Learning Composable Signals for a Cognitive Substrate

Jacob Beal (jakebeal@mit.edu)
MIT CSAIL, 77 Massachusetts Ave, Cambridge, MA 02139 USA

Abstract

According to the cognitive substrate hypothesis, human intelligence arises from the integration of specialist parts that are mostly shared with other mammals. In order for such specialists to cooperate in novel situations, however, they must agree on a system of signals that can describe aspects of the new situation using relations between familiar elements. This poses a problem because many signals can only be learned from experience, yet different specialists may experience the world in very different ways. This paper demonstrates that distributed learning of such signals is possible, and also that the apparent problem is actually a benefit, using a vision specialist and a hearing specialist that together observe a simulated four-way intersection. Using a heuristic method based on Allen's time relations, these two specialists agree on a set of composable signals, and some dynamics of the simulation are captured in the differences in how the two specialists interpret signals.

Introduction

The cognitive substrate hypothesis asserts that our unique human intelligence arises from the integration of specialist parts that are mostly shared with other mammals. If we are to accept this hypothesis, then we must explain how such a team of specialists might cope with the novelty that arises in everyday activities. Even an apparently simple activity, like crossing the street, often involves many different cognitive faculties and a particular arrangement of elements that has never before encountered. The scene in Figure 1, for example, requires vision and hearing sensory data, social reasoning to decide whether the bus will yield, spatial reasoning to know which crosswalk leads toward an unseen destination, language to interpret the construction signs, and so forth. The situation is complex enough that, although all of the elements are familiar, this particular arrangement has likely never before been encountered.

Humans show remarkable competence and flexibility in coping with such situations. If a cognitive substrate model is to do the same, then the specialists must share a system of signals that can describe aspects of a novel situation using relations between familiar elements. Since stoplights and buses are not built into our DNA, many of the signals likely need to be invented and agreed upon by the specialists as they learn about the world.

Agreeing on signals may be challenging because different specialists experience the world in qualitatively different ways: the times when a bus is seen and the times when a bus is heard are related, but not at all the same. Distributed agreement on a signal is an unsupervised learning problem with no clear distributional assumptions to lean on. Moreover, many different potentially distracting elements are present at once and change on time scales spanning several orders of magnitude, from seconds (honking) to minutes (lights) to hours (traffic jams) or even weeks (construction).

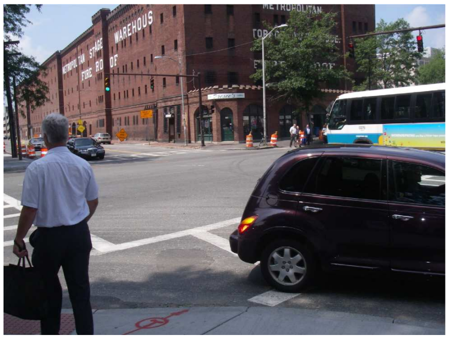


Figure 1: Everyday activities, like crossing the street, often involve many different cognitive faculties and particular arrangements of elements that have never before been seen.

Surprisingly, not only is distributed learning of such composable signals possible, but the differences between specialist's observations is actually a benefit. This is demonstrated using a vision specialist and a hearing specialist that together observe a simulated four-way intersection. Using a heuristic method based on Allen's time relations(Allen, 1983), the two specialists agree on a set of composable signals, and dynamics of the simulation (e.g. hearing an engine predicts a car will soon be seen) are captured in the differences in how the two specialists interpret signals.

Related Work

The cognitive substrate hypothesis is based on recent work in cognitive science. Infant studies show that humans are born with essentially the same cognitive faculties as other mammals—language emerges later (Spelke, 2003). As we mature, these faculties are integrated to produce uniquely human capabilities. For example, children develop the concept of number by combining three faculties—analog magnitude, parallel individuation, and sequence memorization—in a standard developmental sequence (Carey, 2004). In another example, human adults can reorient themselves to find a location specified as a combination of two types of feature, color and geometry, while children less than five years old and rats only use geometry, a single feature, to reorient (Hermer & Spelke, 1996). An explicit statement of the hypothesis, along with one proposal for a set of specialists, can be found in (Cassimatis, 2006).

