful events (such as someone falling from a skateboard or missing a jump from a diving board), and they did tend to underestimate the amount of pain involved. But in these videos people's facial expressions and other

[9] Given that people are generally rather poor at discriminating interoceptive information (as we noted in Chapter 5.2), most of the weight is likely to fall on proprioceptive awareness of one's own mirroring facial and postural behavior. And in that case, as we noted in Chapter 5.1, this aspect of one's awareness counts as interpretive rather than transparent.

behavioral reactions to pain weren't visible, so all subjects would have been forced to estimate the painfulness of the event in some other way. As Danziger and colleagues point out, normal subjects seem to use a simulation of some sort in reaching such judgments. They imagine themselves in a similar situation and monitor the pain reaction that results. The patients with congenital insensitivity to pain, in contrast, would have had to rely on extrapolation from their memories of the amount of pain that people had expressed in other similar situations.

It appears from this study that recognition of pain behavior in other people does not depend upon the experience of pain in oneself (whereas estimates of degrees of pain from situational cues alone, in the absence of pain behavior, might do so to some extent). It is open to Goldman to reply, however, that people with congenital insensitivity to pain might have acquired some alternative route for recognizing the pain of others. He can still claim that in normal subjects recognition of others' pain depends upon introspective awareness of one's own pain. But at least the data demonstrate that simulation of other people's pain experiences isn't *necessary* for the recognition of pain.

3.4. A Common Cause?

In addition to the criticisms of Goldman's (2006) argument made above, two subsequent studies suggest that (at least in the case of fear) one's own emotion might fail to play any causal role in identifying the emotions of others. Rather, both the capacity to feel the emotion and the capacity to recognize the emotion in other people's faces may be results of a common cause. And then a single impairment in the underlying structure would be sufficient to bring about impairments in both first-person experience and third-person recognition.

Atkinson et al. (2007) presented a novel test of fear recognition to two patients with bilateral amygdala damage, who are both severely impaired for the experience of fear and for recognition of fear in people's faces. They presented these subjects with dynamic body-motion stimuli of people expressing fear or other emotions in their behavior (but with facial expressions obscured). They also presented them with static images of body postures typical of fear and other emotions. Much to the experimenters' surprise, both subjects were completely normal in their capacity to identify fear. So it appears that recognizing fear from bodily (as opposed to facial) cues utilizes a different set of resources, and remains undamaged in these subjects. And it follows that in these cases, at least, recognition of others' emotions does *not* depend on a capacity to experience those emotions in oneself.

Even more significant, Adolphs et al. (2005) investigated the way in which a subject with severe bilateral amygdala damage scanned people's faces with her eyes during tasks requiring recognition of emotion from static images. (In fact this is the same subject who had participated in many of the experiments described above, who is incapable of feeling fear herself.) Using eye-trackers, they noticed that she paid much less attention to the eye regions than do normal controls. But these regions are known to be critical

for recognizing the emotion of fear in particular. Indeed, when the subject was instructed to pay attention to the eyes while viewing the pictures, her fear-recognizing capacities became completely normal. But the effects were temporary. When the subject was tested again after an interval, and not given any reminder of the importance of eyes, her capacity to recognize fear in faces diminished dramatically once again.

It would appear from these data that the amygdala plays a necessary role in the creation of the emotion of fear in oneself, and that it also plays a role in directing visual attention to the eyes of other people when the task requires identifying their state of fear. But one's own emotional state doesn't seem to play any role in emotional identification as such. Rather, an intact amygdala is a common cause of both effects. Hence in the case of fear, at least, it would seem that Goldman's (2006) introspection-based simulation theory has been falsified.

3.5. Conclusion

I conclude that while the phenomenon of emotional mirroring is perfectly real, it provides no support for an account of mindreading as grounded in introspection. At most the evidence shows that sensorily-accessible affective and proprioceptive data are among those that the mindreading faculty uses when determining the emotional state of another person. In addition, the most recent evidence suggests that the paired deficits that result from amygdala damage may be results of a common cause. Identification of one's own emotion of fear would appear to be playing no role in enabling one to recognize the emotional state of another. One of the main remaining supports for Goldman's form of inner sense theory has therefore been undermined.

