5

PETER CARRUTHERS

Distinctively Human Thinking

Modular Precursors and Components

This chapter takes up, and sketches an answer to, the main challenge facing massively modular theories of the architecture of the human mind. This is to account for the distinctively flexible, non-domain-specific character of much human thinking. I shall show how the appearance of a modular language faculty within an evolving modular architecture might have led to these distinctive features of human thinking with only minor further additions and non-domain-specific adaptations.

1 On Modularity

To what extent is it possible to see the human mind as built up out of modular components? Before this question can be addressed, something first needs to be said about what a module is, in this context; and also about why the issue matters.

1.1 Fodorian Modularity

In the beginning of our story was Fodor (1983). Against the prevailing empiricist model of the mind as a general-purpose computer, Fodor argued that the mind contains a variety of specialized input and output systems, or *modules*, as well as a general-purpose central arena in which beliefs get fixed, decisions taken, and so on. Input systems might include a variety of visual systems (including face recognition),

Early versions of this chapter were delivered at a preparatory workshop during the first year of the AHRB Innateness and the Structure of the Mind project, held in Newark 12–14 October 2001, and at the "Innateness and the Structure of the Mind" conference that concluded the year, held in Sheffield 3–6 July 2002. I am grateful to those who participated in those discussions for valuable feedback, with special thanks to Mark Baker, Robert Boyd, Richard Samuels, Gabriel Segal, Dan Sperber, Stephen Stich, and John Tooby for comments and suggestions.

auditory systems, taste, touch, and so on; but they also include a language-faculty (which contains, simultaneously, an output/production system, or else divides into input and output subsystems).

In the course of his argument, Fodor provided us with an analysis (really a stipulative definition) of the notion of a module. Modules are said to be processing systems that (1) have proprietary transducers, (2) have shallow outputs, (3) are fast in relation to other systems, (4) are mandatory in their operation, (5) are encapsulated from the remainder of cognition, including the subject's background beliefs, (6) have internal processes that are inaccessible to the rest of cognition, (7) are innate or innately channeled to some significant degree, (8) are liable to specific patterns of breakdown, both in development and through adult pathology, and (9) develop according to a paced and distinctively-arranged sequence of growth. At the heart of Fodor's account is the notion of *encapsulation*, which has the potential to explain at least some of the other strands. Thus, it may be because modules are encapsulated from the subject's beliefs and other processes going on elsewhere in the mind that their operations can be fast and mandatory, for example. And it is because modules are encapsulated that we stand some chance of understanding their operations in computational terms. For, by being dedicated to a particular task and drawing on only a restricted range of information, their internal processes can be *computationally tractable*.

According to Fodor (1983, 2000), however, central/conceptual cognitive processes of belief-formation, reasoning, and decision-making are definitely *amodular* or holistic in character. Crucially, central processes are unencapsulated—beliefs in one domain can have an impact on belief-formation in other, apparently quite distinct, domains. And in consequence, central processes are *not* computationally tractable. On the contrary, they must somehow be so set up that any one of the subject's beliefs can be brought to bear in the solution to a problem. Since we have no idea how to build a computational system with these properties (Fodor has other reasons for thinking that connectionist approaches won't work), we have no idea how to begin modeling central cognition computationally. And this aspect of the mind is therefore likely to remain mysterious for the foreseeable future.

1.2 Central Modularity

In contrast to Fodor, many other writers have attempted to extend the notion of modularity to at least some central processes, arguing that there are modular central/conceptual systems as well as modular input and output systems (Atran, 2002; Baron-Cohen, 1995; Botterill & Carruthers, 1999; Carey, 1985; Carey & Spelke, 1994; Gallistel, 1990; Hauser & Carey, 1998; Hermer-Vazquez et al., 1999; Leslie, 1994; Smith & Tsimpli, 1995; Spelke, 1994). Those who adopt such a position are required to modify the notion of a module somewhat. Since central modules are supposed to be capable of taking conceptual inputs, such modules are unlikely to have proprietary transducers; and since they are charged with generating conceptualized outputs (e.g., beliefs or desires), their outputs cannot be shallow. Moreover, since central modules are supposed to operate on beliefs to generate other beliefs, for example, it seems unlikely that they can be fully encapsulated—at least *some* of the subject's existing beliefs can be accessed during processing by a central module. But the notion of a "module" is not thereby wholly denuded of content. For modules can still be (1) domain specific, taking only domain-specific inputs, or inputs containing concepts proprietary to the module in question, (2) fast in relation to other systems, (3) mandatory in their operation, (4) relatively encapsulated, drawing on a restricted domain-specific database; as well as (5) having internal processes or algorithms that are inaccessible to the rest of cognition, (6) being innate or innately channeled to some significant degree, (7) being liable to specific patterns of breakdown, and (8) displaying a distinctively ordered and paced pattern of growth.

I shall not here review the evidence—of a variety of different kinds—that is supposed to support the existence of central/conceptual modules that possess many of the foregoing properties (see Carruthers, 2003b, for a review). I propose simply to assume, first, that the notion of central-process modularity is a legitimate one; and second, that the case for central modularity is powerful and should be accepted in the absence of potent considerations to the contrary.

1.3 *Massive Modularity*

Others in the cognitive science community—especially those often referred to as "evolutionary psychologists"—have gone much further in claiming that the mind is wholly, or at least *massively*, modular in nature (Cosmides & Tooby, 1992, 1994; Gallistel, 2000; Pinker, 1997a; Sperber, 1994, 1996; Tooby & Cosmides, 1992). Again, a variety of different arguments are offered; these I shall briefly review, since they have a bearing on later discussion. But for the most part in what follows I shall simply assume that some form of massive modularity thesis is plausible, and is worth defending.

(Those who don't wish to grant the foregoing assumptions should still read on, however. For one of the main goals of this chapter is to consider whether there exists any powerful argument *against* massive modularity, premised upon the non-domain-specific character of central cognitive processes. If I succeed in showing that there is not, then that will at least demonstrate that any grounds for rejecting the assumption of massive modularity will have to come from elsewhere.)