Much research on cognitive architectures is compatible with the cognitive substrate hypothesis. Architectures such as SOAR (Wang & Laird, 2006), ACT-R (Anderson et al., 2004), ICARUS (Langley & Choi, 2006), and EPIC (Kieras & Meyer, 1997), to name only a few, are constructed from a set of specialists that cooperate to carry out cognitive tasks.

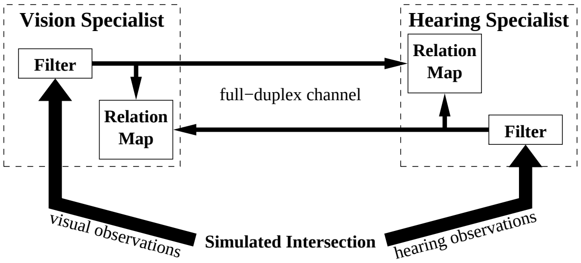


Figure 2: Experimental design: a simulated four-way intersection generates two streams of observations, one for vision and one for hearing. Each stream is filtered to produce a stream of messages. A relation map in each specialist compares the two message streams to discover predictive relations between its observations and message elements used by the other specialist.

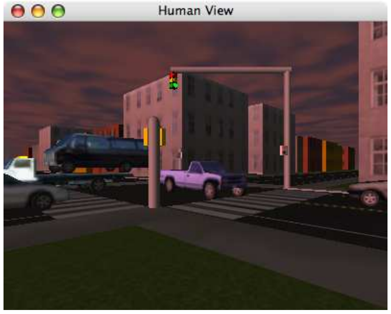


Figure 3: Screenshot of the four-way intersection simulation

PolyScheme (Cassimatis, 2002) is notable for its explicit use of the cognitive substrate hypothesis, and has been used to demonstrate that simple cognitive faculties can be integrated to perform computationally sophisticated reasoning (Cassimatis, Bugajska, Dugas, Murugesan, & Bello, 2007). These systems, however, investigate what integrated specialists can accomplish, rather than how they might learn to communicate or to integrate.

Integration of different modes of input has also been a major subject of study. Kohonen maps (Kohonen, 1989), for example, can be used to organize multiple streams of input into a low-dimensional similarity map. Coen's slices (Coen, 2006) and Roy's cross-modal approach (Roy, 1999) both extract symbols from similarities across a pair of input streams. These approaches focus on identifying and segmenting input, however, and do not produce representations that can be directly used for communicating relations between modes. A notable exception is Minsky's "Emotion Machine" proposal (Minsky, 2006), which is partially implemented in EM-ONE (Singh, 2005) using stories as a basis for integration.

The problem of agreeing on signals has been studied for homogeneous agents by the synthetic languages community (Kirby, 2002b), generally on systems with small vocabularies and slow convergence rates. Particularly notable is Kirby's model of language invention through iterated learning and accumulation of coincidence (Kirby, 1998, 2002a). This is closely related to work by Steels on grounded language acquisition (Steels, 1996), by Yanco on self-configuring communication for mobile robots (Yanco, 1994), and Batali on learning grammar in recurrent neural networks (Batali, 2002). Finally, I have applied these ideas to the cognitive substrate domain for specialists with very similar inputs (Beal, 2002b, 2002a).

Experimental Design

A series of two-specialist experiments provides evidence that distributed agreement on signals is possible despite lack of supervision, differences in observations, and varying time scales. Furthermore, some dynamics of the observed environment are captured as differences in how the two specialists interpret a signal.