4. Unsymbolized Thinking

Recall from Chapter 1.2 that the ISA account predicts that we should be incapable of attributing attitudes to ourselves in the absence of relevant sensory data. All forms of inner sense theory, in contrast (as well as the attitudinal working memory and mental-action accounts discussed in Chapters 6.2 and 6.4), make the opposite prediction. Since they maintain that we can detect our own propositional attitudes through the operations of a special faculty of inner sense or non-sensory working memory, subjects should generally have no need of sensory evidence of any kind when making self-attributions. The presence of behavioral, contextual, or sensory cues should be entirely accidental. The present section will consider some evidence that appears to support inner sense and attitudinal working memory accounts over the ISA theory on just this point.[10]

[10] Chapter 9 will argue, in contrast, that many kinds of metacognitive judgment—such as judgments of learning—are actually dependent upon sensory cues. Hence in these cases, at least, the sensory cues *aren't* accidental.

4.1. Easily Explained Data

The data that seem to support inner sense theories over the ISA account derive from "descriptive experience sampling" studies conducted with normal subjects, using the methodology devised by Hurlburt (1990, 1993). Subjects wear a paging device throughout the day, through which they hear a beep at randomly generated intervals. Subjects are instructed to "freeze" the contents of their consciousness at the very moment of the beep, and to make brief notes about it to be discussed and elaborated at a later meeting with the experimenter. Most normal subjects report, in varying proportions, the occurrence of inner speech, visual imagery, and emotional feelings. But a significant number of subjects also report the presence of "purely propositional," or "unsymbolized," thoughts at the moment of the beep (Hurlburt and Akhter, 2008). In these cases subjects report thinking something highly determinate—such as that they were wondering whether or not to buy a particular box of breakfast cereal—in the absence of any visual imagery, inner speech, or other symbol-like sensory accompaniments.

So far there isn't any difficulty, here, for the ISA account. For the latter doesn't claim that all attributions of thoughts to oneself should be grounded in *imagistic* evidence, of course. Rather, the ISA account claims that self-attributions of thought should depend on the presence of imagistic cues *and/or sensorily-available behavioral or circumstantial* evidence. And what is striking about a good many instances of self-attributed unsymbolized thought is that they occur in circumstances where a third-party observer might have made precisely the same attribution. If you saw someone standing motionless, looking reflectively at a box of breakfast cereal on a supermarket shelf, for example, then you might well predict that she was wondering whether or not to buy it. So the subject who reported entertaining just such a thought when the beep sounded while she was looking at a box of cereal (Hurlburt, 1993) might have arrived at that attribution through swift self-interpretation. Our suggestion can therefore be that when prompted by the beep, subjects turn their mindreading systems on their own behavior and circumstances (together with any sensory or imagistic cues that are present), often enough interpreting themselves as entertaining a specific thought. Provided that the process happens swiftly, then the resulting thought will be self-attributed with all of the phenomenological immediacy and seeming-introspective obviousness as normal.

Consider another example. Siewert (1998) describes a case in which he was standing in front of his apartment door having just inserted his hand into his pocket where he normally keeps his key, finding it empty. Although he neither verbalized nor visualized anything at the time, at that moment he was (he says) wondering where the key could be. And his knowledge of this act of wondering was (he says) immediate, resulting from introspection. But notice, again, that the thought Siewert attributed to himself is exactly what a third-party observer with the same background knowledge might ascribe. For anyone seeing him standing in front of his door fumbling in his pocket, knowing that this is the pocket in which he normally keeps his key while also knowing

that the pocket is empty, might predict that he is wondering where the key might be. And this is especially likely if the observer were also to know that Siewert had just begun to feel anxious, as he reports that he had.

