Representing Cognitive Phenomena in Biological Systems

Joanna J. Bryson
Harvard Primate Cognitive Neuroscience Laboratory
Cambridge, MA USA

1 Introduction

This is a short paper relating representations of intelligence between three fields: psychology, neuroscience and artificial intelligence (AI). I particularly emphasize the role of modularity in these three areas. Because space is limited, I will assume a general familiarity with modular AI architectures, and concentrate on relating them to natural intelligence[1].

2 Modularity in Psychology

I will begin with an incredibly 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 criteria for modularity: that some part of the mind does not have access to some other part of the mind. Given this simple criteria, anyone who accepts the idea of implicit knowledge or unconscious action has already acknowledged that there is some sort of modularity involved in human intelligence.

The second reason this quote is useful 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 Modularity of Mind* by 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 research has followed this lead (recently Coltheart, 1999; Downing et al., 2001; Spelke, in press).

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 (e.g. Minsky, 1985; Brooks, 1991; Albus, 1997). This is because, for Fodor, the purpose of modules is to translate the complexity of raw sensory input into a common representation used by a general-purpose reasoning system which chooses the course of action. Presumably, Fodorian motor modules would translate similarly generic instructions into the complexity of muscular control.

Fodor believes that the mind is constructed of both *vertical* capacities, the afore-mentioned modules specialized to task, and *horizontal* capacities, things like the general-purpose reasoning system. At this high level, his theory is consistent with some AI work. For example, although Minsky (1985) describes single modules (or 'agents') capable of sensing, planning and action, he also describes memory systems and organizational structures (e.g. the B-brain) which are accessible to or have access to all agents. PRS and three-layered architectures (e.g. Georgeff and Lansky, 1987; Bonasso et al., 1997) also have both perception/action modules and monolithic elements such as planners.

3 Modularity in Brains

I would now like to turn from psychology to neuroscience. We have evidence of at least three sorts of modular decomposition in mammal brains[2].

3.1 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 organizations of connectivity. Even if we did not have behavioral evidence (as we do) that the neocortex, cerebellum, hippocampus and so forth perform different functions, we would suspect as materialists and computer scientists that these organs must perform different computations, because of their different structure. This point becomes more obvious when we realize there

[1] See Bryson (2000) for a review of the AI architectures mentioned here.

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

is no particular reason not to extend it to more peripheral organs, such as the spinal chord, the retina or the cochlea.

3.2 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 (Kauffman et al., 2002, e.g.), it may be that such apparent specialization also reflects connectivity, this time toward subcortical brain organs specialized for purposes such as social interaction and navigation (the amygdalic and hippocampal systems respectively.) Some cortical regions are steps along a stream of processing, e.g. 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).

3.3 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 recognized. 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 (Skaggs and McNaughton, 1998, e.g.) might account for individual differences in insight and generalization based on the relative accessibility of two representations.

3.4 Discussion

I would argue that modularity by region could be considered analogous to Fodor's vertical capacities, the things *he* calls 'modules.' They also correspond to AI *behaviors*, as proposed by Brooks (1991), and used widely in modular AI. However, it may take a stream of several cortical areas (for example) to correspond to one Fodorian module, and an even longer stream of processing to create a full Brooksian behavior connecting perception to action.

Modularity by organ is a more analogous to horizontal capacity — organs are often specialized to task rather than perceptual domain and help the agent as a whole. On the other hand, there are many more organs, and their functioning more intertwined with the regular modules, than I think either Fodor or three-layered agent architectures imply. For example, a great deal of semantic knowledge seems to be stored in specialized cortical regions rather than being associated tightly with a planner as is the model of PRS. Minsky's 'Society of Mind' or Soar (Newell, 1990) might be closer models of this knowledge distribution, but neither of these systems have as many sorts of specialized processing as the brain has dedicated organs.

Temporal modularity — modularity by context — is not generally shown in modular AI; it has more in common with traditional computing systems. Modular AI systems tend to have all modules operating continuously in parallel. However, Soar has always had the notion of problem space to constrain search to a particular context (Laird and Rosenbloom, 1996). Developers of a related but simpler system, ACT-R, tried to do away with problem spaces in order to simplify the system, but found them necessary for successful problem-solving (Anderson and Matessa, 1998).

4 Module Coordination and Structured Action Selection

My own AI research has been into the management and design of modular AI. I have come to the conclusion that 1) Semantic and task memory should be stored in specialized representations within behaviors (perception/action vertical modules), and 2) ordering the behavior of such modules is best done using a specialized, horizontal module for sequencing behavior. 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).

I believe that this behavior sequencing is directed by a number of specialized organs in mammals. For example, the affective forebrain systems including the amygdala help redirect attention out of a complex plan sequence in response to urgent environmental stimuli such as loud noises. The amygdala can also learn to respond to frequently salient stimuli such as particular sounds, people or rooms. The basal ganglia has recently been implicated in arbitrating between competing subsystems (Mink, 1996; Redgrave et al., 1999). The periaqueductal gray has been implicated in action sequencing for complex, species-typical tasks (Carlson, 2000; Lonstein and Stern, 1997).

Other horizontal / organ-based biological modules that I believe would have useful analogs for AI systems include the cerebellum, which provides dynamic smoothing between discrete position targets, and the hippocampus, which seems to provide for both episodic memory and task learning (see further discussion in Bryson and Stein, 2001b).

5 Deliberation

Deliberation, or conscious attention to a task, still seems deeply mysterious to me. Although I have been studying planning and modularity with an eye to biological plausibility for over a decade, and although the accessibility difference that determines explicit from implicit knowledge is a key indicator of modularity[3], I still 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).

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, although language may fundamentally *alter* 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, in press). 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.

Most intriguing to me are a number of recent results 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 (Siemann and Delius, 1993; Bechara et al., 1995; Greene et al., 2001). I suspect two things. First, that Dennett (2001) is absolutely right in thinking that, as we come to understand consciousness, we'll realize we have been covering several disparate functions with that one term, none of which are magic, and second, that two of these functions will be focusing search and ordering behavior in time.

References

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

Anderson, J. R. and Matessa, M. (1998). The rational analysis of categorization and the ACT-R architecture. In Oaksford, M. and Chater, N., editors, *Rational Models of Cognition*. Oxford University Press.

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.

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.

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 & Theoretical Artificial Intelligence*, 9(2/3):237–256.

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

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. 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.

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

[3] See the earlier Flombaum et al. quotation.

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.

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

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.

Georgeff, M. P. and Lansky, A. L. (1987). Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, pages 677–682, Seattle, WA.

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.

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

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.

Laird, J. E. and Rosenbloom, P. S. (1996). The evolution of the Soar cognitive architecture. In Steier, D. and Mitchell, T., editors, *Mind Matters*. Erlbaum.

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.

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.

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.

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.

Spelke, E. S. (in press). What makes us smart? Core knowledge and natural language. In Gentner, D. and Goldin-Meadow, S., editors, *Whither Whorf?* MIT Press, Cambridge, MA.