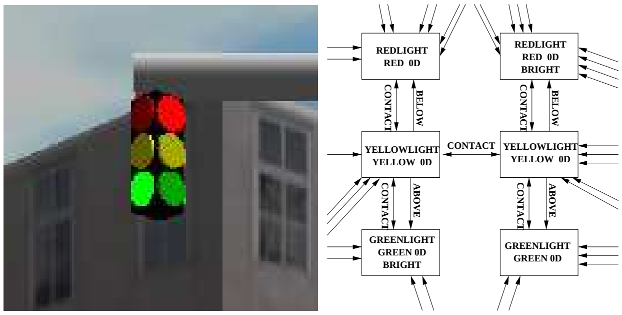
These experiments use a simple system consisting of a simulated four-way intersection observed by a vision specialist and a hearing specialist (Figure 2). Within each specialist, a filter turns each observation into a message describing part of the observation. The specialists exchange messages, which are sets of token pairs, on a full-duplex channel. Finally, a relation map uses heuristic methods to detect predictive relations pairwise between elements in the incoming and outgoing streams of messages.

We stress that the details of the simulation or the specialists should not be regarded as particularly important to the design of this experiment. The simulation is simply a domain that is both readily familiar and exhibits complex behavior at multiple time-scales in two senses. As for the designs of the specialists: it is quite likely that more justifiable and better performing designs could be developed and applied to this question. As yet, however, they have not, and the relatively *ad hoc* designs used here are the result of exploration to determine that the problem is solvable.

Simulated Environment

The experiments use data generated from a simulated four-way intersection. We attempt to obtain realistic enough dynamics and sensory input by building the simulation with well-established tools: dynamics are provided by the Open Dynamics Engine, a well-established open source physics engine, and rendering is done with OpenGL and OpenAL, a standard 3D graphics library and its companion 3D sound library.

The simulator produces scenes containing dozens of objects: streets, sidewalks, buildings, the stoplight and associated poles and walk signals, cars, and people. These objects all interact in many different ways at different time scales and frequencies. For example, the stoplight follows an ordinary cycle, with a walk light and audible walk signal that pedestrians push a button to request. Pedestrians have varying temperaments, clothing, builds and ages, come singly or in groups, often jaywalk, have conversations, meet friends and change their plans, and flee from oncoming cars. Cars



(a) (b)

Figure 4: A fragment of visual observation containing the six lights of the stoplight (shown in close-up in (a)). This portion contains 6 objects, 20 features, and 44 relations (b). A full observation contains many more.

have varying colors, models, driving skill and personality, make right turns on red, yield on left turns, run some yellow and occasional red lights, negotiate right of way, honk when frustrated, and usually drive defensively. The simulator also includes accidents, emergency vehicles, a daily cycle, and more. Screenshots from the simulation are shown in Figure 3.

Observations

Each specialist receives a stream of observations from the same fixed position and orientation, scraped from the simulator at some fixed sampling rate. Each observation is a list of things with properties and relations to one another. For example, a car may be reported to the vision specialist as a thing that looks like a car, occupies a small part of the visual field, is approximately red, is moving to the right, is above something that looks like a road, and so on.

A hearing observation is a list of sources with binary features describing their type, direction, and loudness. A visual observation is a graph of visible objects and their relations (above/below, left/right, forward/back, and contact), plus binary features for each object that describe its type, color, size, and motion. Figure 4 shows an example fragment of a visual observation. For both senses, segmentation and categorization is short-circuited using rendering information from the simulator.

Filter

Each specialist contains a filter that turns observations into messages, where a message is an unordered set of $(feature, marker)$ ordered pairs. Such messages can encode a relation between two objects by pairing each object's features with a marker for its role in the relation. The messages produced by the filter are then transmitted to the other specialist, remapping features and markers consistently to a set of arbitrary tokens in order to prevent any *a priori* structure from affecting interpretation.

Experiments use either a flat filter or a shared-focus filter. The flat filter takes an observation and turns it into a list of the elements it contains, each marked with the same placeholder marker. Thus the existence of relations is reported but their content is not. For example, a blue sky above a red car would be encoded as $\{(above,\bullet),(blue,\bullet),(car,\bullet),(red,\bullet),(sky,\bullet)\}$, losing the information about which object is above which.

The focus filter designates f simulator objects as foci of attention, shifting these foci with a heuristic mechanism that balances reflexive tracking and joint attention (described fully in (Beal, 2007)). The filter then sends messages containing the parts of an observation that directly connect to foci shared by both specialists. Features of a focus are marked $focus_n$ and a relation r between a focus and another object is encoded by marking features of the other object with r_n. For example, a red car beneath a blue sky, with the red car as focus number two, would be encoded as $\{(blue, above_2), (car, focus_2), (red, focus_2), (sky, above_2)\}$.

