Constraining Bayesian Inference with Cognitive Architectures: An Updated Associative Learning Mechanism in ACT-R

Robert Thomson (thomsonr@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA, 15213, USA

Christian Lebiere (cl@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA, 15213, USA

Abstract

Bayesian inference has been shown to be an efficient mechanism for describing models of learning; however, concerns over a lack of constraint in Bayesian models (e.g., Jones & Love, 2011) has limited their influence as being a description of the 'real' processes of human cognition. In this paper, we review some of these concerns and argue that cognitive architectures can address these concerns by constraining the hypothesis space of Bayesian models and providing a biologically-plausible mechanism for setting priors and performing inference. This is done in the context of the ACT-R functional cognitive architecture (Anderson & Lebiere, 1998), whose sub-symbolic information processing is essentially Bayesian. To that end, our focus in this paper is on an updated associative learning mechanism for ACT-R that implements the constraints of Hebbian-inspired learning in a Bayesian-compatible framework.

Keywords: cognitive architectures; Bayesian inference; Hebbian learning; cognitive models; associative learning;

Introduction

Bayesian approaches to reasoning and learning have been successful in such fields as decision-making (Tenenbaum, Griffiths, & Kemp, 2006), language learning (Xu & Tenenbaum, 2007), and perception (Yuille & Kersten, 2006). Most specifically, Bayesian inference has been exceptional in discovering some of the structure of language and word learning with substantially less training than traditional connectionist networks.

Despite their successes, Bayesian models have come under attack for being unconstrained, unfalsifiable, and overly reliant on optimality as an assumption for reasoning (see Jones & Love, 2011; Bowers & Davis, 2012 for an exhaustive review; and Griffiths et al., 2012 for a counter-argument). While these criticisms are not without merit (nor are the Bayesians' rebuttals fully convincing), the issue of constraints remains a critical argument. It is also not a new argument. Over 25 years ago the constraint argument was leveled against the field of connectionism (Fodor & Pylyshyn, 1988). Then it was argued that, via several learning rules and organizing principles, any behavior could theoretically be captured by connectionist networks.

The degree that progress has slowed for the explanatory power of connectionist networks is beyond the scope of this paper; however, constraints on neural network development using a common learning rule in a stable cognitively-plausible architecture have been advanced (O'Reilly, 1998; O'Reilly, Hazy, & Herd, 2012). By corollary, to address similar concerns, the Bayesian movement needs to develop constraints which balance the computational transparency of their models with algorithmic and implementation (i.e., neural) level cognitive plausibility.

Interestingly, ACT-R 6.0 (Anderson et al., 2004) is a cognitive architecture which already uses Bayesian-inspired inference to drive sub-symbolic learning (i.e., to generate and update the activation strength of chunks in declarative memory). The architecture is both constrained by learning rules (e.g., activation equations; base-level learning) and neuro-cognitively justified by many studies (Anderson & Lebiere, 1998; Anderson et al., 2004; Anderson, 2007). While there have been difficulties in adapting some aspects of the Bayesian approach (e.g., in implementations of associative learning), ACT-R serves as an example whereby Bayesian inference can be constrained by a neurally-localized and behaviorally-justified cognitive architecture. In this sense, ACT-R can act as a bridge between all three layers of Marr's tri-level hypothesis.

For the remainder of this paper, we present an overview of the debate over the applicability of Bayes inference to cognition and argue that ACT-R represents the kind of constraint that addresses criticisms against Bayesian models. We will further describe an updated associative learning mechanism for ACT-R that links Bayesian-compatible inference with a Hebbian-inspired learning rule.

Bayesian Inference

The essential feature of Bayesian inference is that it reasons over uncertain hypotheses (H) in probability space (i.e., from 0 – 100% certainty). The Bayes rule is defined as:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

where the posterior probability of an outcome $P(H|D)$ is derived from the likelihood $P(D|H)$ of the hypothesis explaining the data, combined with the prior probability of the hypothesis $P(H)$, and normalized by the probability of the data P(D). Thus, updating one's belief is based on one's prior belief influenced by the likelihood that some new evidence supports this belief. At its core, Bayesian inference is an excellent derivation of the scientific method.

