Modular Representations of Cognitive Phenomena in AI, Psychology and Neuroscience

Joanna J. Bryson

The Artificial models of natural Intelligence (AmonI) Group
University of Bath, Department of Computer Science
Bath, BA2 7AY, United Kingdom
+44 (0) 1225 38 3934; fax: +44 (0) 1225 38 3493

`j.j.bryson@bath.ac.uk`

Abstract

Many architectures of mind assume some form of modularity, but what is meant by the term 'module'? This chapter creates a framework for understanding current modularity research in three subdisciplines of cognitive science — psychology, artificial intelligence and neuroscience. This framework starts from the distinction between *horizontal* modules which support all expressed behaviors vs. *vertical* modules which support individual domain-specific capacities. The framework is used to discuss innateness, automaticity, compositionality, representations, massive modularity, behavior-based and multi-agent AI systems, and correspondence to physiological neurosystems. There is also a brief discussion of the relevance of modularity to conscious experience.

Keywords: modularity, Fodor, Minsky, Chomsky, AI, psychology, neuroscience, systems, representation.

1   Introduction

Many of the architectures of mind described and referred to in this book assume some form of modularity. But what is considered to define a module varies a great deal both within and across the cognitive science disciplines: artificial intelligence (AI), psychology and neuroscience. This chapter is not devoted to any one architecture (though I have one too, which I will describe briefly in the Discussion to make my biases clear), but is rather an overview of the concepts and concerns of modularity. It covers all three of the above disciplines and shows how they relate to one another and to cognition — or at least to cognitive phenomena such as planning, learning, language, emotions, and consciousness.

My hope is that this chapter will serve as a useful primer — in the best case, a Rosetta stone — for both scientists and lay people trying to get a handle on what the various fields of cognitive science might mean by modularity, and how the modular architectures described in this book and elsewhere might correspond to our common understanding of what minds do. It is important to realise that researchers who are experts in one or more of the areas described below may have no awareness of some of the other areas, and therefore make no

effort to reconcile their own theories with the others. Consequently, this chapter contains some substantial redescription in an effort to put these theories into a common framework for comparison.

2   Modularity in Psychology

2.1   Criteria for Modularity

I will begin with an extremely simple definition of modularity from the psychological literature, due to Flombaum et al. (2002):

> "Modularity is the thesis that the mind contains independent input systems that, when engaged, are restricted in the types of information that they can consult."

This definition is useful for two reasons. First, it introduces a very clean criterion for modularity: that some part of the mind does not have access to some other part of the mind, or at least not its 'information'. Given this simple criterion, anyone who accepts the idea of implicit knowledge or unconscious behaviour has already acknowledged that there is some sort of modularity involved in human intelligence.

This is not the only possible characterisation of modularity. Fodor (1983) provides the best-known list of criteria for recognising modularity, some of which are now highly controversial, such as innateness. To be fair, the entire concept of innateness has become controversial because the lifelong interplay between genetics and environment makes many (particularly postmodern) developmental psychologists uncomfortable with the category (for example Thelen and Smith, 1994; Elman et al., 1996; Donnai and Karmiloff-Smith, 2000). Those who do not believe in innateness as a discriminative category in human development often do not believe in modularity either, because Fodor (1983) famously staked so much importance on the innateness criteria. This seems slightly ridiculous when coming from an AI perspective, because innateness has no bearing on the functional or computational characteristics of modularity, but it has had a large impact on the psychological modularity literature.

Psychologists who do believe in modularity are generally concerned with other traits such as automaticity in the presence of appropriate stimuli or brain localisation. Brain localisation I will discuss under neuroscience below. Automaticity is indicated both by speed of processing and by changes in processing due to nearly identical but saliently different stimuli which help the module select its own input. Modules are expected not only to be specialised to a domain, but also to be able to recognise the context in which their domain is present.

An impressive example of this is shown by Tanaka and Farah (1993) in the domain of face recognition. Recognising individuals from their faces is an extremely difficult, highly skilled behaviour which takes years to develop. Children tend to recognise people through superficial cues such as glasses and hairstyle. Adults with sufficient experience make discriminations on subtle differences between faces. Experience is critical: even adults often have trouble discriminating faces from less familiar races, while those with specialised experience, such as farmers and field researchers, can learn to discriminate the faces of other species. Nevertheless, face recognition has often been seen as a candidate

module because we do it quickly with no deliberate access to the process, and because the capacity to recognise faces can be lost through brain damage or stroke independently of any other capacity.

Tanaka and Farah (1993) contribute to this debate by demonstrating implicit, automatic context recognition by the face recognition capacity. Starting with photographs of famous faces, they divide the pictures down the middle, then shift one side slightly with respect to the other. Despite the fact that the misalignment boundary is quite conspicuous — subjects are aware that they see one face that has then slightly skewed — both speed and accuracy of recognition are substantially degraded. The Fodorian modularist's explanation: these slightly altered visual stimuli no longer trigger the 'face recognition' module.

In my own opinion, the most critical attribute of modularity is that individual modules support and are supported by different specialised representations (Bryson and Stein, 2001a; Bryson, 2002). Notice that this is fairly compatible with the inaccessibility definition of Flombaum et al. (2002) — process structure is heavily dependent on representational structure and content, so if we consider minds to be based on process, clearly the different processes might have difficulty accessing each other's knowledge or control (at least directly) if they are based on different representations. However, I came to this conclusion in the course of designing a development methodology for artificial intelligence, which I will discuss further below.

2.2   Fodor: Vertical and Horizontal Modules

The second reason that the Flombaum et al. (2002) quote is a useful introduction to modularity in psychology is the phrase "independent *input* systems". This makes clear the origins of a great deal of the theory underlying modularity in the psychological literature — the book *The Modularity of Mind* Fodor (1983). Although Fodor states that he believes modularity may also exist in motor systems (p. 42) he claims ignorance of these systems and concentrates on perception. An entire school of psychological research has followed this lead (recently Coltheart, 1999; Downing et al., 2001; Spelke, 2003), some (such as Flombaum et al. (2002)) apparently unaware that Fodor's full architecture is actually symmetric with respect to sensing and action.