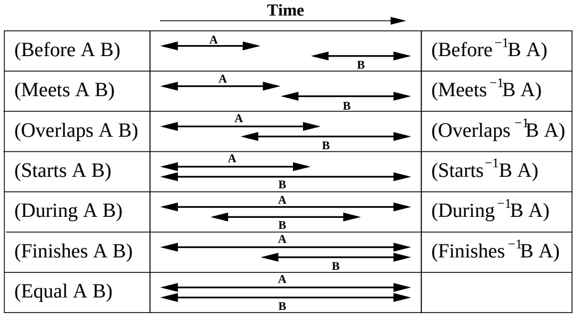


Figure 5: Allen's 13 time relations (Allen, 1983) compare intervals by comparing their start and end times.

Name	Time Relations												
	B	M	O	S	D	F	E	f	d	s	o	m	b
EQUAL	-		+	+	+	+	+	+	+	+	+		-
SUBCLASS				+	+	+		+		-			
SEQUENCE				-	+	+		-	-	-			
CAUSE		-	+	+		+		-		-	-		
ENABLE	+	+	+	+	+	+	-	-	+	-	-		
DISABLE			+	+		-		-	-	-			

Figure 6: A specialist's relation map heuristically interprets the 13 time relations (identified by their first letters, inverses in lower case) as positive, negative, or neutral evidence regarding 11 predictive relations: the 6 above and the inverses of all but symmetric **EQUAL**.

Relation Map

Each specialist's relation map compares incoming and outgoing message sequences to find a set of pairwise predictive relations between features in the specialist's observations and tokens in the signals it receives. We produce agreement between specialists trivially by having the relation maps treat the two sequences symmetrically.

The relation map uses heuristic methods (detailed fully in (Beal, 2007)) to search for 11 feature/token relations, considering every possible pairing and relation in parallel. The streams are interpreted in a time-scale invariant manner by using Allen's time relations(Allen, 1983) (shown in Figure 5) to compare intervals when a feature or token is present in consecutive messages. These time relations are then interpreted heuristically as evidence for or against predictive relations (Figure 6) and independence of evidence increased by only considering the first time relation following a **BEFORE** time relation, which indicates an interval where neither feature nor token is present. Evidence is accumulated using a simple incremental strength measure: starting at zero, positive evidence shifts strength up 1, negative shifts it down 2, rails at +50 and -50 prevent over-saturation, and the relation is considered true whenever the strength is at least 10.

Only **EQUAL** relations represent a direct translation between specialists. Any other predictive relation is a distributed representation of a cross-specialist relation between features: although each specialist acquires the relation, within a specialist it connects the feature to a token with no inherent semantics. When a message containing such tokens is sent between specialists, its interpretation is thus effectively a step of rule-based reasoning using these cross-specialist relations. For example, if the predictive relation **CAUSE** connects engine sounds to seeing a car, then a message from the hearing specialist saying "An engine is to the left of the intersection" is interpreted by the vision specialist as "A car will appear at the left of the intersection."

Experiments

Four sets of experimental data were collected. Collected predictive relations are listed in full in (Beal, 2007).

Content of Relations: To test whether meaningful relations are learned, two simulations were run, each starting at noon, running for 5,000 simulated seconds, and taking observations every 0.5 seconds. For the first run, messages were exchanged using the flat filter; for the second run, messages were exchanged using the focus filter. The number of predictive relations acquired by each specialist was recorded every 100 observations; after each run, each specialist's final set of predictive relations was recorded.

Variation in Sampling Rate: To test whether the results are dependent on sampling rate, ten simulations were run, each starting at noon and running for 5,000 simulated seconds. Each run was observed at four different rates—0.5, 1.0, 1.5, and 2.0 seconds—and messages exchanged using the flat filter. After each observation was processed, the set of relations was used to predict which features would appear or disappear in the next observation, and the cumulative number of unpredicted transitions recorded every 1,000 observations. Predictive relations were recorded as before, and for the first run at each rate, each specialist's final set of predictive relations was recorded.