A difference between Bayesian models and connectionist implementations is that Bayes models of human cognition tend to use richer, more structured, and symbolic knowledge than connectionist models, which tend to use more distributed representations operating over less structured input. This level of inference places Bayesian models at the computational level of Marr's tri-level hypothesis, whereas cognitive architectures and connectionist networks operate

more at the algorithmic level (Marr, 1982). By remaining at a higher level of description, it is argued that Bayesian descriptions of cognitive behaviors are better understood as a framework for explaining cognition as opposed to an explanation of how cognitive operations and representations should behave in a given task (Tenenbaum et al., 2011).

This higher level of description leads to many of the criticisms leveled against Bayesian models. We wish to address three related criticisms of Bayesian models: (1) they are unconstrained; (2) they are unfalsifiable; and (3) there is little neuro-scientific evidence to support Bayesian theory. It is easy to see how (2) and (3) follow from (1) since without constraint, it is theoretically possible to redefine the priors and hypothesis space of the model to curve fit to any data. Part of the issue with (3) is that Bayesian description tends to operate at the computational level, yet be described in stronger, more algorithmic terms (e.g., probabilistic population codes; Ma et al. 2006).

These criticisms have led to Bayesian theory being criticized as a *'just-so'* story (i.e., that the Bayesian framework commits the *ad hoc* fallacy; Bowers & Davis, 2012). However, rebuttals by Griffiths et al. (2012), rather than addressing these criticisms in a constructive manner, countered with essentially a *'you-too'* argument. Griffiths et al. (2012) argued that curve-fitting models to data is not an exclusive sin of Bayesian models, however, the transparency with which Bayesian models do so make them easy targets. In fact, as Griffiths et al. counter, criticisms (1) and (2) may be leveled against any model or architecture with sufficient parametric degrees of freedom (which they implicitly argue is a feature of most or all existing models). This argument against architectures had previously been espoused by Roberts and Pashler (2000) over a decade ago.

In a recent Science article by Tenenbaum et al., (2011) Bayesian inference is defined as being synonymous with probabilistic inference. This leads to criticism (2). The difficulty with making 'Bayesian' and 'probabilistic' synonymous terms is that any algorithm that approximates probabilistic reasoning can be argued to be approximating Bayesian inference and thus be essentially Bayesian. Conversely, any Bayesian algorithm that does not successfully reproduce human data can lead to the argument that the issue isn't with the Bayesian algorithm per se, but in the transformation of data into a probability space (e.g., by not having the correct priors or correct hypotheses) or in the lack of human-like limitations of the algorithms to carry out the computations. It is for this reason that some have argued that probabilities are "epistemologically inadequate" (McCarthy & Hayes, 1969).

Instead of offering more criticisms, we wish to offer solutions. The issue with constraints is that, even if Bayesian models do not have too many parameters, there is effectively unlimited freedom in setting priors and the hypothesis space (which greatly influences the performance on the model). What is needed is a way to constrain the generation of the initial probability space and set of algorithms to carry out inference for a set of models. For instance, Kruschke (2008) reviewed two Bayesian models of learning backward blocking in classical conditioning, the first using a Kalman filter (Dayan, Kakade, & Montague, 2000) and the other using a noisy-logic gate (Danks, Griffiths, & Tenenbaum, 2003). Both models gave substantively different predictions, with the Kalman filter model unable to reproduce human behavior.

Furthermore, there are several tasks whose results do not readily fit within a naïve Bayesian explanatory framework. For instance, simple Bayesian models do not capture violations of the sure-thing principle. Given a random variable x that has only two possible outcomes A or B, naïve Bayesian inference requires $p(x)$ to fall between $p(x|A)$ and $p(x|B)$. A violation occurs when $p(x) > p(x|A)$ and $p(x) > p(x|B)$ or vice versa. Shafir and Tversky (1992) showed this violation of the sure-thing principle in a prisoner's dilemma task. Finding these unintuitive results that naïve Bayes models do not easily address, and finding constrained parameter learning rules (such as the noisy-logic gate) provides much needed constraints and falsifiability to the Bayesian framework. Rather than being seen as anti-Bayesian results, these models should be seen as shaping the boundaries of Bayesian explanatory power.