4.2. Harder Cases

A great many of the examples of unsymbolized thinking in the literature can be handled in this sort of way, as involving swift self-interpretation from background knowledge together with observations of behavior and current circumstances. But not quite all of them can. For instance, at the time of the beep one subject—Abigail—reported that she was wondering whether her friend Julio would be driving his car or his truck when he came to collect her later that day (Hurlburt and Akhter, 2008). This thought seemed to occur in the absence of any inner speech or visual imagery. Yet there was nothing in the subject's immediate circumstances or behavior from which it could be derived, either.

What cannot be ruled out, however, is that the thought in question was self-attributed because it made the best sense of sensory activity that had been taking place just *prior* to the moment "frozen" by the beep. So what seems like awareness of an unsymbolized thought might really be a belief formed by the mindreading system from interpretation of imagistic activity that had been occurring just previously. For example, Abigail might have recently entertained two memory images deriving from previous experience, in one of which Julio arrives in his car and in the other of which he arrives in his pickup truck, perhaps combined with a feeling of uncertainty. Alternatively, shortly before the beep she might have rehearsed in inner speech the sentence, "Will Julio be driving his car or his truck?" Either set of events would have led Abigail's mindreading faculty to formulate the higher-order belief that she is wondering whether Julio will be driving his car or his truck. This belief might remain at the time of the beep, and be passed along to executive and language systems for report, although memory of the previous sensory imagery that gave rise to it has been lost.

Note that this proposal is by no means arbitrary. For subjects are instructed to focus on, and report, *only* what is occurring at the moment of the beep. And we know that memory for sensory experience fades rapidly when not attended to. (Think, again, of dreams, which disappear rapidly from memory unless fixed in attention.) Moreover, the sound of the beep itself will attract attention, of course, and in some cases this may have the effect of hastening still further the loss of the subject's memory for earlier imagery (especially if the latter is weak and fragmentary).[11]

[11] Indeed, attention to the beep may serve to "backward-mask" one's previous experience, in the manner noted in Chapter 5.3. For in general when a briefly presented stimulus is followed swiftly by another that attracts one's attention, then all memory of the former tends to be lost (Breitmeyer and Ogmen, 2000). Note, however, that such unconsciously experienced stimuli can still prime related thoughts and behaviors. This raises the possibility that self-attributed unsymbolized thoughts might result from self-priming by previous imagery that has been backward-masked by the sound of the attended-to beep.

How might these alternative explanations be tested? Hurlburt's methodology makes no provision for collecting data on experiences occurring in the seconds prior to the beep. So one might suggest extending the subjects' task to report, not just experience concurrent with the beep, but also conscious events from the moments before. However, this extended task is likely to overwhelm people's working memory capacities. Another possible, but indirect, test would be to look for correlations between the extent to which different subjects report unsymbolized thoughts (with quantities of inner speech and visual imagery controlled for) and the speed of their mindreading abilities in third-person tasks. Since subjects will only have the illusion of introspecting an unsymbolized thought if they can reach an interpretation smoothly and swiftly from contextual or briefly-presented sensory data, then one might predict that there should be a positive correlation.

Hurlburt and Akhter (2008) concede the possibility that attributions of unsymbolized thought to oneself might result from swift and unconscious self-interpretation. But they present the following consideration against such an idea. Many subjects are initially quite reluctant and hesitant when describing instances of unsymbolized thought in follow-up interviews. (According to Hurlburt and Akhter, this is because they hold a folk-belief that all conscious thinking is accompanied by images of one sort or another.) This suggests that subjects did *not* arrive at their beliefs about unsymbolized thinking through self-interpretation, Hurlburt and Akhter say. But explicitly held folk theories are one thing, assumptions built into the operations of the mindreading faculty are quite another. And there is no reason to think that the latter will share all of the explicit theoretical beliefs adopted by the folk. Hence the mindreading system might have no hesitation in attributing a thought to the self in the absence of any presently-accessible sensory cues, even though the person in whom that system resides does so hesitate. I conclude that the introspection-sampling data do not, as yet, provide evidence that the ISA theory cannot accommodate.