Variation in Extrinsic Activity: To test whether the results are dependent on extrinsic activity, ten simulations were run for each of three different start times: midnight (1/10 car and pedestrian activity), 8am (double car activity), and 3pm (double pedestrian activity). Each of the 30 runs lasted 5,000 simulated seconds, taking observations once every 0.5 seconds and exchanging messages using the flat filter. Prediction and relations were recorded as for the sampling rate data-set.

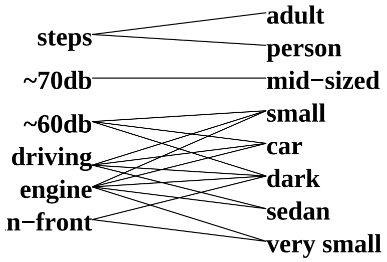


Figure 7: The focus filter produces three clusters of strongly associated features.

Convergence: To test that the set of relations does eventually converge, a simulation was run starting at noon and continuing for 20,000 simulated seconds, taking observations once every 0.5 seconds and exchanging messages using the flat filter. Prediction and relations were recorded as for the sampling rate data-set.

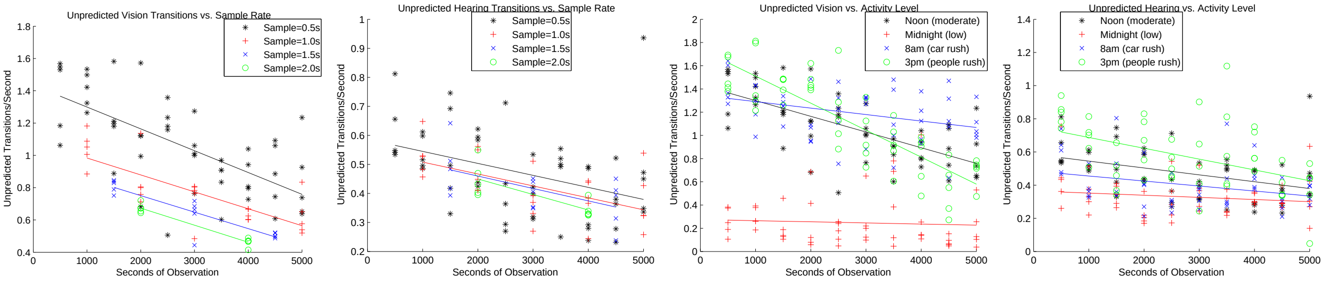
Results and Analysis

Analysis shows that the experimental data is consistent with the desired result: rapid agreement on a set of signals that can describe aspects of a new situation using relations between familiar elements. Furthermore, the differences in interpretation between the two specialists (relations other than **EQUAL**) capture some dynamics of the simulation.

Meaningful relations are acquired even when using the focus filter, which discards a large, shifting portion of the observations. The relations describe phenomena occurring at variable lengths over a wide variety of time scales. Finally, neither sampling rate nor activity level makes a significant difference in the speed of acquisition or the type or usefulness of relations acquired.

Content of Relations: In the flat filter content run, the specialists learn identical sets of 156 relations on 91 feature/token pairs, out of a possible 15576 relations on 1416 pairs. All six types of relation are represented among the 156 learned: there are 78 **ENABLE** relations, 47 **SUBCLASS** relations, 18 **EQUAL** relations, 9 **DISABLE** relations, 2 **CAUSE** and 2 **SEQUENCE** relations. Of these, only 6 are of dubious correctness, while the rest capture valid simulator structure and dynamics, including:

- The walk light is equivalent to the audible walk signal.
- The "don't walk" light is followed by the audible walk signal, then disappears.
- A moderately loud sound is always followed by the appearance of a car, and sometimes the car subclasses truck, van, or SUV. The sound often leads to motion away from the observer (the car crossing the intersection).