One argument for massive modularity appeals to considerations deriving from evolutionary biology in general. The way that evolution of new systems or structures characteristically operates is by "bolting on" new special-purpose items to the existing repertoire. First, there will be a specific evolutionary pressure—some task or problem that recurs regularly enough and that, if a system can be developed that can solve it and solve it quickly, will confer fitness advantages on those possessing that system. Second, some system that is targeted specifically on that task or problem will emerge and become universal in the population. Often, admittedly, these domain-specific systems may emerge by utilizing, coopting, and linking together resources that were antecedently available; hence they may appear quite inelegant when seen in engineering terms. But they will still have been designed for a specific purpose, and are therefore likely to display all or many of the properties of central modules, outlined earlier.

A different—though closely related—consideration is negative, arguing that a general-purpose problem-solver *couldn't evolve*, and would always by out-competed by a suite of special-purpose conceptual modules. One point here is that a general-purpose problem-solver would be very slow and unwieldy in relation to any set of domain-specific competitors, facing, as it does, the problem of combinatorial explosion as it tries to search through the maze of information and options available to it. Another point relates more specifically to the mechanisms charged with generating desires. It is that many of the factors that promote long-term fitness are too subtle to be noticed or learned within the lifetime of an individual; in which case there couldn't be a general-purpose problem-solver with the general goal "promote fitness" or anything of the kind. On the contrary, a whole suite of fitness-promoting goals will have to be provided for, which will then require a corresponding set of desire-generating computational systems (Tooby, Cosmides, & Barrett, chapter 18 here).

The most important argument in support of massive modularity for my purposes, however, simply reverses the direction of Fodor's (1983, 2000) argument for pessimism concerning the prospects for computational psychology. It goes like this: the mind is computationally realized; amodular, or holistic, processes are computationally intractable; so the mind must consist wholly or largely of modular systems. Now, in a way Fodor doesn't deny either of the premises in this argument; nor does he deny that the conclusion follows. Rather, he believes that we have independent reasons to think that the conclusion is false; and he believes that we cannot even *begin* to see how amodular processes could be computationally realized. So he thinks that we had better give up attempting to do computational psychology (with respect to central cognition) for the foreseeable future. What is at issue in this debate, therefore, is not just the correct account of the structure of the mind but also whether certain scientific approaches to understanding the mind are worth pursuing.

Not all of Fodor's arguments for the holistic character of central processes are good ones. (In particular, it is a mistake to model individual cognition too closely on the practice of science, as Fodor does. See Carruthers, 2003a). But the point underlying them is importantly correct. And it is this that is apt to evince an incredulous stare from many people when faced with the more extreme modularist claims made by evolutionary psychologists. For we *know* that human beings are capable of linking together in thought items of information from widely disparate domains; indeed, this may be *distinctive* of human thinking (I shall argue that it is). We have no difficulty in thinking thoughts that link together information across modular barriers. How is this possible, if the arguments for massive modularity, and against domain-general cognitive processes, are sound?

1.4 A Look Ahead—The Role of Language

We are now in position to give rather more precise expression to the question with which this chapter began; and also to see its significance. Can we finesse the impasse between Fodor and the evolutionary psychologists by showing how non-domain-specific human thinking can be built up out of modular components?

If so, then we can retain the advantages of a massively modular conception of the mind—including the prospects for computational psychology—while at the same time doing justice to the distinctive flexibility and non-domain-specific character of some human thought processes.

This is the task that I propose to take up in this chapter. I shall approach the development of my model in stages, corresponding roughly to the order of its evolution. This is because it is important that the model should be consistent with what is known of the psychology of other animals, and also with what can be inferred about the cognition of our ancestors from the evidence of the fossil record.

I should explain at the outset, however, that according to my model, it is the language faculty that serves as the organ of intermodular communication, making it possible for us to combine contents across modular domains. One advantage of this view is that almost everyone now agrees (1) that the language faculty is a distinct input and output module of the mind, and (2) that the language faculty would need to have access to the outputs of any other central/conceptual belief or desire forming modules, in order that those contents should be expressible in speech. So in these respects language seems ideally placed to be the module that connects together other modules, if this idea can somehow be made good sense of.

Another major point in favor of the proposal is that there is now direct (albeit limited) empirical evidence in its support. Hermer-Vazquez and colleagues (1999) have proposed and tested the thesis that it is language that enables geometric and object-property information to be combined into a single thought, with dramatic results. This evidence is reviewed and extended in Shusterman and Spelke (chapter 6 here) and so does not need to be elaborated upon here.

2 Animal Minds

What cognitive resources were antecedently available, before the great-ape lineage began to evolve?

2.1 *The Model*

Taking the ubiquitous laboratory rat as a representative example, I shall assume that all mammals, at least, are capable of thought—in the sense that they engage in computations that deliver structured (propositional) belief-like states and desire-like states (Dickinson, 1994; Dickinson & Balleine, 2000). I shall also assume that these computations are largely carried out within modular systems of one sort or another (Gallistel, 1990, 2000). For after all, if the project here is to show how non-domain-specific thinking in humans can emerge out of modular components, then we had better assume that the initial starting-state (before the evolution of our species began) was a modular one. I shall assume, however, that mammals possess some sort of simple non-domain-specific practical reasoning system, which can take beliefs and desires as input, and then figure out what to do (I shall return to this in a moment). Simplifying greatly, one might represent the cognitive organization of mammals as depicted in figure 5.1 (I shall return to the simplifications shortly).

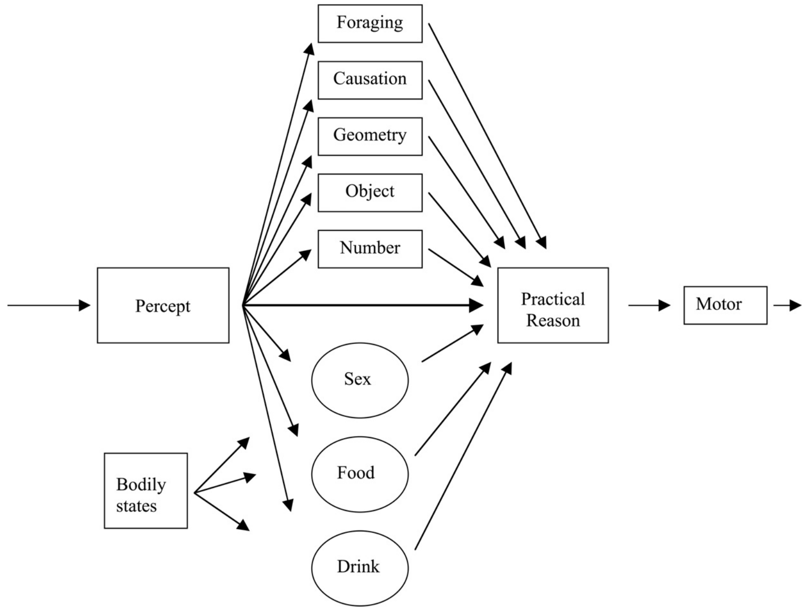

FIGURE 5.1 Rats (mammals?).