Finally, while there is contested neuro-scientific evidence as to neural assemblies firing probabilistically, this does not necessarily imply a Bayesian implementation-level explanation, but instead implies the softer claim of a Bayesian-compatible behavioral explanation of neural phenomena, especially when the Bayesian inferences are justified within a neurally-plausible cognitive architecture.

In considering many of the criticisms of Bayesian theory, it is important to note that more research needs to be done to find constraints. As we previously argued, connectionist networks were not sufficiently constrained until sufficient model testing was performed and architectures developed using a common learning rule and constrained set of parameters. For the Bayesian framework, we argue that all of criticisms (1) – (3) can be addressed by situating Bayesian inference within a cognitive architecture, and furthermore that ACT-R 6 is already such an architecture.

The ACT-R Architecture

ACT-R is a computational implementation of a unified theory of cognition. It accounts for information processing in the mind via task-invariant mechanisms constrained by the biological limitations of the brain. ACT-R 6 includes long-term declarative memory and perceptual-motor modules connected through limited-capacity buffers. Each module exposes a buffer, which contains a single chunk, to the rest of the system. A chunk is a member of a specific chunk type, and consists of a set of type-defined slots containing specific values.

The flow of information is controlled by a procedural module implemented using a production system, which operates on the contents of the buffers and uses a mix of parallel and serial processing. Modules may process information in parallel with one another. So, for instance, the visual and motor modules may both operate at the same

time. However, there are two serial bottlenecks in process. First, only one production may execute during a cycle. Second, each module is limited to placing a single chunk in a buffer.

Each production consists of if-then condition-action pairs. Conditions are typically criteria for buffer matches, while the actions are typically changes to the contents of buffers that might trigger operations in the associated modules. The production with the highest utility is selected to fire from among the eligible productions. In general, multiple production rules can apply at any point. Production utilities, learned using a reinforcement learning scheme, are used to select the rule that fires.

When a retrieval request is made to declarative memory (DM), the most active (highest A_i) matching chunk is returned:

$$A_i = B_i + S_i + P_i + \mathcal{E}_i$$

where activation A_i is computed as the sum of base-level activation (B_i), spreading activation (S_i), partial matching (P_i) and stochastic noise (ε_i). Spreading activation is a mechanism that propagates activation from the contents of buffers to declarative memory proportionally to the strength of association between buffer contents and memory chunks. Partial matching is a mechanism that allows for chunks in memory that do not perfectly match a retrieval request to be recalled if their activation overcomes a similarity-based mismatch penalty.

ACT-R as a Constrained Bayesian Architecture

ACT-R's sub-symbolic activation formula approximates Bayesian inference by framing activation as log-likelihoods, with base-level activation (B_i) as the prior, the sum of spreading activation and partial matching as the likelihood adjustment factor(s), and the final chunk activation (A_i) as the posterior. The retrieved chunk has an activation that satisfies the maximum likelihood equation.

ACT-R provides the much needed constraint to the Bayesian framework through the activation equation and production system. The calculation of base-levels (i.e., priors) occurs within both a neurally- and behaviorally-consistent equation:

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right)$$

where n is the number of presentations for chunk i, t_j is the time since the j^{th} presentation, and d is a decay rate (community default value is .5). This formula provides for behaviorally-relevant memory effects like recency and frequency, while providing a constrained mechanism for obtaining priors (i.e., driven by experience). Thus, we can address the constraint criticism (1) through this well justified mechanism (see Anderson et al., 2004).

In addition, the limitations on matching in the production system provide constraints to the hypothesis space and kinds of inferences which can be made. For instance there are constraints on the kinds of matching that can be accomplished (e.g., no disjunction, matching only to specific chunk types within buffers) and, while user-specified productions can be task-constrained, the production system can generate novel productions (through proceduralization) using production compilation. In addition, the choice of which production to fire (conflict resolution) also constrains which chunks (i.e., hypotheses) will be recalled (limiting the hypothesis space), and are also subject to learning via production utilities.

In production compilation, a new production is formed by unifying and collapsing the conditions of the production, and possibly automatizing a given memory retrieval. This new production has a unique utility and can be considered an extension of the hypothesis space; perhaps with enough learning compiled productions are more analogous to overhypotheses (Kemp, Perfors, and Tenenbaum, 2007).