Fodor himself cites Chomsky (1980) and Gall (1825, the originator of phrenology), as his main inspirations. Dawkins (1976) and Hume (1748) also give highly relevant discussions. But I will use Fodor as a basis for describing psychological modularity both because of his influence in psychology and because of his relevance to modular AI.

Fodor introduces the terms 'horizontal' vs. 'vertical' to describe two different sorts of decomposition for intelligence. For Fodor *horizontal* decompositions are those which identify processes which underlie all of cognition, such as memory, attention, perception, and judgement. *Vertical* decompositions identify particular skills or faculties, such as mathematics, language, and metaphysics, which each have their own characteristic processes of memory, attention and so forth (Fodor, 1983, pp. 14–21). Roughly speaking, evidence for horizontal decomposition is the extent to which, for a particular individual, performance across all domains is correlated, while evidence for vertical decomposition is the extent to which it is not. For example, it might turn out that individuals who have good memories tend to be able to remember things well across any domain; this would indicate that memory is a horizontal module. On the other hand, if how good an individual is a mathematics

in no way predicts how good they are language, then this is evidence that both mathematics and language are vertical modules.

Fodor believes that only certain parts of human intelligence are decomposed in the vertical sense; those parts being perception and action. In Fodor's system, a number of semi-autonomous perceptual modules run simultaneously giving quick, automatic analysis of the perceptual scene. Each module recognises its own best input, and effectively trumps any other module trying to process that input. The output of perception modules is in the 'language of thought'. This output is operated on by a horizontal reasoning system that then chooses an action. The chosen action is then presumably produced by a vertical action module, though as I've mentioned such action-skill modules are little researched or discussed in the Fodorian modularity literature. But we would expect such a module to take 'language of thought' as input and to generate patterns of muscular control as output.

Even if Fodorian psychology research did consider motor as well as perceptual modules, it would never consider the sorts of tightly-coupled perception-motor modules prevalent in artificial intelligence (for example Minsky, 1985; Brooks, 1991b; Albus, 1997, I discuss these further below). This is because, for Fodor, the purpose of vertical modules is to reduce the complexity of the real world into a common representation used by a horizontal, general-purpose reasoning system.

2.3   Massive Modularity and Evolutionary Psychology

The examples Fodor initially proposed of vertical modules (e.g. language and mathematics) are far higher-level skills than most Fodorian psychological modularists currently ascribe to modules. This is because of another characteristic Fodor himself attributed to modules: that they are atomic. This means that, for Fodor, modules are not composed of further modules. Since language has been demonstrated to have many independent constituent parts, it does not meet this Fodorian criteria, despite being one of the modules he originally discussed the most. Fodor believed this atomicity was necessary to his vision of lightning-fast, automatic, parallel modules vying with each other to interpret the world for the general-purpose reasoning system (and, presumably, to translate the general-purpose reasoning back out into actions in the world.)

Other researchers, coming particularly from evolutionary psychology, have a very different understanding of modularity (for example Cosmides and Tooby, 1994; Evans and Zarate, 1999; Carruthers, 2003). These researchers focus primarily on understanding why humans show greater computational abilities in some cognitive domains than others. That is, problems which are computationally equivalent are easier or harder to solve depending on the domain being reasoned about. For example, people are better at reasoning about relationships when they are expressed in terms of social characteristics and obligations than when they are presented as logical abstractions. People are also more capable of doing arithmetic involving fractions if problems are expressed in terms of the currency local to their country (this was easier to demonstrate before the British converted to a decimal system of change for the British Pound).

Here again there is a diversity of opinion about innateness. For some researchers, the leading indication that a module exists is if non-human primates are shown to share the specialised capacity. For example, the recent results indicating monkeys expect equivalent compensation as their peers for performing the same task (Brosnan and de Waal, 2003) is

taken as evidence for a cheater-detection module. Others consider any specialised capacity, such as face recognition described above, to be an indication of a module. They are happy to believe that modules develop or are learned. Such acquired modules could explain both the increased cognitive capacities of mature animals and their relative inflexibility — essentially a general-purpose learning substrate consolidates into regions of specialised skills and representations. Bates (1999) provides such an account of language learning, although she is not normally associated with 'massive modularity'. *Massive modularity* is the term applied for those who believe the adult mind consists perhaps entirely of specialised (vertical) skill modules.

3 Modularity in Artificial Intelligence

I mentioned briefly above that Fodor's theory of modularity was strongly influenced by contemporaneous work by, for example, Chomsky (1980). Chomsky's influence extends not only into linguistics and philosophy but also into computer science, particularly in artificial intelligence. Two of Chomsky's colleagues at MIT working in AI also made contemporaneous contributions to modularity research which have resulted in the widespread adoption of modularity in certain areas of AI.

Since the mid-1990s, modular approaches have dominated the development of 'autonomous' AI systems such as mobile robots or virtual reality (VR) characters (Hexmoor et al., 1997; Kortenkamp et al., 1998; Sengers, 1999; Thórisson, 1999; Bryson, 2000). These systems share with humans and other animals the characteristic of needing to be able to coordinate a large range of intricate expressed behaviours, many of which are only applicable in some of the variety of contexts the system may find itself in. These contexts include an environment which changes independently of the actions of the intelligent system and in ways the system cannot control.

This section offers a brief overview of four distinct approaches to modularity that have been developed in AI in the last twenty years. For more extensive reviews, see Bryson (2000, 2001).