Here I am imagining a variety of input modules collapsed together under the heading "percept" for purposes of simplicity. (Of course I don't think that vision, audition, and so on are all really one big module; it is just that the differences between them don't matter for present purposes, and so don't need to be represented.) What *are* represented separately on the input side, however, are a set of systems for monitoring bodily states, which play an important role in the generation of desires (hunger, thirst, and so on). Then at the output end, I imagine a variety of motor-control systems collapsed together for our purposes under the heading "motor." And in between these two, I imagine a variety of belief- and desire-generating central modules, together with a practical reasoning system that receives its inputs from them (as well as from perception).

I assume that the practical reasoning system in animals (and probably also in us) is a relatively simple and limited-channel one. Perhaps it receives as input the currently strongest desire and searches among the outputs of the various belief-generating modules for something that can be done in relation to the perceived environment that will satisfy that desire. So its inputs have the form DESIRE [Y] and BELIEF [IF X THEN Y], where X should be something for which an existing action-schema exists. I assume that the practical reasoning system is *not* capable of engaging in other forms of inference (generating new beliefs from old), or of combining together beliefs from different modules; though perhaps it *is* capable of chaining together conditionals to generate a simple plan—for example, BELIEF [IF W THEN X], BELIEF [IF X THEN Y] → BELIEF [IF W THEN Y].

As for the modules that appear in the diagram, there is pretty robust evidence for each of them—at least, qua *system* if not qua *modular* system.[1] Thus there is plenty of evidence that rats (and many other animals and birds) can represent approximate numerosity (Gallistel, 1990); and there is evidence from monkeys, at least, that simple exact additions and subtractions can be computed for numbers up to about 3 (Dehaene, 1997; Hauser, 2000). Moreover, there is the evidence provided by Cheng (1986) that rats have a geometrical module, which is specialized for computing the geometrical relationships between the fixed surfaces in an environment (Gouteux & Spelke, 2001), and which they use especially when disoriented. In addition, there is the evidence collected by Dickinson and Shanks (1995) that rats make judgments of causality that closely mirror our own (including, apparently, essentially the same dispositions toward *illusions* of causality, in certain circumstances).

My guess is that many of the beliefs and desires generated by the central modules will have partially indexical contents. Thus a desire produced as output by the sex module might have the form "I want to mate with *that* female," and a belief produced by the causal-reasoning module might have the form "*That* caused *that*." So if the practical reasoning system is to be able to do anything with such contents, then it, too, would need to have access to the outputs of perception, to provide anchoring for the various indexicals—hence the bold arrow in figure 5.1 directly from percept to practical reason. The outputs of the practical reasoning system are likely to be indexical too, such as an intention of the form "I'll go *that* way."

2.2 *Adding Complexity to the Model*

One way figure 5.1 is oversimplified is that it probably radically underestimates the number of belief- and desire-forming modules that there are. This is especially true on the desire side, where of course all animals will have systems for generating pains/desires to avoid current noxious stimuli; and all animals will have systems for generating various emotions, such as anger (normally involving a desire to attack), fear (normally involving a desire to retreat), and so on. In addition, among social animals there will be systems for generating desires for such things as status. Similarly on the belief side, there will often be systems for kin-recognition and for computing degrees of relatedness, systems for recognizing an animal's position in a dominance hierarchy, and so on.

Another way figure 5.1 is probably oversimplified is that there may well exist informational relationships among the various belief-forming modules, in particular.[2] Thus one of the main functions of the numerosity module is to provide

[1] The only case in which there is direct robust evidence of modularity that I know of concerns the geometric system, which does appear to be isolated in the right kind of way from the rest of cognition. See Cheng (1986) and Hermer and Spelke (1994).

[2] Note that this means that the thesis of this chapter isn't that *no* integration of central-modular outputs takes place without language. Rather, the claim is that the mind's capacity to combine together central-modular contents will have been limited, prior to the evolution of language, and that the appearance of language makes such cross-modular integration well-nigh ubiquitous.

inputs to the foraging system, helping to calculate rates of return from various sources of food (Gallistel, 1990). I have not attempted to represent these in the diagram, partly for simplicity, partly because I have no fixed views on what the relevant informational relationships among modules actually are.

Another source of complexity that goes unrepresented in figure 5.1 is that each modular system presumably has some sort of domain-specific memory function attached. For central/conceptual modules don't just generate information "online," of course, for use in current practical reasoning. They are also going to be implicated in learning, and in generating new standing-state beliefs. So a more accurate version of figure 5.1 should probably show each central module as dividing into two components—a processing subsystem for generating domain-specific information and a domain-specific memory store for recording (some of) that information. Presumably, too, the processing subsystem should be able to access its proprietary memory store in the course of its computations, hence providing a constraint on the degree to which it is informationally encapsulated.

The final sort of oversimplification I want to mention is that there should probably also be links between (some of) the belief modules and (some of) the desire modules. For example, one would expect that information about degrees of relatedness (generated by the kin module) should be available as input to the module charged with generating sexual desire, suppressing the processes that would normally produce such desires in appropriate cases. And one might expect that information about rich or unusual sources of food (generated by the foraging module) might be taken as input by the hunger-generating module, sometimes causing a desire for food where there was none antecedently. And so on. In addition, one might expect that the content of whatever happens to be the currently strongest desire should have an impact upon the belief-generating modules, directing them to search for information that might help to satisfy that desire.

Although figure 5.1 is pretty simple, therefore, I don't really want to say that animal minds are that simple. The relevant claim to take away from the discussion is just that in all mammals (and so, *a fortiori*, in those mammals that were the immediate ancestors of the great ape lineage) there is probably a complex layer of belief- and desire-generating modules located between the various perceptual systems and some sort of limited-channel practical reasoning system.

3 Earlier Species of *Homo*

What changes began to occur in the basic mammalian cognitive architecture, during the evolution of the great apes and the transition to modern *Homo sapiens*?