In summary, the conflict resolution and production utilities algorithms both constrain the hypothesis space and provide an algorithm for learning how the space will evolve given experience, constrained within the bounds of a neurally-consistent functional cognitive architecture. This bridges Bayesian inference from a computational-level framework within an algorithmic-level architecture. However, this argument for constraint is not without criticisms (some of which will be addressed in the *Discussion*). As an example of increasing constraints and grounding mechanisms, we will now present an updated associative learning mechanism in ACT-R.

Associative Learning

Associative learning - the phenomenon by which two or more stimuli are associated together - is ubiquitous in cognition, describable as both a micro (Hebbian learning between neurons) and macro (classical and operant conditioning) feature of behavior. Associative learning is a flexible and stimulus-driven mechanism which instantiates many major phenomena such as classical conditioning, context sensitivity, non-symbolic spread of knowledge, and pattern recognition (including sequence learning and prediction error). At the neural level, associative learning is the process by which *cells that fire together, wire together*.

In its simplest form, Hebbian learning can be described as: $\Delta W_{ij} = x_i x_j$, where W_{ij} is the synaptic strength of the connection between neurons i and j, and x_i and x_j are the inputs to i and j (Hebb, 1949). When both i and j are active together, W_{ij} is strengthened. While the traditional Hebbian rule was unstable due to a lack of mechanisms to control for weakening of connections (i.e., long-term depression; LTD) or to set a maximum state of activation (i.e., to implement a softmax equation; Sutton & Barto, 1998), several variants have addressed these issues to provide a stable learning rule.

At a macro level, associative learning is a mechanism where, when a stimulus is paired with a behavior, future presentation of the stimulus primes this behavior. Models of classical conditions are a common macro-level application of associative learning. At this level, associative learning allows animals and humans to predict outcomes based on prior experience with learning mediated by the degree of match between the predicted outcome and the actual result (Rescorla & Wagner, 1974; Pearce & Hall, 1980).

While macro-level models are normally processed at a more symbolic level, micro-level sub-symbolic processing can capture statistical regularities from the environment without recourse to explicitly coding context information. There is evidence that humans do not explicitly encode positional information when sequentially recalling a list of items, yet ACT-R's model of list memory required explicit position information to drive recall (Anderson et al., 1998).

Despite being a pervasive factor of human intelligence, associative learning is no longer directly implemented in ACT-R. One reason for this absence is due to difficulties in scaling models in its Bayesian implementation of associative strengths, which treated both the activation strength and associative strength of knowledge elements (e.g., chunks) as likelihoods of successful recall.

Bayesian Associative Learning Rule

Associative learning was deprecated in ACT-R 5 due to a lack of scalability in spreading activation as the number of chunks in a model increased and as new productions fired (i.e., new contexts generated). Instead, a simpler spreading activation algorithm was used. The reason for this was that the Bayesian formula used to calculate strength of association (S_{ji}) led to some unintended consequences which would render larger and longer-running models unstable.

In ACT-R 4/5, the strength of association (S_{ji}) represented the log likelihood ratio that chunk N_i was relevant given context C_j:

$$S_{ji} = ln\left(\frac{P(N_i|C_j)}{P(\bar{N}_i|C_j)}\right) = \frac{P(N_i)}{P(\bar{N}_i)} \prod_j \frac{P(C_j|N_i)}{P(C_j|\bar{N}_i)}$$

When C_j is usually not in the context when N_i is needed, $P(N_i|C_j)$ will be much smaller than $P(\bar{N}_i|C_j)$ and the S_{ji} will be very negative because the log-likelihood ratio will approach 0. In a long-running model, these chunks may have been recalled many times without being in context together, leading to strongly inhibitory S_{ji}.

Once a connection was made, the initial prior S_{ji} was set by the following equation:

$$S_{ji} = \ln(m/n)$$

where m is the total number of chunks in memory and n is the number of chunks which contain the source chunk j. This ratio is an estimation of the likelihood of retrieving chunk i when j is a source of activation. As a convenience unconnected chunks were set at 50% likelihood.[1]

As can be seen from the previous two equations, given sufficient experience or sufficient numbers of chunks in the model, these context-ratio equations specify that S_{ji} values will become increasingly and unboundedly negative as more chunks are present in the model and more unique contexts experienced. This is a direct result of S_{ji} reflecting the statistics of retrieval of chunk j given that source i is in the context, and is a version of the Naïve Bayes Assumption.