Moreover, Hurlburt himself (2009) suggests that unsymbolized thinking is consistent with the ISA model. For to say that someone is engaged in unsymbolized thinking is to say that there is no sensory awareness of any imagistic symbols, at the time of a self-attributed thought. But "sensory awareness," for Hurlburt, is a technical term, referring to sensory information that is at the focus of attention. He therefore suggests that attributions of unsymbolized thought may result from the apprehension of some "sensory bits," so long as those sensory fragments are not organized into a coherent, central, thematized sensory awareness of the sort that would be revealed in a standard introspection-sampling interview. It is quite possible, therefore, that people engaged in unsymbolized thinking do have fragmentary imagistic awareness at the moment of the beep that could aid in a process of self-interpretation, leading to the attribution to themselves of a particular thought. Since subjects are unaware of the self-interpretation process, but find themselves inclined to attribute a specific thought to themselves, they will have the sense that they are consciously thinking that thought in an unsymbolized way.

Indeed, Hurlburt (2009) goes further, claiming that the introspection-sampling data actually *supports* a self-interpretive model. But here I think he oversteps the mark. His reasoning is that introspection-sampling subjects never report any awareness of an attitude at the moment of the beep, except in the early stages of training (in which case they are inclined to back off such claims in discussion with the interviewer). But in making this claim Hurlburt must have in mind medium-term or standing-state attitudes like *intending to go out to dinner this evening*, or *believing that the economy will soon turn around*. For introspection-sampled subjects *do* report *momentary* attitudes in cases of unsymbolized thinking, and they *don't* back off these claims. One will report *wondering* something (as in the case of Abigail, described above), whereas another will report *wishing* for something or *worrying* about something, and so on. But the ISA account of self-knowledge doesn't claim only that *standing* attitudes are attributed to oneself through interpretation, of course. On the contrary, the thesis extends also to activated ones like judging, wondering, wishing, and worrying.

4.3. A Double-Edged Sword

Although even the "harder" data on unsymbolized thinking discussed in Section 4.2 can be explained by an ISA theorist, that explanation requires us to adopt an auxiliary assumption. This is that in such instances there are always sufficient sensory cues occurring near the time of the beep to enable self-interpretation, although those cues are forgotten or remain unnoticed. The upshot is that the ISA theory is somewhat weakened. For even if the auxiliary assumption in question is a plausible one, our only direct reason for believing it, at this point, is that it enables us to preserve the ISA theory in the face of apparent counter-evidence.

I want to emphasize, however, that the data on unsymbolized thinking are a double-edged sword, and that they actually pose equal or even greater problems for inner sense theories (as well as for attitudinal working memory and action-awareness accounts). The source of the problem is the patterning of the data across subjects. Only some people ever report unsymbolized thoughts, and they only do so some of the time (Heavey and Hurlburt, 2008). Many people's reports suggest that their thoughts are *always* expressed in sensory images of one sort or another. (And even those people who do report unsymbolized thoughts also tend to report imagistic forms of thinking as well.) Why should this be so, if people possess a faculty of inner sense (or an attitudinal working memory system) that enables them to detect their attitudes directly? For in that case one would expect *everyone* to report unsymbolized thoughts with high frequency.

Someone might try to deny that the patterning in the data is reliable, building on one of the points noted in Section 4.2. This is that some people are initially quite hesitant to report instances of unsymbolized thinking, presumably because the very idea of such thoughts conflicts with their preconceived theories. Perhaps those who *never* report such thoughts are simply those in whom this reluctance is strongest. This suggestion is implausible, however. For Hurlburt and colleagues go to great lengths to

emphasize to their subjects that they should set aside any preconceptions about their conscious experience and provide faithful and accurate reports, whatever the content of the latter. And this emphasis is repeated in interviews with subjects over many days (Hurlburt and Schwitzgebel, 2007). Moreover, Schwitzgebel (2007) was able to test the effects of people's preconceptions on their introspective reports in his own introspection-sampling study by gathering data on the former beforehand, and was able to find no biasing effects. Admittedly this study wasn't concerned with the question of unsymbolized thinking, but rather with the question of the richness of experience. But it does at least show that theoretical preconceptions don't always bias subjects' reports.