3.1 *Deepening Modules*

At some point in the evolution of the great-ape lineage—whether in the common ancestor of ourselves and the chimpanzees or perhaps later, during the development of *Homo*—changes would have begun to occur. These were not initially changes of an architectural sort, I suggest. Rather, some of the existing suite of modules were deepened and enlarged, rendering their processing much more sophisticated; and perhaps some new modules were added, such as the social-exchange/cheater-detection module investigated by Cosmides and Tooby (1992; Fiddick et al., 2000). Thus some sort of social relationships module gradually developed into the beginnings of a mind-reading module; the foraging module became increasingly sophisticated, developing into a system of naive biology; the causal reasoning system developed into a form of naive physics; the object-property system expanded greatly to allow for many more object categorizations; and so on. The result is represented in figure 5.2.

I don't mean to claim that all of these changes occurred at the same time or for the same reason, of course. There is a complex story eventually to be unraveled here about the differing pressures—and no doubt their complex interactions—that led to the gradual deepening of different modules at various points in our ancestry. And for any given such change, it is still highly controversial at what evolutionary stage it first occurred.[3] Nor do I wish to claim that *all* modules have undergone a similar sort of transformation. On the contrary, it may well be that we still operate with essentially the same system for calculating approximate numerosities that is present in rats, for example.

By the time of *Homo ergaster* some quarter of a million years ago, all of the relevant changes would surely have taken place. Members of this group of species were plainly much smarter than present-day chimpanzees. They were able to move out of Africa into a wide variety of environments throughout Asia and Europe, including extremely harsh subtundra habitats. They had sophisticated stone-tool technologies. They were capable of adapting quickly to, and extracting relevant information concerning, wide variations in flora and fauna (Mithen, 1990). And all the evidence points to highly social creatures, capable of sustaining the complex social and personal relationships necessary for survival in such harsh environments, for the rearing of children with increasingly long periods of maternal dependency, and so on. (See Mithen, 1996, for a review of the evidence.)

Some of the data suggest, however, that members of *Homo ergaster* were *not* capable of the main elements of distinctively human thinking (Mithen, 1996).[4] Specifically, they weren't capable of creative thinking, or of generating radically new ideas. On the contrary, their stone-tool industries, for example, displayed long periods of stasis, with no significant changes of design over tens of thousands of years. And they don't appear to have been capable, as we are, of conjoining together ideas across modular boundaries. There is no sign that ideas concerning naive physics and ideas from naive biology could be combined to lead to the

[3] See, e.g., Povinelli (2000), for evidence concerning the relative shallowness of the mind-reading and naive physics modules possessed by our nearest cousins, the chimpanzees.

[4] Others have argued that distinctively human thinking emerged much earlier than the first arrival of *Homo sapiens sapiens* 100,000 years ago (McBrearty & Brooks, 2001), claiming that appearances to the contrary are an artifact of small sample sizes. If these views should prove to be correct, then they would only make my task that much easier, since they would allow greater time for the elements of distinctively human thinking to evolve together with language. I prefer to work with the more demanding assumption of late emergence.

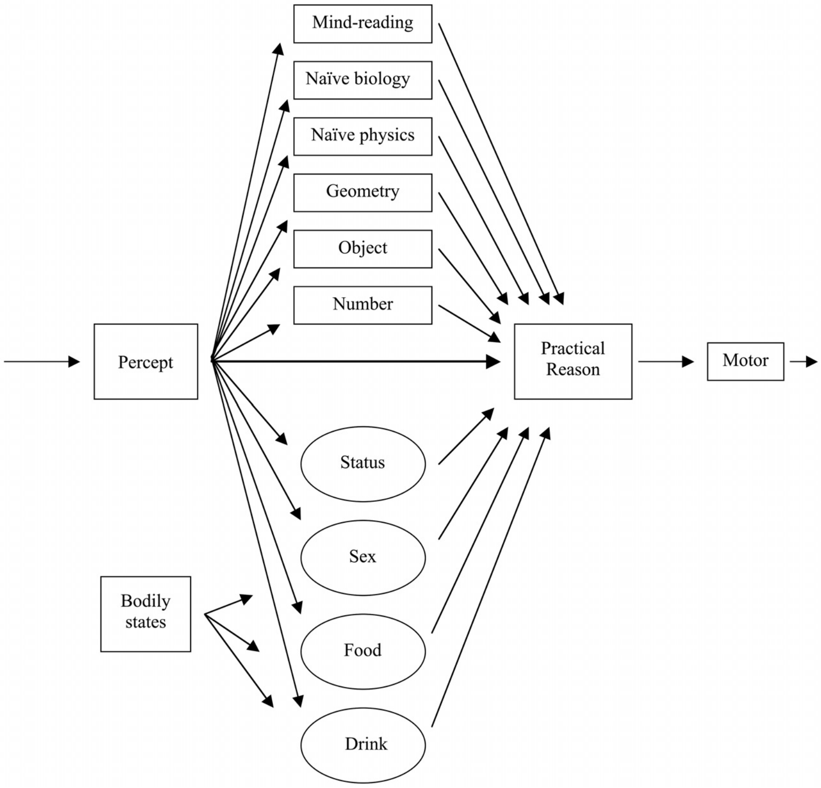

FIGURE 5.2 Homo ergaster (great apes?).

development of specialist stone hunting tools, such as we find in connection with *Homo sapiens sapiens*. Nor is there any evidence of analogical linkages between animal and social domains, such as we find in modern totemism, in the famous lion-man figurine from Hohlenstein-Stadel in Germany, and so on. It is for these reasons that I say the basic mammalian cognitive architecture was unchanged in members of *Homo ergaster* and before.

3.2 *Developing Imagination*

There is one further point I want to pick up on, resulting from the deepening of modules. This is that the extensive development and enriching of the object-property system would have made possible simple forms of sensory imagination. For the evidence is that imagery deploys the same top-down neural pathways in our perceptual systems that are deployed in normal perception for purposes of object-recognition (Kosslyn, 1994). As the number and range of object-categorizations available to our ancestors greatly increased (as it plainly did), so increasing pressure would have been put on the mechanisms concerned with object-recognition, leading to further strengthening of the top-down pathways used to "ask questions" of degraded, incomplete, or ambiguous input. It seems quite likely, then, that *Homo ergaster* would have been capable of generating visual and other images, even if this capacity was rarely used outside of the demands of object-recognition.