The issue is with the ratio-driven global term (C_j) which alters S_{ji} values for a chunk whenever a new chunk is added and/or production fires, and is magnified by the log-likelihood calculation which penalizes the inevitable low context ratio in long-running models.

Spreading Activation in ACT-R 6

Due to the abovementioned issues with scalability, associative learning was deprecated in ACT-R and a simpler spreading activation function was implemented that does not activation, but instead spreads a fixed amount of activation:

$$S_{ji} = smax - \ln(fan_{ji})$$

where *smax* is a parameterized set spread of association (replacing the m term from the previous equation), and fan_{ji} is the number of chunks associative with chunk j (the n term). Fan_{ji} is traditionally the number of times chunk j is a slot value in all chunks in DM and represents interference.

With a default *smax* usually between 1.5 and 2 (Lebiere, 1999), this means that a chunk can appear as a value in 6-8 chunks before becoming inhibitory. In the context of a modeling a single session psychology experiment this may be reasonable, but if ACT-R models long-term knowledge effects, then S_{ji} will become inhibitory for most chunks.[2]

As previously discussed, associative learning is a ubiquitous mechanism in both human and animal cognition, which serves as a kind of statistical accumulator which is applicable at both the micro (neural) and macro (cognitive) behavioral level. It seems that to abstract this essential learning mechanism, we are losing out on the exact kind of human-model comparisons that might provide evidence for these much-needed constraints. Perhaps, it is in part for this reason that ACT-R (and other cognitive architectures) have had their explanatory power limited due to a lack of newer, more complex models being built from extant successful models (ACT-R Workshop, 2012).

To both reconcile the difficulties in previous implementation of associative learning and show how we can constrain Bayesian-compatible inference in a cognitive architecture, we will now present a Hebbian-inspired associative learning rule influenced by spike-timing dependent plasticity (STDP; Caporale & Dan, 2008).

Hebbian-Inspired Associative Learning Rule

The major issues with the Bayesian associative learning rule were the reliance on ratio-driven log-likelihoods and the fact that context (C_j) was a global term which altered S_{ji} whenever a new chunk was created and whenever a production fired. This is due to the fact that low log-likelihoods become strongly inhibitory, and the generation of context-based ratios necessitates low-likelihoods in a long-running model. In short, this Bayesian account based on the Naïve Bayes Assumption does not adequately capture some of the features of associative learning such as locally-driven strengthening of associations and bounded decay.

An alternative framework is to eliminate the ratio function and remove the global nature of context, while also moving to a frequency-based algorithm instead of a probability-based algorithm. The former removes the aforementioned

[1] Before C_j appears in a slot of N_i, the total probability of retrieving a chunk unconnected to C_j is 0 (which means $S_{ji} = -\infty$).

[2] After presenting this at the 2012 ACT-R Workshop, a flag was written in ACT-R to set a floor of 0 in the S_{ji} computation.

issues with scalability, while the latter eliminates $S_{ji} = \lim_{x \to 0}(\ln x) = -\infty$, where x is the likelihood. That said, a benefit of using log-likelihood in probability space is that there is no need to squash activation strength (e.g., use a softmax rule to keep S_{ji} values from overwhelming B_i in the activation equation) because likelihoods cannot go above 100% while frequency-based Hebbian activations can theoretically grow unbounded. Thus, the switch to frequencies is about reshaping the range of S_{ji} values and making S_{ji} independent of changing global context.

Basing associative learning on frequencies also adds a more Hebbian flavor to the algorithm. Learning, rather than being a global property of the system (as in the Bayesian mechanism) is instead a local property based on co-occurrence and sequential presentation. As previously discussed, our Hebbian-inspired mechanism is influenced by STDP. Unlike traditional Hebbian implementations which simply give a bump to association so long as the pre-synaptic and post-synaptic neurons both fire within a given temporal window, in STPD if the pre-synaptic neuron fires before the post-synaptic then the association is strengthened (long-term potentiation; LTP). Conversely, if the post-synaptic neuron fires before the pre-synaptic then the association is inhibited (long-term depression; LTD).

