CogSci to AI: It's the Brainware, Stupid!

Jacob Beal and **Gerald Jay Sussman**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge MA 02139

Abstract

Current modularization techniques fail when applied to hard AI problems. But cognitive science shows that the mind has modules specialized for particular functions. Unlike current engineered modules, the modules of the mind learn to communicate with each other as a child matures. Kirby's ideas on language evolution, combined with constraints derived from neuroanatomy, yield a new mechanism for integrating modules into a system: a communications bootstrapping system in which two agents build a shared vocabulary capturing information common to their mutual experience, including cross-module knowledge about the world.

The Problem of Modularity

Modularity and abstraction are among the crown jewels of computer science. They allow engineers to take fantastically complex problems and decompose them into myriad subproblems which a well-trained human can comprehend well enough to solve. Without them, we cannot build complex artifacts like modern word processors: it would be surprising if building a mind was any less difficult.

Our current techniques for modularity have failed in AI. Although the field has produced many elegant and powerful systems, these systems are brittle and inflexible. They are difficult to build upon or apply to problems of a class other than those they were designed to solve.

Should we abandon modularity? Cognitive science's study of naturally occurring minds suggests otherwise: recent evidence indicates that the human brain likely contains specialized modules,(Kanwisher 1998; Spelke 2003) and that infants have the same core cognitive systems as other mammals, with our unique mental capabilities arising from cooperation between modules.(Spelke 2003)

The critical issue in building a mind is not the modules from which is it composed, but the process by which they integrate. Whereas traditional computer science modules are "born" connected, the modules of the human brain appear to learn to communicate over time.

For example, human adults can re-orient themselves to find a location specified as a combination of color and geometry, while rats and infants only use geometry to re-orient.(Hermer & Spelke 1996) Moreover, the transition between infant and adult capabilities correlates with production of the words "left" and "right"(Hermer-Vasquez, Moffett, & Munkholm 2001) and adult re-orientation is impaired by simultaneous performance of a language task but not a rhythm task.(Hermer-Vasquez, Spelke, & Katznelson 1999)

We envision a mind composed of a few dozen specialized modules, each bidirectionally connected to a few other related modules. For example, we might have an audition module connected to a language module, which connects to trajectory, which connects to vision.

The modules then integrate by learning to communicate about shared experiences. Differences in perspective between modules will result in vocabulary entries that capture cross-module relations in the external world, allowing the collection of modules to act as an integrated whole rather than merely a union of individual parts.

Kirby's Language Learners

This approach is inspired by Simon Kirby's work on language evolution.(Kirby 1998; 2002)[1] Kirby has been investigating the origins of human language, with the idea that universal grammar might not be encoded explicity, but rather emerge from the interaction of a community of learners with a simple induction algorithm.

In Kirby's experiments, a student and teacher are given the same semantic expression. The teacher generates an utterance, occasionally inventing random babble to fill gaps. The student assumes the teacher is making sense and tries to organize the teacher's utterances into a grammar. Every once in a while, a coincidence in the random utterances lets the student make a generalization. Later, the student becomes the teacher, and the generalization is no longer coincidence. The coincidences build up until they collide and transition to a stable grammar with words in subject, verb, and object roles and a vocabulary entry for each semantic symbol.

Consider this not as a model of language evolution, but as a model for the development of cooperation among modules of the mind.

Design Hardware, Not Software

Software is too flexible. Hardware constraints force us to build realizable systems and give us hints about the engineering principles that nature exploits to construct minds.

[1] See also work by Batali(2002) and Steels(1996)

Bug-for-bug compatibility between humans and AI is unimportant: the Wright brothers learned wing-warping from birds, but their airplane had no feathers.

Hardware designers build physical mechanisms out of fixed numbers of well-characterized devices at a relatively high level of abstraction—a hardware designer can think in terms of register banks and ALUs, not the gates of which they are composed. Devices are not created and destroyed dynamically, although they may go unused. As hardware designers, we build minds from devices subject to constraints inferred from the biology of the brain.