In fact, however, there is evidence of the use of rotated visual images by members of *Homo ergaster* some 400,000 years ago. This comes from the fine symmetries that they were able to impose upon their stone tools, while using a reductive technology that requires the planning of strikes some moves ahead. For Wynn (2000) makes out a powerful case that this can only be done if the stone-knapper is able to hold in mind an image of the desired shape that the stone would have when seen from the other side, rotating it mentally in such a way as to compare it with the shape of the stone now confronting him.

Then, given that members of *Homo ergaster* were capable of forming and manipulating mental images outside of the context of object-recognition, it may well be the case that they also used such images for purposes of *mental rehearsal* more generally. If they could form an image of an action they were about to perform, for example, then that image would be processed by the input systems in the usual way, and made available to the suite of central modules, some of which might then generate further predictions of the consequences of that action, and so forth. At any rate, this sort of mental rehearsal looms large in the cognition of our own species, as I will show hereafter; so it is interesting to note that it may well have been available to some of our more immediate ancestors as well.

4 The Emergence of Language

Most people think that language was probably a late-emerging capacity in the hominid lineage. Some people go so far as to put the emergence of language at the time of the "creative explosion" of the upper Paleolithic period, just 40,000 years ago and well after the appearance of anatomically modern humans some 60,000 years earlier (Noble & Davidson, 1996). Others wonder cautiously whether the Neanderthals might have had language (McBrearty & Brooks, 2001). But most are inclined to put the emergence of grammatical, syntax-involving, natural language with the first appearance of our species—*Homo sapiens sapiens*—about 100,000 to 120,000 years ago, in southern Africa (Bickerton, 1990, 1995; Stringer & Gamble, 1993; Mithen, 1996).

It does seem quite likely that some later species of *Homo ergaster* (including the Neanderthals) may have spoken a form of what Bickerton (1990, 1995) calls "protolanguage," similar to pidgin languages and the languages spoken by two-year-old children. This would be a system of spoken signs, with some distinction between nouns and verbs, perhaps, but with little other grammatical structure. Such "languages" have considerable utility (there is quite a lot that you can communicate using a pidgin language, for example), but they place considerable demands on the interpretational—mind-reading—skills of their hearers. This is because utterances that consist only of strings of nouns and verbs tend to be multiply ambiguous. Indeed, it may well be that the increasing use of protolanguage was one of the major pressures leading to the evolution of a full-blown sophisticated mind-reading capacity as we now know it (Goméz, 1998).

4.1 A Language-Involving Architecture

It seems likely, then, that at some point around the cusp of the first appearance of *Homo sapiens sapiens*, a system for processing and producing full-blown grammatical language began to emerge. I assume, as is now conventional, that this system divides into a core knowledge-base of grammatical and phonological knowledge, subserving separate production and comprehension systems. The result is depicted in figure 5.3, with all of the previous belief- and desire-generating modules now collapsed together for simplicity (and now with a double arrow between them to accommodate the fact, acknowledged earlier, that some belief modules deliver their outputs as input to some desire modules, and so forth).

At the protolanguage stage, I presume that the messages to be communicated were either the domain-specific outputs of one or other of the conceptual modules or the results of practical reasoning (such as an intention to act). So the causal sequence would go like this: first there exists a domain-specific propositional thought, generated by a central module, which the agent wants to communicate.[5] The agent then marshals a learned vocabulary and the resources of the mind-reading system to produce an utterance that is likely to convey that thought to a hearer, given the context. And in order for the hearer to be able to do anything with that thought, it has to be made available to the various belief- and desire-generating central systems. (At this stage, agents have no other inferential resources available to them, I am supposing.)

Similarly, with the emergence of the modern language-faculty, at least initially: each spoken sentence would be an encoding into grammatical language of a thought that is the output of a central module (or of the practical reasoning system); and each comprehended sentence would be made available to the full suite of central modules. The language faculty, then, is a unique kind of module, producing a radical new architecture to cognition. This isn't just because it is simultaneously both an input and an output module (though that is part of the

[5] Does the desire to communicate these domain-specific thoughts presuppose that there is some system—presumably the mind-reading system—that has access to the outputs of all the others? If so, then it might be said there was *already* a system capable of linking together the outputs of all modules prior to the evolution of a language faculty, namely, the mind-reading system. However, that a system can take any contents as input doesn't mean that it is capable of combining those contents together into new thoughts, or of deriving arbitrary inferences from those inputs. Moreover, at least two other mechanisms to underpin these early forms of communication can be envisaged that are much more modest in their requirements. One is that people should be disposed to express in language information that is highly *salient*. The other is that they might operate via a form of subvocal *rehearsal*, of the sort that arguably becomes ubiquitous in contemporary humans (see hereafter). That is, people might rehearse potential utterances in imagination, selecting those that have the greatest number of relevant *effects* (upon themselves). It is far from obvious that either of these proposals should require intermodular communication to be taking place already at this early stage.

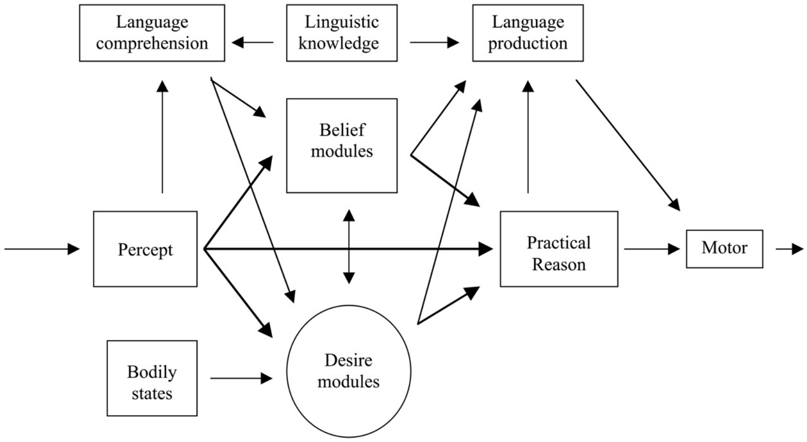

FIGURE 5.3 Archaic homo sapiens?

explanation). It is also because it is a module that—uniquely—feeds into, and draws outputs from, all of the central modular systems. This makes it a kind of "supermodule." It also means that there is a sense in which it isn't domain specific, since it can draw inputs relating to any domain. But in another sense it *is* domain specific. For the language faculty isn't interested in the *contents* of the thoughts it receives per se. Its job isn't to draw inferences from a belief as input to generate other beliefs, for example. Rather, its job is just to formulate that thought into a syntactically acceptable sentence. Since the function of the language faculty is to produce and decode linguistic utterances, *that* is its proper domain.