(a) Sampling Variation (Vision) (b) Sampling Variation (Hearing) (c) Activity Variation (Vision) (d) Activity Variation (Hearing)

Figure 8: Over time, the number of unpredicted transitions trends downwards, as shown by the linear regression on each data set. Neither sampling rate nor extrinsic activity has a significant effect on percentage rate of improvement.

- Cars are always moderately loud.
- When the walk light is visible, engine idling is heard.
- Engine idling is only heard when there are are cars visible.
- Sounds directly in front come from a car.

We can thus see that the relations learned involve many time spans, some short like the passage of a car through the intersection, some moderate like stoplight signals, and some fairly long like all the times when some car is audible. The relations also involve both common features like hearing a car idle, which happens in 72.5% of the samples, and rare ones like hearing the audible walk signal, which happens in only 5.5% of the samples. Moreover, these results are not qualitatively affected by changes in sampling rate or extrinsic activity: although the size of the final set varies (from a minimum of 97 to a maximum of 176) and the particulars of the relations captured vary as well, all runs capture some important simulator structure and dynamics and acquire only a handful of dubious relations.

In the focus filter content run, both specialists acquire an identical set of 448 relations. The greater number is largely due to breaking up the intervals when extremely common phenomena, like the sound of people walking, are present. The focus filter relations capture no cross-object dynamics, consistent with the much reduced coverage of the messages, but do capture a set of strongly associated features (connected by at least four relations) that form the three clusters shown in Figure 7. These clusters correspond roughly to cars, people, and things passing close to the observer, and offer a base that a more sophisticated filter might use to guide cross-object relation discovery.

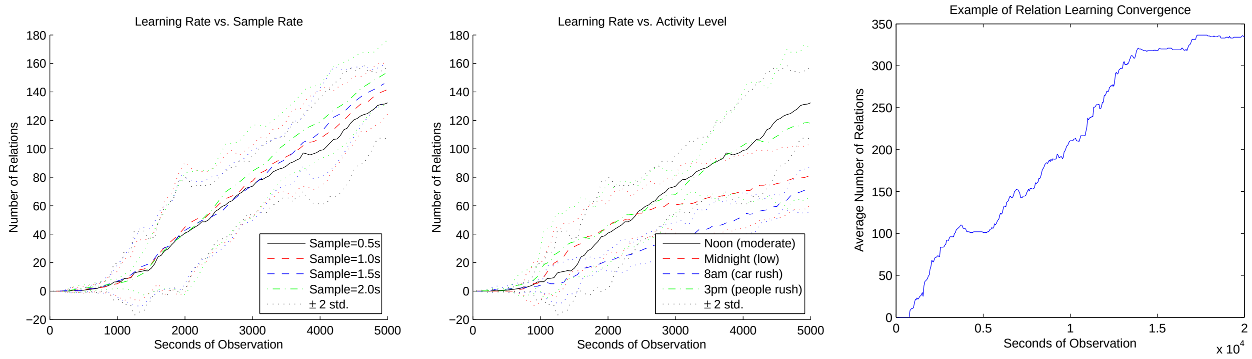
Prediction Quality: The ability of relations to capture simulator dynamics is also illustrated by their ability to predict changes in a specialist's observations. Figure 8 shows the rate of unpredicted transitions over time for each specialist under the various sampling and extrinsic activity conditions, plus a linear regression for each data set to measure the trend of improvement. Note that the measure is extremely noisy, due to the long-duration variations in simulator behavior. The number of unpredicted transitions trends downward in every case, and although the trend lines are different, the noise in the measure means that there is no significant difference observed in the percentage rate of improvement, either between specialists or between different conditions.

Learning Rate: During each run, there is an initial pause while example begin to accumulate, followed by a rapid climb. Sampling rate has no significant impact on the learning rate (Figure 9(a)). Extrinsic activity has a small effect (Figure 9(b)): the noon runs finish slightly larger than the midnight and 8am runs. This may hint that an intermediate rate of activity is best for learning, but the significant difference between the conditions is to small to base such a conclusion on these results. The number of relations does not appear to grow indefinitely: the long recording plateaus at around 14,000 seconds (Figure 9(c)), though another late set of weak or rare relations might still be building up strength.