3.1 Modules as Agents

The first well-known modular model of mind at least described by an AI researcher is the *Society of Mind* (Minsky, 1985). Although the book was published in 1985, Minsky had been working on and presenting the idea for some time before that (Doyle, 1983). Compared to Fodor's, Minsky's proposal is more substantially vertical, although it still has some horizontal elements. An individual's actions are determined by simpler individual agencies, which are effectively specialists in particular domains. Minsky's agencies *are* compositional — they exploit hierarchy for organisation. For example, the agency of play is composed of agencies of block-play, doll-play and so forth. Arbitration between agencies is also hierarchical, so the play agency competes with the eat agency for the individual's attention. Once play establishes control, the block and doll agencies compete.

Minsky's agents have both perception and action, but not memory, which is managed by a shared facility — presumably 'horizontal' to Fodor, though one that is still modularly decomposed. Memory (K) agencies are interconnected both with each other and with the

other, actor (S) agents. K agents and S agents can each activate the other type as well as others of their own type. Keeping the whole system working requires another horizontal faculty: the 'B brain' which monitors the main (A) brain for internally obvious problems such as redundancy or feedback cycles.

Minsky's model attempts to account for all of human intelligence, but has never been fully implemented. The existent systems described in the book, for example the learning system of Winston (1975), were for the most part fairly traditional, monolithic, single-problem AI systems with centralised control. Masters and Ph.D. students routinely resolve to fully and properly implement the Society of Mind model, but there is to date no widely-accepted canonical implementation.

3.2   Modules as Finite State Machines

In contrast, the term "behaviour-based artificial intelligence" (BBAI) was invented to describe a simplified but fully-implemented system, originally used to control mobile robots. This was the subsumption architecture (Brooks, 1986, 1991b). The subsumption architecture is purely vertical. The modules were originally each finite state machines (see Figure 1), and arbitration between them was conducted exclusively by wires connecting the modules — originally literally (Connell, 1990), but soon as encoded in software. Each wire could connect one module to another's input or output wires, the signal of which the first module could then either monitor, suppress or overwrite.
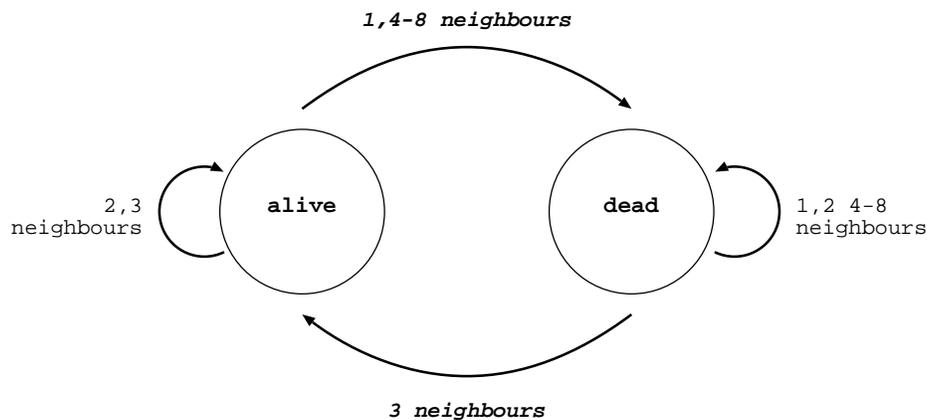
Figure 1: A finite state machine is an enumerated set of all the possible states the module can be in, plus the complete list of possible transitions between states, each labelled with the condition that would lead the module to make that transition. Example figure is for the cells in Conway's Game of Life (Gardner, 1970).

Brooks initially asserted that most apparent horizontal faculties (memory, judgement, attention, reasoning) were actually abstractions 'emergent from' (used to describe) an agent's expressed behaviour, but had no place in the agent's actual control (Brooks, 1991b, p. 146–147). However, his system was rapidly extended to have learning systems either inside modules or local to layers of modules (e.g. Matarić, 1990; Brooks, 1991a). My own opinion is that this is precisely where learning belongs, in specialised representations in the heart of modules. Unfortunately, what might have been a promising approach has generally been overlooked by most critics and followers of the subsumption architecture — they

were most enthralled by the attractive simplicity and radicalism of Brooks' deemphasis of representation and centralised control. In fact, many researchers are still convinced that Brooks' robots are 'stateless' (have no memory), despite the fact that finite *state* machines are at the core of his architecture, and do serve as the short-term memory necessary to react to events after they are sensed (for example, collisions with obstacles.)

3.3 Modules as Slaves and Bitmaps

Of the researchers who did *not* immediately adopt "no representation" as a mantra, most attributed the impressive success of Brooks' approach to the fact that he had created abstracted primitives — the semi-autonomous action/perception modules. Because these primitive units could sort out many of the details of a problem themselves, they made the composition of intelligence under *any* approach easier (Malcolm et al., 1989). Thus modular behaviour systems have been incorporated as a component into a large variety of AI architectures, many of which still maintain centralised, logic-based planning and learning systems (for example Gat, 1991; Bonasso et al., 1997). In fact, due to the difficulty of reasoning about relatively autonomous components, some systems have reduced behaviours to 'fuzzy rules' (Konolige and Myers, 1998) or vector fields (Arkin, 1998) which can be more easily composed.

Despite the lack of commonality of such approaches to Brooks' original ideal, they are still often called either behaviour-based or hybrid behaviour-based systems. Further, by the late nineties, the work of these researchers had so far outstripped that of the 'pure' BBAI researchers that two significant publications declared these hybrid approaches to have been conclusively demonstrated superior to pure BBAI (Hexmoor et al., 1997; Kortenkamp et al., 1998).

It is interesting to note that the systems with simplified, easily composed modules (e.g. Konolige and Myers, 1998; Arkin, 1998) are the AI systems closest to Fodor's ideal, although often the modules are for action, not perception. But they are simple, quick, one-step mappings from a goal constructed by a centralised / horizontal planning system to a set of motor commands to achieve it. On the other hand, they have lost many of the engineering advantages that Minsky and Brooks considered critical to modular AI. Intelligence is no longer decomposed entirely into simple elements. The planning systems are generally as elaborate as any in AI; they simply reason about more powerful elements.