4.2 *Interfacing Language and Other Modules*

What kind of interface would need to have been built to enable the language faculty to communicate with the central modules? On the production side, this is (initially) relatively straightforward, at least in the sense of meshing with classical accounts of sentence production (e.g., Levelt, 1989). For each of the central modules would already have been charged with producing propositional outputs. The task for the language faculty is just that of mapping these outputs onto a sentential structure.[6]

[6] In fact this task seems likely to be somewhat more complex than is often supposed. For although the geometric module will deliver outputs that are propositional—in the sense of having combinatorial structure of some sort—it seems unlikely that those outputs will already be such as to contain concepts like "left" and "right." (This may be the reason why such words are so difficult for children to learn. See Shusterman & Spelke, chapter 6 here.) So those outputs will need to be transformed into the appropriate conceptual structures before the process of encoding into language can take place.

But how does comprehension work? How does the comprehension subsystem of the language faculty provide inputs for the central modules? *Some* of these modules would already be set up to accept propositional inputs from *some* other central modules. But this wouldn't by any means provide for global availability of propositional contents. Nor would this provide any obvious way for the comprehension subsystem to take a sentence with a content that crosses modular boundaries (once that becomes possible—see hereafter) and to "carve it up" into appropriate chunks for consumption by the relevant domain-specific central modules.

There are perhaps a number of different ways this problem could have been solved, in principle. But I suspect that the way it was *actually* solved was via the construction of mental models. There is quite a bit of evidence of the role of mental models in discourse comprehension (see Harris, 2000, for reviews). And a mental model, being an analog quasi-perceptual structure, has the right format to be taken as input by a suite of central modules that were already geared up to receive perceptual inputs. So I suspect that the process goes something like this: upon receipt of a sentence as input, the comprehension system sets about constructing an analog model of its content, accessing semantic knowledge, and perhaps also relevant background beliefs. The resulting structure is then presented to all central modular systems as input. (These structures might also be stored in existing perceptual memory systems, in effect creating a virtual non-domain-specific memory system. See sec. 5.)

4.3 *Combining Contents in Language*

Returning now to the question of how domain-specific thoughts are encoded by the production subsystem of the language faculty—how can such thoughts be combined into a single non-domain-specific sentence? Some aspects of this are relatively easy to get a handle on. Suppose that the output of the geometric module is the thought THE FOOD IS IN THE CORNER WITH THE LONG WALL ON THE LEFT, while the output of the object-property system is the thought THE FOOD IS BY THE BLUE WALL.[7] Our problem is to understand how these two thoughts can be combined together to produce the single non-domain-specific sentence "The food is in the corner with the long blue wall on the left." Given that we are supposing that there is already a system for encoding thoughts into language, this reduces to the problem of understanding how this sentence might be generated from the two sentences "The food is in the corner with the long wall on the left" and "The food is by the blue wall."

Two points are suggestive of how this might be done. One is that natural language syntax allows for multiple embedding of adjectives and phrases. Thus one can have "The food is in the corner with the *long* wall on the left," "The food is in the corner with the *long straight* wall on the left," and so on. So there are already

[7] I here follow the usual convention of using capitals to designate sentences of Mentalese, reserving quotation marks to designate sentences of English.

"slots" into which additional adjectives—such as "blue"—can be inserted. The second point is that the reference of terms like "the wall," "the food," and so on will need to be secured by some sort of indexing to the contents of current perception or recent memory~~,~~—in which case it looks like it would not be too complex a matter for the language production system to take two sentences sharing a number of references like this, and to combine them into one by inserting adjectives from one into open adjective-slots in the other. And there would surely have been evolutionary pressure from the demands of swift and efficient communication for the language faculty to evolve just such a capacity.

5 Distinctively Human Thinking

We are already in a position to see how the addition of a language module to the preexisting modular architecture might provide *one* of the distinctive elements of human thought, namely, its capacity to combine together contents freely across modular domains. But we have, as yet, said nothing to suggest why tokens of natural language sentences should qualify as *thoughts*. From the fact that we can express, in speech, contents that cross modular domains, it doesn't yet follow that we can *reason with* or otherwise make use of those contents in any of the ways distinctive of thinking.

5.1 *Using Language in Thought*

As a first step toward seeing how the language faculty might underpin distinctively human thinking, recall a point made earlier, that modular input and output systems have substantial back-projecting neural pathways that make possible different forms of sensory and motor imagery; and that such images are processed by perceptual input-systems in the usual way, just as if they were percepts. Assuming that the same is true for language, then sentences formulated by the production subsystem could be displayed in auditory or motor imagination, hence become available to the comprehension subsystem that feeds off perceptual inputs and, via that, to all of the various central-process modules.

Cycles of activity would thus become possible, as follows. In response to perceptual or linguistic input, the central modules generate a variety of domain-specific outputs. These are made available to the language faculty, which combines some of them into a sentence that is displayed in imagination, processed by the comprehension subsystem, and made available to the central modules once again. The latter process the resulting input, generating new domain-specific output, which is again made available to the production subsystem of the language faculty, which formulates some of it into a new sentence; and so on. While there is no reason to think that this could be the *whole* of human thinking, it does suggest a way in which—given sufficient cycles of domain-specific activity—new non-domain-specific ideas and beliefs might be generated, which could go well beyond anything manifest in the initial input.