It is reasonable to assume, then, that the data are reliable, and that it really is the case that many people never experience unsymbolized thought. The challenge for inner sense theorists and attitudinal working memory theorists is to explain why this should be so. It isn't difficult to explain why some people should regularly report inner speech while others hardly do so at all, or why some people should never report visual imagery or emotional feelings (Hurlburt, 1993; Heavey and Hurlburt, 2008). For such phenomena depend upon the global broadcast of sensory representations, and will consequently be attention-dependent. Indeed, we have already had occasion to note in Chapter 4.4 that conscious, "System 2," thinking and reasoning is highly idiosyncratic in its patterning and contents, depending upon people's habits of attention and mental rehearsal. But it isn't so easy to extend such an account to explain why some people should never experience unsymbolized thinking, from the perspective of inner sense theory. This is because *everyone* has attitude-events like judgments and decisions, of course, and it would be quite remarkable if there were individual differences in the extent to which this is true. So, all of the weight must be thrown onto the attentional component of the explanation sketched above: it would have to be said that some people don't report their unsymbolized thoughts, not because they don't have them, but because they don't pay any attention to them.

Notice, however, that this would be tantamount to saying that people have a faculty of inner sense (or an attitudinal working memory system) that goes unused. If true, this would really be quite surprising. Since any such faculty will involve mechanisms that are complex and costly to maintain, we have argued that they must have been subject to significant selection pressure. But it seems unlikely that the need for monitoring and control functions, and/or for flexible forms of thinking and reasoning, should no longer be operative in the modern world (even in a subset of the population). On the contrary: literate learning-dependent societies such as ours should place a premium on such functions. It is therefore exactly as if we had found significant numbers of people who never experience any episodic memories, despite retaining the underlying capacity for such memories. This, too, would be puzzling in just the same way. It would require us to believe that a complex and important cognitive mechanism is lying dormant and unused.

It seems, therefore, that the patterning in the introspection-sampling data raise significant problems for inner sense theories (and also for attitudinal working memory and action-awareness accounts). Moreover, these problems seem quite severe. At the very least we can conclude that such theories will need to adopt some or other auxiliary assumption in order to accommodate the data. And in contrast with the ISA theory's handling of the data on unsymbolized thinking, it is far from clear what auxiliary assumption could plausibly serve.

4.4. Introspection During Speech

Before concluding this section, let me briefly discuss what Hurlburt calls "partially unworded speech" (Hurlburt and Schwitzgebel, 2007).[12] Although comparatively rare, its mere existence might be thought to count against the ISA theory of self-knowledge and in favor of some form of inner sense account. In cases of partially unworded speech, introspection-sampled subjects report tokens of fragmentary inner speech at the time of the beep, but they nevertheless feel that they know the complete content of the underlying thought or speech intention (that is, they have a conscious sense of what words belong in the missing parts of the token). For example, when a token of inner-hearing belonging to one subject—Melanie—was interrupted by a beep, she reported having a clear sense of how the sentence would otherwise have ended. She was in her car, and had just realized that she had forgotten to take off the parking brake as she tried to move off. She heard, "Why can't I . . . " just at the moment of the beep. She reported knowing that the sentence was going to end with " . . . remember about the parking brake?" Do cases like this suggest that there is introspective access to one's thoughts beyond the imagery that one experiences, as Schwitzgebel claims (Hurlburt and Schwitzgebel, 2007)?

They surely do not. For according to the ISA account, self-interpretation doesn't just draw on evidence from internal imagery, but also on facts about the agent's behavior and circumstances (in the latter regard operating much like third-person mindreading). And then just as a third party observing Melanie might predict, given her situation, that her episode of inner speech would involve the parking brake, so Melanie is able to make that prediction about her own experience. Thus as long as the content of an inner-speech episode could be inferred from a subject's situation, partially unworded speech poses no threat to the ISA model. And indeed, consistent with that model, Hurlburt reports that the most frequent experience of inner speech involves simply the speech itself, with no conscious sense of what is about to be said (Hurlburt and Schwitzgebel, 2007). Moreover, he tells us that when people are beeped in the midst of speaking they generally do *not* have conscious awareness of what they are intending to say.