3.4 Agents as Modules

At the other end of the modular-complexity spectrum are multi-agent systems (MAS) (Weiß, 1999; Wooldridge and Ciancarini, 2001). Here, the modules composing the system *are* agents, but not in Minsky's sense. Rather, these agents were meant at least initially to be themselves complete software systems — often the agents themselves use the sort of hybrid behaviour-based architectures just described (d'Inverno et al., 1997; Guzzoni et al., 1997).

MAS practitioners generally consider themselves to be modelling not individual minds, but societies. They nevertheless typically do have 'horizontal' modules / agents / components for connecting agents with complementary needs and abilities together (directory agents) or for enforcing behavioural norms of participants.

In some senses, MAS are actually closer to BBAI than the so-called hybrid behaviour-based systems. Each agent performs a particular task, and may have its own private knowledge store and representations which are presumably well suited to its function. However, to date there are a few fundamental differences between a MAS and a single, modular agent. These differences are due to issues of communication and arbitration between modules / agents. The MAS community is concerned with interoperability between unspecified numbers and types of agents, and with distribution across multiple platforms. This creates an administrative overhead not necessary for a single, modular agent. Where MAS are in fact limited to a single platform and a relatively fixed architecture, I suspect their engineers may in fact be taking the wrong approach, and should consider them to be modular single agents. But this is a topic for another paper (Bryson, 2003).

It is important to realise that, despite their high profile in some reseearch communities, MAS are not yet a proven technology (Edmonds, 2002). Unlike behaviour-based and hybrid behaviour-based systems, they do not yet have an extensive commercial application base.

3.5   Summary: AI and Mental Modules

AI provides us with working models of both Fodorian modular decomposition (in the form of hybrid architectures) and of massive modularity (in the form of more strictly modular architectures, such as behaviour-based AI and multi-agent systems.) This provision has been largely unintentional, though certainly influenced by some concerned researchers' theories of the nature of natural intelligence. But because they have been built into working systems they have been subjected to a special kind of selective pressure. These systems need to work, and in order to work they need to be relatively easy to design and debug. Thus the quest for success in AI leads to a sort of selective pressure for parsimony yet at the same time a need to be able to handle the complexity of the real world.

The net result of all this experience seems to be the following:

- Modularity is an important attribute for systems that have to interact with a complex, changing environment. It is used widely for mobile robotics, virtual reality and user interfaces. It has also often been suggested for managing networked resources (whether load balancing or exploiting e-services on the Internet), but these applications are not yet well established.

- *Pure* modularity is difficult to manage. If modules are both autonomous and simple, they tend to interfere with each other. Most modular systems now have some sort of behaviour arbitration. These systems run the gamut from top-down control by a reasoning system to negotiated solutions where each module acts as a voter. Some architects Blumberg (including 1996); Gat (including 1998); Sloman and Logan (including 1999); Bryson and Stein (including 2001a) think that intermediate architectures which reflect both top-down and bottom-up information will ultimately prevail. Such architectures require additional specialised structures, akin to Fodor's horizontal modules.

- Nevertheless, extremely quick, simple modules typical of Fodor's description of vertical modules are *not* the norm, although some examples of such an approach do ex-

8

ist. If hand-coded BBAI continues dominating applications (or is replaced by MAS), then this will be evidence that, for AI at least, it makes more sense for modules to be more intricate, mapping sensing clear through to action. To this extent, AI supports a model more like massive modularity.

For a more complete (if older) analysis along these lines, see Bryson (2000), or the slightly updated version of that work in Chapter 3 of Bryson (2001). I should say that, although this section has concentrated mostly on the pragmatic aspects of AI, that this is not meant to undermine the work by some philosophers and psychologists to work within the AI discipline at creating and understanding complete models of mind. Besides Minsky, see in particular Sloman and Logan (1999) as well as many of the chapter authors in this book.

4   Modularity in Neuroscience

We have evidence of at least three sorts of modular decomposition in mammal brains[1]: modularity by organ within the brain, by region within an organ, and by context or time. In this section I will describe each of these in more detail.

4.1   Modularity by Organ

We know that different parts of the central nervous system have radically different structure, in terms of different component cells, different amounts of connectivity, and different organisations of connectivity. Even if we did not have behavioural evidence (as we do) that the neocortex, cerebellum, hippocampus and so forth perform different functions, we would suspect as materialists — and given our understanding of computation in networks — that these organs must perform different computations, because of their different structure and connectivity. This point becomes even more obvious when we realise there is no particular reason not to extend the concept of organ modularity to more peripheral organs, such as the spinal cord, the retina or the cochlea.

The brain is normally considered to have three parts: the fore-, mid- and hindbrain (Carlson, 2000). Speaking roughly, the hindbrain seems necessary for coordinated action — animals that are missing much of their hindbrain produce jerky, uncoordinated actions and may have difficulty with balance. The midbrain contains much of the more basic or primitive control, including the encoding of complex species-typical behaviours. A cat with an intact hind- and midbrain can go through the motions of pouncing, stretching, sleeping and eating, but does not necessarily perform these behaviours in appropriate contexts. The forebrain is associated with connecting behaviours to contexts, or in more cognitive terms, with goal-oriented behaviour. In primates at least, this includes inhibiting more reflexive actions so that more complex strategies or learning can have a chance to achieve activation (Hauser, 1999).

This sort of task decomposition indicates that the best parallel within the Fodorian framework to organ-level modularity may well be horizontal decomposition. Organs seem dedicated to a sort of processing, not a particular context for perception or action.