For starters, humans have only about one gigabyte of DNA. Much of that is devoted to cell metabolism and other organs besides the brain, yet a specification of the detailed interconnections in a mature brain would surely take more than a gigabyte. Further, although the overall form of a biological system is quite robust, much variation occurs during morphogenesis and still yields a functioning system, so a design should not depend on precision interconnection of parts. Encoded information must be robust against noise and damage and must be adapted throughout the life of the system. Thus the details of communication among modules are not predetermined and must be learned dynamically.

Evolution generally develops new structures by duplicating an existing structure, then modifying the duplicated copy. It is cheap to make many copies of a structure: millionfold parallelism is reasonable. Sequential processing, on the other hand, is extremely expensive, and no more than a few dozen serial steps should lie between sensation and reaction.

Finally, the anatomy of the brain shows that most connections between areas are bidirectional, and that motor areas are active during the interpretation of sensory data.(Ullman 1996) Thus, mechanisms should be bidirectional, so that the same piece of knowledge can be used for both interpretation and production. Constraints like those used by Yip and Sussman to learn phonological rules are a good example of such a system.(Yip & Sussman 1998)

Communications Boostrapping Captures World Knowledge in Linguistic Agreements

We apply these hardware constraints to Kirby's basic ideas to yield communication bootstrapping(Beal 2002), in which two agents learn a dictionary of encodings capturing commonalities in their experience.

In the communication bootstrapping system, two agents are connected to each other and the external world by thick, twisted bundles of wires. Scenes in the world are thematic role-frames, delivered to each agent as a signal for role on wires corresponding to the feature playing that role. An agent invents encodings for features by choosing a random sparse set of wires on the bus connecting it to its partner, and likewise invents encodings for roles.

Agents agree on an encoding using modified Hebbian learning. Successful learning depends upon three things: sparseness in the appearance of features, sparseness in the encoding of symbols, and strong correlations between the features of the two agents.

Given these things, the two agents rapidly agree upon symbols to express equivalent sets of features to one another: the expected time for the two agents to acquire a shared symbol is 10 exposures to scenes involving the features it expresses. Communication bootstrapping also degrades gracefully in the face of damage or parameter variance.

With multiple modules, each specialized to a particular aspect of cognition and connected together via channels supported with bootstrapping protocols, the shared symbols may then become a medium for expressing more general knowledge about the world and solving more complicated problems than can be solved by any particular module.(Beal 2004)

References

Batali, J. 2002. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In Briscoe, T., ed., *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press. chapter 5.

Beal, J. 2002. Generating communications systems through shared context. Master's thesis, MIT.

Beal, J. 2004. Leveraging language into learning. PhD Proposal.

Hermer, L., and Spelke, E. 1996. Modularity and development: the case of spatial reorientation. *Cognition* 61:195–232.

Hermer-Vasquez, L.; Moffett, A.; and Munkholm, P. 2001. Language, space and the development of cognitive flexibility in humans. *Cognition* 79:263–299.

Hermer-Vasquez, L.; Spelke, E.; and Katznelson, A. 1999. Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology* 39:3–36.

Kanwisher, N. 1998. The modular structure of human visual recognition: Evidence from functional imaging. In Sabourin, M.; Craik, F.; and Robert, M., eds., *Advances in psychological science, Vol. 2: Biological and cognitive aspects*. Psychology Press. 199–214.

Kirby, S. 1998. Language evolution without natural selection: From vocabulary to syntax in a population of learners. Technical report, Language Evolution and Computation Research Unit, University of Edinburgh.

Kirby, S. 2002. Learning, bottlenecks and the evolution of recursive syntax. In Briscoe, T., ed., *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press. chapter 6.

Spelke, E. 2003. What makes humans smart? In Gentner, D., and Goldin-Meadow, S., eds., *Advances in the Investigation of Language and Thought*. MIT Press.

Steels, L. 1996. Emergent adaptive lexicons. In Maes, P., ed., *SAB96*. Cambridge, MA: MIT Press.

Ullman, S. 1996. *High-Level Vision*. MIT Press. chapter Sequence Seeking and Counter-Streams.

Yip, K., and Sussman, G. 1998. Sparse representations for fast, one-shot learning. Technical Report MIT AI Lab Memo 1633, MIT.