Conclusions

Differences between specialists in a cognitive substrate need not make it hard for them to agree on expressive, composable signals. We have seen such learning demonstrated using heuristic methods based on Allen's time relation and a vision specialist and hearing specialist observing a simulated four-way intersection. Furthermore, the differences between the two specialists may be beneficial, as the predictive relations acquired to connect their signals turn out to capture interesting structure and dynamics from the simulation.

These experiments show that this apparently difficult problem is quite tractable. With more careful study—particularly of how joint attention and signal agreement can build off one another—we may expect to see great improvement in the quality of cross-specialist relations discovered. These distributed representations are perhaps the most important direction for future exploration, as they may allow the automation of routine integration, dramatically simplifying the construction of systems and models containing many specialist components. If so, further study of this capability is important not only for the cognitive substrate hypothesis, which depends intimately on integration, but for the broader field of cognitive architectures and perhaps software engineering in general.



(a) Sampling Rate (b) Extrinsic Activity (c) Long Training

Figure 9: Relations are acquired rapidly following an initial pause. The rate is not affected significantly by sampling rate (a) and minimally by extrinsic activity (b). After long training, the number of relations plateaus, appearing to converge (c).

Acknowledgements

This work was supported by NSF Grant #6898853. Portions of this paper appear in (Beal, 2007).

References

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, *26*(11), 832–843.

Anderson, J., Bothell, D., Byrne, M., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (chap. 5). Cambridge University Press.

Beal, J. (2002a). An algorithm for bootstrapping communications. In *International conference on complex systems (iccs 2002).*

Beal, J. (2002b). *Generating communications systems through shared context*. Unpublished master's thesis, MIT.

Beal, J. (2007). *Learning by learning to communicate*. Unpublished doctoral dissertation, MIT.

Carey, S. (2004, Winter). Bootstrapping and the origin of concepts. *Daedalus*, 59–68.

Cassimatis, N. (2002). *Polyscheme: A cognitive architecture for integrating multiple representation and inference schemes*. Unpublished doctoral dissertation, MIT.

Cassimatis, N. (2006). A cognitive substrate for achieving human-level intelligence. *AI Magazine*, *27*(2), 45–56.

Cassimatis, N., Bugajska, M., Dugas, S., Murugesan, A., & Bello, P. (2007). An architecture for adaptive algorithmic hybrids. In *Aaai-07.*

Coen, M. (2006). *Multimodal dynamics: Self-supervised learning in perceptual and motor systems*. Unpublished doctoral dissertation, MIT.

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

Kieras, D., & Meyer, D. (1997). An overview of the epic architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, *12*, 391-438.

Kirby, S. (1998). *Language evolution without natural selection: From vocabulary to syntax in a population of learners* (Tech. Rep.). Language Evolution and Computation Research Unit, University of Edinburgh.

Kirby, S. (2002a). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (chap. 6). Cambridge University Press.

Kirby, S. (2002b). Natural language from artificial life. *Artificial Life*, *8*, 185–215.

Kohonen, T. (1989). *Self-organization and associative memory* (3rd ed.). Berlin: Springer-Verlag.

Langley, P., & Choi, D. (2006). A unified cognitive architecture for physical agents. In *Aaai-06.*

Minsky, M. (2006). *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon & Schuster.

Roy, D. (1999). *Learning from sights and sounds: A computational model*. Unpublished doctoral dissertation, MIT.

Singh, P. (2005). *Em-one: An architecture for reflective commonsense thinking*. Unpublished doctoral dissertation, MIT.

Spelke, E. (2003). What makes humans smart? In D. Gentner & S. Goldin-Meadow (Eds.), *Advances in the investigation of language and thought.* MIT Press.

Steels, L. (1996). Emergent adaptive lexicons. In P. Maes (Ed.), *Sab96.* Cambridge, MA: MIT Press.

Wang, Y., & Laird, J. (2006). *Integrating semantic memory into a cognitive architecture* (Tech. Rep. No. CCA-TR-2006-02). University of Michigan.

Yanco, H. (1994). *Robot communication: issues and implementations*. Unpublished master's thesis, MIT.