---

[1]Most of this discussion is true of vertebrate brains in general, but I am most familiar with primate brains so I restrict my claims.

4.2 Modularity by Region

Even within an organ which is fairly structurally homogeneous (at least in considerations likely to affect the nature of its computations) there are differences in function. In some cases these seem to be determined primarily by connectivity: for example, the primary auditory and visual cortices are areas of the neocortex that most directly receive the sensory input of the two systems. It has been suggested that other regions are modular by function, such as the 'fusiform face area' or the 'parahippocampal place area' (Downing et al., 2001). However, given the amazing diversity of cortical computation even in single regions (for example Kauffman et al., 2002, who show that the 'visual cortex' is necessary for learning Braille, see below), it may be that such apparent specialisation also reflects connectivity. For example, the 'fusiform face area' may reflect links to subcortical brain organs specialised for purposes such as social interaction (perhaps the amygdalic system), and the 'parahippocampal place area' may reflect known links to navigation in the hippocampal system.

Some cortical regions are steps along a stream of processing, for example regions dedicated to identifying low-level features such as line orientations (Hubel, 1988), or to higher-level concepts such as categories of objects or tasks (Freedman et al., 2001) or personal identity (Perrett et al., 1992). If we are trying to map these regions into the sorts of modularity expressed by BBAI or massive modularity, we would have to consider a column of such representations as a single module, or need to hypothesise ensembles of modules acting in a coordinated manner.

4.3 Modularity by Context

Even within a given region, the semantics of a particular cell's firing seems to be dependent on the context in which it fires. This has been demonstrated in the hippocampus (Kobayashi et al., 1997), in sensory cortices mapping receptive fields (Sen et al., 2001), and in the prefrontal cortex (Asaad et al., 2000). I believe that the extent of the consequences of this *temporal* modularity have not been fully recognised. It may be that some computations are mutually exclusive because their representations cannot be active at the same time. Further, individual differences in developing these representations might account for individual differences in insight and generalisation based on the relative accessibility of two representations. See for example Skaggs and McNaughton (1998) for their account of individual differences in rats' ability to discriminate two similar rooms.

The concept of temporal modularity is in marked contrast to the descriptions of Fodor or the rationale for BBAI and of MAS, which all claim that the reason their modular approach is useful is because all modules are constantly active. However, some more established (and less conventionally modular) systems of cognitive modelling such as Soar (Newell, 1990) and ACT-R (Anderson, 1993), have found it necessary to create 'problem spaces' so that the system chooses actions only from an appropriate subset of available actions. The same is true of blackboard systems (Hayes-Roth, 1985). Such strategies are also to be found buried in the details of some well-known BBAI architectures (for example Blumberg, 1996).

Modularity by context must be vertical, since it is necessarily context / task specific. It also involves a significant amount of processing structure, encompassing from perception

to action. For example, emotional states might be seen as contextual modules — the state of fear reduces the problem of behaviour arbitration to a choice between fight or flight, yet all of an animal's senses and all of its muscles are available to effect these actions.

Although this sort of modularity violates Fodor's notions both by the lack of parallelism and by the end-to-end nature of its control, these are quite similar to the violations already mentioned in describing behaviour based AI and MAS. As such, the concept of temporal modularity might be very interesting to evolutionary psychologists and other massive modularists.

4.4   Summary: Modularity in Brains

As usual when looking at a real, evolved systems, the picture of modularity gets much more messy when looking at the brain. There is nevertheless strong evidence for modularity of some sort. Recall the definition of modularity I championed in the first section — there can be no question that various organs of the brain and various regions of organs have specialised, externally inaccessible representations and processes. However, task-based activation incorporates many disparate parts of the brain. Rather than dividing into horizontal *or* vertical modules, it seems that much of mental processing can be pictured as a criss-cross of activations between horizontal *and* vertical modules. Even V1, long 'known' to be the area of human neocortex associated with very basic interpretations of retinal (visual) data, has now been implicated for learning Braille (a tactile but still spatial skill) in both blind and sighted patients (Kauffman et al., 2002). Interestingly, this also requires contextual modularity — sighted patients find learning Braille much easier if they are blindfolded.

Why should the brain be modular? I suspect that evolution has found it useful for the same reason as software engineers have (Parnas et al., 1985; Coad et al., 1997): to combat the explosive combinatorial complexity of searching for the right solution to hard problems. During both evolution (for the species) and development (for the individual), preferred synaptic organisations are learned for recognising regularities that are useful for controlling actions and achieving goals. These regularities may arise from the external environment, as communicated by the senses, or from neighbouring neural systems — from the neuron's perspective there is no difference. This view of the development of modularity in the individual is similar to that of Bates (1999) and to some extent to that of Karmiloff-Smith (1992), whose work emphasises the developmental aspects of skill specialisation. On the species level, it echos the Livesey (1986) account of the evolution of brain organisation.

5   Discussion

The bulk of this chapter has been about the current state of the art in modularity. In this discussion I turn at least a bit more speculative. To begin with I will describe my own work (alluded to earlier) in behaviour coordination. Then I will dive into a brief discussion of the extent to which I have failed to discuss many of the cognitive entities generally associated with mind.

5.1 Module Coordination and Structured Action Selection

Although my main personal motivation is understanding how brains and minds work, the bulk of my research to date has been into the management and design of modular AI systems. This is because I think that modular AI systems are a good platform for modelling and thus understanding natural intelligence (Bryson and Stein, 2001b; Bryson, 2001).

Here is a brief summary of my current conclusions about the engineering of modular AI systems:

1. Semantic and task memory should be stored in specialised representations within modules. That is, specialised memory should be stored with the processes that exploit it. This is roughly consistent with the massive-modularist approach to psychology, and a slight elaboration on the position of BBAI, as described earlier.