What, then, are the other main elements of distinctively human thinking that need to be explained? One, surely, is *creativity*. Humans are capable of creating new ideas that don't just go *beyond* the input but appear to be wholly unrelated to it. Humans engage in fantasy and pretence in which they create imaginary worlds quite

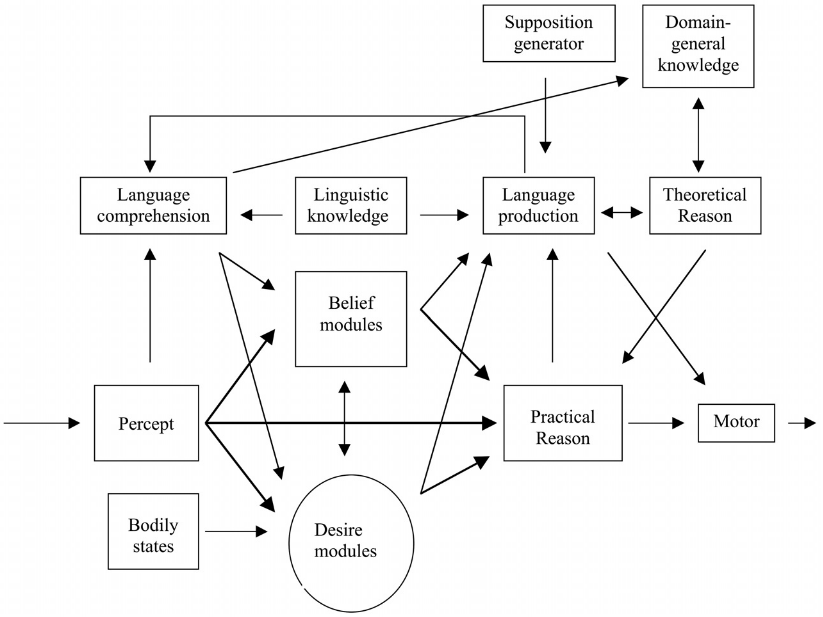

FIGURE 5.4 Homo sapiens sapiens.

unlike the real world. And humans are capable of forms of insight in which new ideas or new theories are produced that radically extend anything previously available. These capacities are not yet accounted for on the foregoing model.

Another main element in distinctively human thinking, however, concerns what humans *do* with new ideas once they are produced. Plainly, they can be remembered; so we need some sort of non-domain-specific memory system. But they can also be *evaluated*. We can take a new idea and decide whether or not it is a good one. Or we can consider two competing hypotheses and judge which of them is the better, and so on. When these functions are added to the architecture of figure 5.3, we get something like that depicted in figure 5.4.

Here four main elements have been added to the previous language-involving architecture. First, an arrow has been added backward from language production to language comprehension, enabling cycles of linguistic and domain-specific cognitive activity to occur in "inner speech." Second, a box for non-domain-specific memory has been added, taking input both from language comprehension (so that people can believe and remember what they are told) and from theoretical reason (see hereafter). Third, a supposition generator has been added, providing input to the language production system. Its function is to generate new sentences whose contents aren't produced from the outputs of the various central modules. Fourth, a box for theoretical reason has been added, which takes inputs from language production and domain-general memory, and which provides outputs to both domain-general memory and to practical reason, so that decisions on which sentences to accept can be both recalled and acted upon.

How radical would these departures be from the previous modular architecture, as represented in figure 5.3? And how plausible is it that these new functions could make their appearance within a relatively short time-span subsequent to (or coincident with) the evolution of the language faculty? Providing for the first two functions should have been relatively simple, as I have already shown. Thus there is every reason to think that the language faculty, like other input and output systems, would have been set up in such a way as to make it possible to display output-sentences in imagination, so that they can then be consumed by the input comprehension subsystem; hence making possible cycles of modular activity of the sort envisaged earlier. Moreover, if the comprehension subsystem operates by constructing analog mental models, as suggested earlier, then the results could be stored in existing perceptual memory systems—thus de facto creating a system of domain-general knowledge, given that the sentences comprehended can have non-domain-specific contents. But what of the supposer? And what of a faculty of theoretical reason? Is it plausible that such domain-general functions could have been built within the time-frame available, and that their operations should be computationally tractable?

5.2 *Supposing and Pretending*

In the case of the supposer, there is some reason to think that a simple initial disposition to generate new sentences for consideration—either at random or drawing on similarities and analogies suggested by perceptual or other input—might be sufficient. I have argued elsewhere (Carruthers, 2002b) that it is just such a disposition that gives rise to the ubiquitous and distinctive phenomenon of *pretend play* in human children; and that the function of such play may be to practice and hone a capacity for relevant and fruitful creative thinking. Here I shall be brief.

Consider the case of a young child pretending that a banana is a telephone. The overall similarity in shape between the banana and a telephone handset may be sufficient to activate the representation *telephone*, albeit weakly. If the child has an initial disposition to generate an appropriate sentence from such activations, she will then construct and entertain the sentence "That is a telephone." This is then comprehended and processed, accessing the knowledge that telephones can be used to call people, and that Grandma is someone who has been called in the past. If Grandma is someone whom the child *likes* talking to, then this may be sufficient to initiate an episode of pretend play. By representing herself *as* making a phone call to Grandma (using the banana), the child can gain some of the motivational rewards of talking to her. The whole sequence (including the initial generation of the supposition "That is a telephone") is then reinforced, making it more likely that the child will think creatively again in the future.

From such simple beginnings one can imagine that children gradually build up a set of heuristics for generating fruitful suppositions—relying on perceptual and other similarities, analogies that have proved profitable in the past, and so on. And with such a suppositional faculty up and running, the generative powers of the cognitive system represented in figure 5.4 would become radically transformed, becoming much less dependent upon perceptual and spoken input for its operations, and arguably displaying just the kinds of creativity in thought and behavior that we humans evince.

5.3 *Inference to the Best Explanation*

As for the faculty of theoretical reason, we need first to consider what such a faculty should contain. As I envisage it, a theoretical reasoning faculty is basically a faculty of inference to the best explanation, of the sort employed in science. While no one any longer thinks that it is possible to codify the principles involved, it is generally agreed that the good-making features of an explanation include such features as: *accuracy* (predicting all or most of the data to be explained, and explaining away the rest); *simplicity* (being expressible as economically as possible, with the fewest commitments to distinct kinds of fact and process); *consistency* (internal to the theory or model); *coherence* (with surrounding beliefs and theories, meshing together with those surroundings, or at least being consistent with them); *fruitfulness* (making new predictions and suggesting new lines of inquiry); and *explanatory scope* (unifying together a diverse range of data). Such principles are routinely employed in everyday life as well as science, of course, in thinking about a wide range of subject matters. And it is arguable that hunter-gatherers concerned with tracking prey will employ just such principles in the course of a hunt (Carruthers, 2002a; Liebenberg, 1990). So such a faculty very probably has a considerable ancestry, and would have been of vital adaptive significance.