This theory of neural plasticity was adapted to our modeling approach by assuming that the sources of activation from chunks in buffers act similarly to pre-synaptic firings, and the set of chunks in the buffers at the time the new chunk is retrieved is similar to post-synaptic firings. The associative learning rule fires when a request is made to retrieve a chunk from declarative memory. First, a positive phase occurs (LTD; or Hebbian) where the current contents of the buffers spread activation and a new chunk is retrieved. The association between this new chunk and the sources of activation are strengthened according to standard Hebbian learning rules. However, once this new chunk is placed in the retrieval buffer, a negative phase occurs (LTP; or anti-Hebbian) where the retrieved chunk will negatively associate with itself and with its context. In formal terms:

$$\Delta S_{ji} = \alpha \cdot F\left(N_i | C_j^{pre}\right)$$
$$\Delta S_{ji} = -\alpha \cdot F\left(N_i | C_j^{post}\right)$$

where α is a Hebbian learning term, $F\left(N_i | C_j^{pre}\right)$ is the context of source chunks C_j^{pre} at the time of the retrieval request for chunk N_i, and $F\left(N_i | C_j^{post}\right)$ is the context of chunks C_j^{post} after chunk N_i has been retrieved. Note that only changes in context will have a net ΔS_{ji} due to the balanced positive and negative learning phase. Furthermore, these associations are not symmetric (i.e., $S_{ji} \sim= S_{ij}$).

This balanced Hebbian/anti-Hebbian mechanism is geared towards developing a local, scalable learning rule while maximizing neural plausibility by incorporating a negative inhibitory learning phase. We argue that this inhibitory phase, while seemingly unintuitive[3], is actually a relevant and necessary mechanism to account for refractory periods in neural firings.

An advantage of this Hebbian-inspired implementation is that it avoids the inhibitory associations of low log-likelihoods, but the learning rule requires a form of softmax equation (either driven by expectation or more simple decay/inhibition) to keep S_{ji} values from overwhelming base-level B_i (i.e., from the likelihood overwhelming the prior, in Bayesian terms). At the micro/neural level, softmax approximates a maximum likelihood, while at a macro/behavioral level, softmax simulates learning as expectation violation. In Bayesian terms, the more active (c.f., likely) the existing association between chunks $A \rightarrow B$, then the less marginal increase in S_{ji} when chunk A is a source in the retrieval of chunk B.

There are several beneficial effects from this kind of implementation. The first is that the mechanism is more balanced and geared towards specializing associative activations rather than just increasing all activations. Thus, the mechanism is more stable as it grows (i.e., it will not tend towards all associations becoming either strongly excitatory or inhibitory; S_{ji} doesn't vary with number of chunks in memory). Second, since the retrieved chunk self-inhibits, this reduces the chance that it will be the most active chunk in the following retrieval request (due to recency effects), which can cause models to get into self-feedback loops. In short, this inhibition leads to a natural refractory period for retrieving a chunk. Third, by self-inhibiting and spreading activation to the next context, it provides a forward momentum for the serial recall of chunks. Combined with recency and frequency of base level, this provides a mechanism for automatic serial recall of lists without the need for coding of explicit positional information (something required in prior models of list memory; Anderson et al., 1998) and marking of previously retrieved chunks through finst-like mechanisms. The uniqueness of the subsequent context drives order effects.

There are still, however, several design decisions and more empirical justification required in order to strengthen the constraint argument. Currently, the softmax learning term is based on ACT-R's base-level learning equation. However, several candidate equations need to be compared against human performance data to determine the best possible match. Furthermore, existing models of list memory and sequence learning need to be re-envisioned in terms of the new associative learning mechanism.

In summary, this balanced Hebbian/anti-Hebbian learning mechanism avoids the issues of scalability (e.g., runaway activations) that have been associated with prior implementations of associate learning in ACT-R. In addition, this mechanism is constrained by neural plausibility constraints, can still be discussed in Bayesian-compatible terms, and fits within the Bayesian description of ACT-R's sub-symbolic activation.