2. Ordering the behaviour of such modules is best done using a specialised, horizontal module for sequencing behaviour. This sequencing module is not a full planning system, but rather a system for running established reactive plans (see Bryson and Stein, 2001a, for further details). As such, it is similar to the interface between the basal forebrain in mammals and parts of the midbrain implicated with storing action sequences (see further Mink, 1996; Redgrave et al., 1999; Carlson, 2000; Lonstein and Stern, 1997).

These are the systems that I have the most experience of building myself, but I believe there are other 'horizontal' modules that will be of general use that my systems are just not yet sophisticated enough to require. I suspect that it is useful to have coordination and smoothing of motor command conducted by a hindbrain-analogue module. There is already some evidence for this in the literature, for example the motor command of the Ymir architecture (Thórisson, 1999) or the work of various groups in modelling the cerebellum. To date, some of the smoothest control of complex robotics have come from monolithic 'dynamical systems' based mathematical control (for example Atkeson et al., 1997) but I suspect breaking this approach into modules will make it easier to scale it up to more complex tasks.

I also suspect that real agents need a hippocampus analog. The hippocampus seems to be an organ with highly-indexical, context-dependent relations which rapidly learns associations. This organ has been implicated as necessary for episodic memory, but also for navigation. My own current research is leading me to believe that these may follow from each other — that the organ may have originally been specialised to learning navigation, but became adapted for task learning in general, and episodic memory may have become one of the features of this capability. Unfortunately, it may not be a good idea for artificial agents to be built to include a primate-like capacity for task learning, as such a system would necessarily both slow down their processing and reduce their reliability (Bryson and Hauser, 2002). But it *is* clear that any human-like mind would need this kind of capacity.

5.2 Deliberation

Episodic memory is occasionally called 'declarative memory' because in humans it is the kind of memory you can talk about (have conscious access to), but this is a strange term in a sense since there is evidence of other animals having memories of isolated experiences,

but not of them declaring much of anything. Episodic memory is useful locally for keeping place within a task, for example knowing which parts of a familiar maze you have explored already *today* (assuming the maze is rebaited every day.) It might be useful long term as a source of cases for case-based reasoning, or as raw data to be compiled statistically into probability frameworks so that expectations can be used for planning or to disambiguate noisy sensory data.

On the other hand, the role of deliberation (or conscious attention to a task) still seems deeply mysterious to me. As already mentioned the accessibility difference that determines explicit from implicit knowledge *is* a key indicator of modularity. But I see no systematic difference (other than qualia) between conscious and unconscious thought other than a marked increase in cortical activity (Dehaene et al., 2001, 1998).

Unlike some researchers in AI, I am not convinced that consciousness is isomorphic with having self-knowledge — although clearly having a good representation of oneself is useful to planning. Nor is it with having language. Language almost certainly fundamentally *alters* the nature of consciousness, both by allowing shorthand concept reference in what is clearly a limited-capacity system, and by increasing coherence as a consequence of language's sequential temporal nature (Spelke, 2003; Bryson, 2002). But I could easily construct an AI straw-being that might have either or both of these attributes but not seem particularly more alive or aware than any other AI system.

The consciousness-related data that are currently intriguing me most are a number of recent results from a variety of laboratories (Siemann and Delius, 1993; Bechara et al., 1995; Greene et al., 2001) showing that:

1. humans can learn complex tasks without explicitly understanding them, and further

2. humans who *do* gain an explicit understanding show *no performance difference* from those who do not.

I suspect that two things are true. First, that I believe Dennett (2001) is absolutely right in suggesting that, as we come to understand consciousness, we will realise we have been covering several disparate functions with that one term, none of which are magic. Second, I believe that two of these functions will turn out to be focusing search for action selection and ordering behaviour in time.

6 Conclusions

This chapter has been an introduction to the idea of modularity as approached from three different disciplines: psychology, AI and neuroscience. It has attempted to create a common framework for discourse between these fields, leveraging (but not necessarily supporting) the decomposition originated by Fodor (1983) between the notion of horizontal modules, which affect all of an agent's intelligence, and vertical modules, which are specialised skill areas showing disassociated abilities or deficits. Many topics have been touched on only lightly here, hopefully the bibliography can fill in more details for anyone interested.

Modularity is a key aspect of mind — it explains or at least describes our inability to access all of our intelligence, which in turn justifies the hacks we use to control our own behaviour — for example, not having a whole box of cookies in the house, or not

having even one drink if we know it would lead to more drinks than it's safe to have before driving. Modularity also casts an interesting light on the fundamental mental problem of planning or action selection. If our minds are modular, then choosing the next action is not necessarily something done by visiting all possible alternatives, but may instead be a matter of arbitrating between a number of alternative courses of action proposed by modules. These modules are then themselves engaged in a similar problem, but over a more limited set of goals and possible courses of action.

In conclusion, I expect modular models of intelligence will continue to dominate both AI and the natural intelligences, though not necessarily such clean and simplistic models as we began with. One of the current fundamental problems of AI is to enable the automatic learning of modular representations. This is difficult because it requires a general-purpose representational substrate which will almost certainly be slow and inefficient. It should also be noticed that the brain is *not* a general-purpose representational substrate, but incorporates an enormous amount of genetic bias which enables learning in animals. Similarly, as systems neuroscience comes to dominate the biological attempts to understand intelligence, there will be increased demands for models which somehow hide or abstract from the complexity of real, messy modules yet allow for the interconnectivity that makes the whole thing work.

Acknowledgements

References

Albus, J. S. (1997). The NIST real-time control system (RCS): an approach to intelligent systems research. *Journal of Experimental and Theoretical Artificial Intelligence*, 9(2/3):147–156.

Anderson, J. R. (1993). *Rules of the Mind*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Arkin, R. C. (1998). *Behavior-Based Robotics*. MIT Press, Cambridge, MA.

Asaad, W. F., Rainer, G., and Miller, E. K. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, 84:451–459.