There is some reason to think that a good proportion of these principles would come to us "for free" with the language faculty, however. (This point is argued more fully in Carruthers, 2003c.) For a strong case can be made for the vital role of considerations of *relevance* in the production and comprehension of speech (Sperber & Wilson, 1986, 1995). And there are two basic determinants of relevance, on such an account. First, utterances are relevant to the extent that they minimize the processing effort required to generate new information from them. Second, utterances are relevant to the extent that they issue in large amounts of new information. One would therefore expect that, when these principles are turned inwards, coopted for use in deciding whether or not to *accept* (believe) an internally generated sentence, they would lead to a preference for *simple but fecund* theories. That is, we should prefer statements that yield as much information as possible (unifying or predicting the maximum possible extent of data) but do so economically.

The other main strands in inference to the best explanation are then consistency and coherence with surrounding theories. There is no reason to think that this should require the introduction of anything radically new into the cognitive system, I think. Consistency with other beliefs can be checked by running the sentence that is up for acceptance back through the comprehension system, building a model of its content that can be compared with those already stored in non-domain-specific memory, and making its content available to the various domain-specific modules and their associated memory systems. Coherence can be checked by forming a conjunction of the sentence in question and any other theoretical belief, subjecting that conjunction to the processes just described.

If this account is along the right lines, then it is somewhat misleading to talk about a "faculty" of inference to the best explanation, and to represent it with a box in figure 5.4. For it doesn't have to be a functionally separate system, with a distinct neural realization. Rather, it is a sort of "virtual" faculty, built out of the operations of other systems. For there would already have had to be in place some system for deciding whether or not to believe a sentence received as input—that is, for deciding whether or not to accept the testimony of another person. And the relevance-theoretic preference for simple but fecund statements would already have been built into the language-interpretation system. What you get when imaged sentences of natural language are created by the supposition generator and cycled through the system would thus be a functional equivalent of a faculty of inference to the best explanation.

It appears, then, that none of the additions and changes necessary to transform the figure 5.3 architecture into the figure 5.4 architecture is especially demanding; nor is it implausible that those changes might have taken place within the relatively short time-frame available—either coincident with, or within a few tens of thousands of years of, the evolution of a language faculty. In which case it would seem that the main elements of distinctively human thinking can be secured from domain-specific modular components with a minimum of additional non-domain-specific apparatus. All that is needed, in effect, is a non-domain-specific memory system supervening on existing perceptual memory systems, and a disposition to generate new suppositions/sentences for consideration. The remainder of the new elements in the figure 5.4 architecture can be secured by coopting resources already available.

5.4 *Outstanding Problems*

Of course it would be foolish of me to pretend that all of the problems involved in understanding distinctively human cognition have now been addressed, let alone solved. For one thing, there remains the question of how some central-modular outputs rather than others get selected for encoding into language. Would this require the existence of some sort of general problem-solving executive system, overseeing the operations of all the other systems? If so, then the prospects for modeling human cognition in computational terms would not be looking too bright. For another thing, there remains the question of how the practical reasoning system can direct or moderate the activity of the central modules and the supposer, in such a way that those systems are directed toward the generation of contents that might prove useful in satisfying existing goals.

There is some reason to hope that the former problem can be understood in terms of the *salience* of different modular contents, where this might be modeled in terms of intermodular competition for scarce cognitive resources (Sperber, chapter 4 here). And one might expect that the latter problem could be addressed in terms of the operations of a variety of *attentional* systems, which either direct the various modules to work on some aspects of perceptual input rather than others, or cue those modules to be interested in certain sorts of contents rather than others, or both.

Perhaps a more serious problem, for my account, is to explain how domain-general knowledge can become *practical*. For all that I have really done so far is to explain how domain-general *sentences* might be generated and accepted. But how do these sentences then get to have an impact upon practical reasoning, and upon action? One option would be to say that there is a distinct parser/interface for the practical reasoning system that can take a natural language sentence as input and produce a representation in the right format to be processed in practical reasoning.

But this isn't a very attractive option for me, since it multiplies the number of computationally serious mechanisms that would need to be postulated in explaining how language comes to be the medium of intramodular integration. But it is still a possible option. After all, pressures of efficiency in communication alone might have been enough to explain the increasing use of language to combine the outputs of a number of different modules. And then there might have been selective advantages if the practical reasoning system could evolve a language interface so that it could take these crossmodular inputs directly, using them as a basis for action.

The more attractive option, for me, is to use a combination of three ideas: (1) cycles of linguistic activity in inner speech, (2) the use of mental models in speech comprehension, and (3) the access of the practical reasoning faculty to perceptual inputs. Here, then, is how the story might go. The crossdomain sentence "The toy is in the corner with a long blue wall on the left" is constructed and displayed in auditory imagination, thereby being taken as input by the language comprehension subsystem. That system sets to work to build a mental model of its content, where such a model is an analog quasi-perceptual representation. This model is then in the right format to be taken as input by the practical reasoning faculty, which must always have had access to perceptual outputs to underpin highly indexical planning in relation to the perceived environment. ("I'll take *that* one," "I'll go *that* way," "I'll fit *that* through *there* and then move it just *so*.") Then the practical reasoning faculty has access to both of the items of information that it needs (*long wall on left*, and *blue wall*) in order to achieve the goal of retrieving the toy, embedded within a single representation.

Admittedly, this story does emphasize that the role of mental models in my account is something of a hostage to fortune. Might it require the existence of some sort of General Problem Solver to construct a mental model from a given sentence as input? I hope not; and I don't see why it should; but I can't here demonstrate that it doesn't. However, investigation of these and other issues must await some future occasion. All that I can claim to have done here is to sketch a modular architecture that holds out the *promise* of understanding human cognition in modular and computationally tractable terms.

6 Conclusion

I have argued that it is both possible and plausible that distinctively human thinking should be constituted out of modular components, specifically components that implicate natural language. If this is the case, then those who argue against the thesis of massive modularity on the grounds that it cannot account for the non-domain-specific character of much human thought will have to find other reasons for their continued opposition. In fact, it looks like one can have one's massive modularity *and* one's non-domain-specific thinking and reasoning too. In addition, those who are already committed to believing in some form of massive modularity hypothesis now have an architectural proposal to investigate further.