Discussion

This paper has described how a functional cognitive architecture can constrain Bayesian inference by tying

[3] Some have found the notion of a chunk being self-inhibitory very unintuitive, because it conflicts with the idea that a chunk should be maximally similar to itself and self-activating.

neurally-consistent mechanisms into Bayesian-compatible sub-symbolic activations. This combination of grounded implementation- and algorithmic-level functions into cognitive-level Bayesian inference defuses many criticisms of Bayesian inference, and provides a launch-point for future research into constraining the Bayesian framework across all three levels of Marr's hypothesis. An example of this research was provided by examining a novel implementation for associative learning in ACT-R. In addition to the sub-symbolic layer being driven by Bayesian mathematics, it is also compatible with neural localization and the flow of information within the brain.

It has been argued that ACT-R's numerous parameters don't really provide the kind of constraint necessary to avoid the criticisms discussed in this paper (Tenenbaum et al., 2011). However, the use of community and research-justified default values, the practice of removing parameters by developing more automatized mechanisms (such as the associative learning replacing spreading activation), and the development of common modeling paradigms mitigates these criticisms by limiting degrees of freedom in the architecture and thus constraining the kinds of models that can be developed and encouraging their integration. In summary, the evolution of the architecture is not a process of invalidation, but instead moving towards more constrained and more specific explanations.

As we have argued, the associative learning mechanism is an attempt to increase constraint within the architecture and promote a broader explanatory power to numerous cognitive phenomena. This mechanism is geared towards specializing associative strength to capture both symbolic and non-symbolic associative learning. A major contribution of this mechanism is its balance between Hebbian (LTP) and anti-Hebbian (LTD) learning at each retrieval request, which provides numerous benefits over traditional Hebbian and Bayesian implementations.

Acknowledgments

This work was conducted through collaboration in the Robotics Consortium sponsored by the U.S Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016; and by Intelligence Advanced Research Projects Activity via DOI contract number D10PC20021. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review, 8,* 629-647.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Qin, Y. (2004). An integrated theory of Mind. *Psychological Review, 111,* 1036-1060.

Anderson, J. R., Bothell, D., Lebiere, C. & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, 38, 341-380.

Anderson, J. R., and Lebiere, C. (1998). *The atomic components of thought*, Erlbaum, Mahwah, NJ.

Bowers, J. S., & Davis, C. J. (2012). Bayesian Just-So Stories in Psychology and Neuroscience. *Psychological Bulletin, 138 (3)* 389-414.

Caporale, N., & Dan, Y. (2008). Spike Timing-Dependent Plasticity: A Hebbian Learning Rule. Annual Review of Neuroscience, 31, 25-46.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *WIREs Cognitive Science*, 1, 811-823.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical casual learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural information processing system.* MIT Press: Cambridge, MA.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention, *Nature Neuroscience*, 3, 1218-1223.

Deneve, S. (2008). Bayesian spiking neurons II: Learning. *Neural Computation, 20,* 118-145.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition, 28 (1),* 3-71.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got Their Beliefs (and What Those Beliefs Actually Are): Comment on Bowers and DAvids (2012). *Psychological Bulletin, 138 (3),* 415-422.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10 (3),* 307-321.

Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception and Psychophysics, 35(6),* 536-542.

Kruschke J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning and Behaviour, 36 (3),* 210-226.

Lebiere, C. (1999). The dynamics of cognition: An ACT-R model of cognitive arithmetic. *Kognitionswissenschaft.*, 8 (1), pp. 5-19.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9,* 1432-1438.

O'Reilly, R. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Science, 2 (11),* 455-462.

O'Reilly, R., Hazy, T. E., & Herd, S. A. (2012). The Leabra Cognitive Architecture: How to Play 20 Principles with Nature and Win! In S. Chipman (Ed) *Oxford Handbook of Cognitive Science*, Oxford: Oxford University Press.

Pearce J. M. and Hall G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87,* 532-552.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical Conditioning II,* Appleton-Century-Crofts.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358-367.

Shafir, E. & Tversky, A. (1992) Thinking through uncertainty: nonconsequential reasoning and choice. *Cognitive Psychology* 24: 449-474.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, 10, 309–318.

Varma, S. (2011). Criteria for the Design and Evaluation of Cognitive Architectures. *Cognitive Science, 35 (7),* 1329-1351.

Xu, F., & Tenenbaum, J. B. (2007). Word Learning as Bayesian Inference, *Psychological Review, 114 (2),* 245-272.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Science, 10,* 301–308.