Atkeson, C. G., Moore, A. W., and Schaal, S. (1997). Locally weighted learning for control. *Artificial Intelligence Review*, 11:75–113.

Bates, E. (1999). Plasticity, localization and language development. In Broman, S. and Fletcher, J. M., editors, *The changing nervous system: Neurobehavioral consequences of early brain disorders*, pages 214–253. Oxford University Press.

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., and Damasio, A. R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science*, 269(5227):1115–1118.

Blumberg, B. M. (1996). *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, MIT. Media Laboratory, Learning and Common Sense Section.

Bonasso, R. P., Firby, R. J., Gat, E., Kortenkamp, D., Miller, D. P., and Slack, M. G. (1997). Experiences with an architecture for intelligent, reactive agents. *Journal of Experimental and Theoretical Artificial Intelligence*, 9(2/3):237–256.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2:14–23.

Brooks, R. A. (1991a). Intelligence without reason. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, pages 569–595, Sydney.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Brosnan, S. F. and de Waal, F. B. M. (2003). Monkeys reject unequal pay. *Nature*, 425:297–299.

Bryson, J. J. (2000). Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(2):165–190.

Bryson, J. J. (2001). *Intelligence by Design: Principles of Modularity and Coordination for Engineering Complex Adaptive Agents*. PhD thesis, MIT, Department of EECS, Cambridge, MA. AI Technical Report 2001-003.

Bryson, J. J. (2002). Language isn't quite *that* special. *Brain and Behavioral Sciences*, 25(6):679–680. commentary on Carruthers,"The Cognitive Functions of Language", same volume.

Bryson, J. J. (2003). Where should complexity go? Cooperation in complex agents with minimal communication. In Truszkowski, W., Rouff, C., and Hinchey, M., editors, *Innovative Concepts for Agent-Based Systems*, pages 298–313. Springer.

Bryson, J. J. and Hauser, M. D. (2002). What monkeys see and don't do: Agent models of safe learning in primates. In Barley, M. and Guesgen, H. W., editors, *AAAI Spring Symposium on Safe Learning Agents*.

Bryson, J. J. and Stein, L. A. (2001a). Modularity and design in reactive intelligence. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 1115–1120, Seattle. Morgan Kaufmann.

Bryson, J. J. and Stein, L. A. (2001b). Modularity and specialized learning: Mapping between agent architectures and brain organization. In Wermter, S., Austin, J., and Willshaw, D., editors, *Emergent Neural Computational Architectures Based on Neuroscience.*, pages 98–113. Springer.

Carlson, N. R. (2000). *Physiology of Behavior*. Allyn and Bacon, Boston.

Carruthers, P. (2003). The cognitive functions of language. *Brain and Behavioral Sciences*, 25(6).

Chomsky, N. (1980). Rules and representations. *Brain and Behavioral Sciences*, 3:1–61.

Coad, P., North, D., and Mayfield, M. (1997). *Object Models: Strategies, Patterns and Applications*. Prentice Hall, 2nd edition.

Coltheart, M. (1999). Modularity and cognition. *Trends in Cognitive Sciences*, 3(3):115–120.

Connell, J. H. (1990). *Minimalist Mobile Robotics: A Colony-style Architecture for a Mobile Robot*. Academic Press, Cambridge, MA. also MIT TR-1151.

Cosmides, L. and Tooby, J. (1994). Origins of domain specificity: the evolution of functional organization. In Hirschfeld, L. A. and Gelman, S. A., editors, *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press.

Dawkins, R. (1976). Hierarchical organisation: A candidate principle for ethology. In Bateson, P. P. G. and Hinde, R. A., editors, *Growing Points in Ethology*, pages 7–54. Cambridge University Press, Cambridge.

Dehaene, S., Kerszberg, M., and Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Acadamy of Science, USA*, 95:14529–34.

Dehaene, S., Naccache, L., Cohen, L., Le Bihan, D., Mangin, J.-F., Poline, J.-B., and Rivire, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, 4(7):678–680.

Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition*, 79:221–237.

d'Inverno, M., Kinny, D., Luck, M., and Wooldridge, M. (1997). A formal specification of dMARS. In Singh, M. P., Rao, A. S., and Wooldridge, M. J., editors, *Proceedings of the 4th International Workshop on Agent Theories, Architectures and Languages*, pages 155–176, Providence, RI. Springer.

Donnai, D. and Karmiloff-Smith, A. (2000). Williams syndrome: From genotype through to the cognitive phenotype. *American Journal of Medical Genetics*, 97(2):164–171.

Downing, P. E., Liu, J., and Kanwisher, N. (2001). Testing cognitive models of visual attention with fmri and meg. *Neuropsychologia*, 39:1329–1342.

Doyle, J. (1983). A society of mind. Technical Report 127, CMU Department of Computer Science.

Edmonds, B. (2002). A review of "reasoning about rational agents" by Michael Wooldridge. *Journal of Artificial Societies and Social Simulation*, 5(1). http://jasss.soc.surrey.ac.uk/5/1/reviews/edmonds.html.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness. A Connectionist Perspective on Development*. MIT Press, Cambridge, MA.

Evans, D. and Zarate, O. (1999). *Introducing Evolutionary Psychology*. Icon Books Ltd., Cambridge.

Flombaum, J. I., Santos, L. R., and Hauser, M. D. (2002). Neuroecology and psychological modularity. *Trends in Cognitive Sciences*, 6(3):106–108.

Fodor, J. A. (1983). *The Modularity of Mind*. Bradford Books. MIT Press, Cambridge, MA.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–316.

Gall, F. J. (1825). *Sur l'origine des qualits morales et des facults intellectuelles de l'homme: et sur les conditions de leur manifestation*. J. B. Baillire, Paris.

Gardner, M. (1970). Mathematical Games: The fantastic combinations of John Conway's new solitaire game 'Life'. *Scientific American*, 223(4):120–123.