[12] This book is constructed as a dialog between the two authors, with some chapters written by Hurlburt, some chapters written by Schwitzgebel, and some chapters reporting their discussions.

These latter points could bear some emphasis. For they suggest that people may have no access to their own intentions in speaking, just as the ISA theory predicts. Consider a case in which someone is beeped in the midst of an item of inner or outer speech, then, where the content of that speech isn't predictable from the context (as in most cases it is not). In these circumstances the ISA theory predicts that subjects should have no immediate sense of how the speech-episode would have continued, since they would lack any evidence on which to base such a prediction. Since the data appear to bear this out, the ISA theory is to that extent confirmed. From the perspective of inner sense theories, in contrast, there is no reason to think that one's speech intentions wouldn't be introspectable, and so the data constitute another anomaly.

4.5. Conclusion

I conclude that the introspection-sampling data provide little or no support for inner sense theories of self-knowledge (nor for attitudinal working memory models). In particular, subjects who report unsymbolized or partly-worded thoughts at the moment of the beep may actually be relying on self-interpretation, grounded in prior imagistic activity, current imagery of a non-symbolic sort, and/or knowledge of current behavior and circumstances. Provided that the mindreading system does its work swiftly and unconsciously, subjects will simply find themselves with the powerful intuition that they were entertaining (or were in the process of entertaining) a specific thought at the moment of the beep, but with no awareness of how this intuition is arrived at.

Admittedly, the ISA theory is forced to appeal to an ancillary hypothesis in order to accommodate all of the data, and this is a hypothesis for which we currently lack direct evidence. It is that instances of reported unsymbolized thought will always involve sensory cues sufficient to enable the mindreading faculty to do its interpretive work (but ones that aren't recalled by the subjects). Inner sense theories, in contrast, can take the data at face value. Taken in isolation this provides *some* reason to prefer the latter. But the ancillary hypothesis in question is by no means arbitrary. In part this is because we already know that in many instances of unsymbolized thought there are sufficient contextual and/or behavioral cues demonstrably present. But the hypothesis also coheres well with what we know about the effects of attention on memory for experience. In addition, inner sense theories (as well as attitudinal working memory and action-awareness theories) face their own problems in explaining why only *some* people should ever report unsymbolized thoughts. These theorists, too, will need to appeal to some or other auxiliary hypothesis to explain the patterning of the data.

I conclude, therefore, that the introspection-sampling data require ISA theorists to pay an additional cost, which is quite small; and competing theories, too, by no means get to endorse the data for free. On the contrary, they are also required to pay an additional cost. This looks, at worst, like a stand-off in the competition between the two sorts of approach, and at best as a further reason to prefer the ISA theory to any form of inner sense theory or attitudinal working memory account.

5. Conclusion

Our examination of the claim that we possess one or more special, non-interpretive, channels of information to our own attitudes is by no means complete. For crucial predictions remain to be evaluated. In particular, we need to consider evidence relating to the claim that these channels of access were fashioned by evolution for purposes of self-monitoring and cognitive control. This will happen in Chapter 9. And we also need to consider whether there exist patterns of dissociation between self-knowledge and other-knowledge of the predicted sorts. This will be discussed in Chapter 10. Moreover, we need to consider direct evidence *against* inner sense views (as well as against most other forms of transparent-access theory), relating to failures and inadequacies in our knowledge of our own propositional attitudes. This will take place in Chapter 11.

At this point, however, we can fairly conclude that the case in support of inner sense theories is distinctly underwhelming. For the various positive arguments that we have examined are either flawed, or fail to count significantly against the competing ISA account. The latter, in contrast, is currently supported by considerations of simplicity and explanatory scope, as well as by its nice coherence with surrounding theories in cognitive science. It also either predicts, or is consistent with, all of the evidence we have reviewed up to now. Although this isn't a final verdict in favor of the ISA theory, our discussion to date suggests that it is much the more promising account.