Gat, E. (1991). *Reliable Goal-Directed Reactive Control of Autonomous Mobile Robots*. PhD thesis, Virginia Polytechnic Institute and State University.

Gat, E. (1998). Three-layer architectures. In Kortenkamp, D., Bonasso, R. P., and Murphy, R., editors, *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems*, pages 195–210. MIT Press, Cambridge, MA.

Greene, A. J., Spellman, B. A., Dusek, J. A., Eichenbaum, H. B., and Levy, W. B. (2001). Relational learning with and without awareness: transitive inference using nonverbal stimuli in humans. *Memory & Cognition*, 29(6):893–902.

Guzzoni, D., Cheyer, A., Julia, L., and Konolige, K. (1997). Many robots make short work. *AI Magazine*, 18(1):55–64.

Hauser, M. D. (1999). Perseveration, inhibition and the prefrontal cortex: A new look. *Current Opinion in Neurobiology*, 9:214–222.

Hayes-Roth, B. (1985). A blackboard architecture for control. *Artificial Intelligence*, 26(3):251–321.

Hexmoor, H., Horswill, I., and Kortenkamp, D. (1997). Special issue: Software architectures for hardware agents. *Journal of Experimental and Theoretical Artificial Intelligence*, 9(2/3).

Hubel, D. H. (1988). *Eye, Brain and Vision*. Freeman.

Hume, D. (1748). *Philisophical Essays Concerning Human Understanding*. Andrew Millar, London.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Change*. MIT Press, Cambridge, MA.

Kauffman, T., Theoret, H., and Pascual-Leone, A. (2002). Braille character discrimination in blindfolded human subjects. *Neuroreport*, 13(5):571–574.

Kobayashi, T., Nishijo, H., Fukuda, M., Bures, J., and Ono, T. (1997). Task-dependent representations in rat hippocampal place neurons. *JOURNAL OF NEUROPHYSIOLOGY*, 78(2):597–613.

Konolige, K. and Myers, K. (1998). The Saphira architecture for autonomous mobile robots. In Kortenkamp, D., Bonasso, R. P., and Murphy, R., editors, *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems*, chapter 9, pages 211–242. MIT Press, Cambridge, MA.

Kortenkamp, D., Bonasso, R. P., and Murphy, R., editors (1998). *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems*. MIT Press, Cambridge, MA.

Livesey, P. J. (1986). *Learning and Emotion: A Biological Synthesis*, volume 1 of *Evolutionary Processes*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Lonstein, J. S. and Stern, J. M. (1997). Role of the midbrain periaqueductal gray in maternal nurturance and aggression: *c-fos* and electrolytic lesion studies in lactating rats. *Journal of Neuroscience*, 17(9):3364–78.

Malcolm, C., Smithers, T., and Hallam, J. (1989). An emerging paradigm in robot architecture. In *Proceedings of the International Conference on Intelligent Autonomous Systems (IAS)*, volume 2, pages 545–564, Amsterdam. Elsevier.

Matarić, M. J. (1990). A distributed model for mobile robot environment-learning and navigation. Technical Report 1228, Massachusetts Institute of TechnologyArtificial Intelligence Lab, Cambridge, Massachusetts.

Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Progress In Neurobiology*, 50(4):381–425.

Minsky, M. (1985). *The Society of Mind*. Simon and Schuster Inc., New York, NY.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge, Massachusetts.

Parnas, D. L., Clements, P. C., and Weiss, D. M. (1985). The modular structure of complex systems. *IEEE Transactions on Software Engineering*, SE-11(3):259–266.

Perrett, D. I., Hietanen, J. K., Oram, M. W., and Benson, P. J. (1992). Organisation and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London*, 335:25–30.

Redgrave, P., Prescott, T. J., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023.

Sen, K., Theunissen, F. E., and Doupe, A. J. (2001). Feature analysis of natural sounds in the songbird auditory forebrain. *Journal of Neurophysiology*, 86(3):1445–1458.

Sengers, P. (1999). *Anti-Boxology: Agent Design in Cultural Context*. PhD thesis, School of Computer Science, Carnegie Mellon University.

Siemann, M. and Delius, J. D. (1993). Implicit deductive reasoning in humans. *Naturwissenshaften*, 80:364–366.

Skaggs, W. and McNaughton, B. (1998). Spatial firing properties of hippocampal CA1 populations in an environment containing two visually identical regions. *Journal of Neuroscience*, 18(20):8455–8466.

Sloman, A. and Logan, B. (1999). Building cognitively rich agents using the Sim_agent toolkit. *Communications of the Association of Computing Machinery*, 42(3):71–77.

Spelke, E. S. (2003). What makes us smart? Core knowledge and natural language. In Gentner, D. and Goldin-Meadow, S., editors, *Advances in the Investigation of Language and Thought*. MIT Press, Cambridge, MA.

Tanaka, J. W. and Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, 46A(2):225–245.

Thelen, E. and Smith, L. B. (1994). *A Dynamical Systems Approach to Development of Cognition and Action*. MIT Press, Cambridge, MA.

Thórisson, K. R. (1999). A mind model for multimodal communicative creatures & humanoids. *International Journal of Applied Artificial Intelligence*, 13(4/5):519–538.

Weiß, G., editor (1999). *Multiagent Systems*. The MIT Press, Cambridge, Massachusetts.

Winston, P. (1975). Learning structural descriptions from examples. In Winston, P., editor, *The Psychology of Computer Vision*. McGraw-Hill Book Company, New York.

Wooldridge, M. J. and Ciancarini, P. (2001). Agent-Oriented Software Engineering: The State of the Art. In Ciancarini, P. and Wooldridge, M. J., editors, *First International Workshop on Agent-Oriented Software Engineering*, volume 1957 of *LNCS*, pages 1–28. Springer, Berlin.