A Model of Infant Causal Perception and its Development

Harold Henry Chaput (chaput@cs.utexas.edu)
Department of Computer Sciences, Taylor Hall 2.124 [C0500]
The University of Texas at Austin
Austin, TX 78712-1188 USA

Leslie B. Cohen (cohen@psy.utexas.edu)
Department of Psychology, Mezes Hall 330 [B3800]
The University of Texas at Austin
Austin, TX 78712 USA

Abstract

The acquisition of infant causal perception has been the center of considerable debate, and some have attributed this phenomenon to an innate causal module. Recent studies, however, suggest that causal knowledge may develop in infants through experience with the environment. We present a computational model of causal knowledge acquisition built using the Constructivist Learning Architecture, a hierarchical self-organizing system. This system does a remarkably good job of developing causal perception from a component view to a holistic view in a way that mirrors data from habituation studies with human infants.

Causal Perception in Infants

Causal perception has been the focus of philosophical inquiry for centuries, but it received its first notable psychological investigation by Michotte (1963). He presented adults with a scene in which one billiard ball struck another stationary ball, resulting in the launching of the stationary ball, and the halting of the moving ball. The subjects, naturally, described this scene as a "causal" event. But by manipulating the launching event (along spatial or temporal dimensions), Michotte could affect a subjects' likeliness of perceiving causality. One can alter the *spatial* component of the event by introducing a gap between the two balls, so that agent and the object never actually touch. Also, one can alter the *temporal* component by introducing a delay between the moment of contact and the moment of launching. As these components deviated from zero gap and zero delay, adult subjects were less likely to classify the event as "causal." These events are illustrated in Figure 1.

Since then, researchers have combined these events with habituation techniques to demonstrate the presence of causal perception in infants. One such study, by Leslie (1984), was able to demonstrate 6 1/2-month-old infants' ability to discriminate between different launching events. Leslie further demonstrated that infants' responses were based, in part, on the causality of the event. For example, infants habituated to a causal event would dishabituate to a non-causal event (e.g. from direct to gap), or vice versa. But infants would not dishabituate to the same degree if the causality remained constant between events (e.g. from delay to delay+gap). Leslie then claimed that these results, since they came from such young infants, were the product of an innate "causal module".

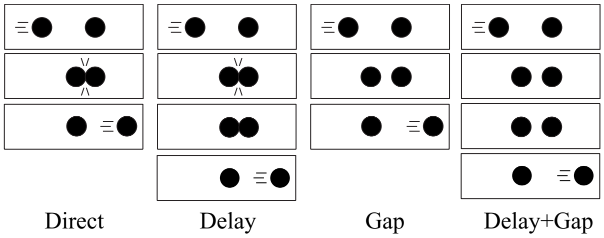


Figure 1: Four different launching events.

More recent studies, though, have cast doubt on a nativist view of causal perception. Cohen and Amsel (1998) performed a similar experiment on 6 1/4-month-old infants, and re-affirmed their ability to discriminate causal events from non-causal events. But they then ran the same experiment on 5 1/2-month-old and 4-month-old infants and found that these younger infants were not able to discriminate between launching events strictly on the basis of causality. Rather, infants responded to the spatial and temporal *components* of the event. Unlike the older infants, younger infants would respond to the introduction or removal or either a delay or a gap, regardless of how this change impacted the causality of the event.

Cohen and Amsel (1998) posited that these results indicated a developmental component to causal perception. It is this progression from component to high-level concept that we are interested in modeling. The development of causality is just one instance of a more general part-to-whole progression that can be seen in a variety of cognitive developmental domains. Cohen (1998) pointed out numerous studies of developmental cognition that fit this general framework, and proposed an *information processing* approach to cognitive development as an alternative to nativism. Rather than being born with a set of innate perceptual modules, infants start with very low-level perceptual capabilities. This approach is summarized by the following set of propositions (Cohen & Cashon, 2000):

1. Perceptual/cognitive development follows a set of domain-general information processing principles.
2. Information in the environment can be processed at a number of different levels of organization.
3. Higher (more holistic) levels can be defined in terms of the types of relations among lower (parts) levels.
4. Development involves progressing to higher and higher levels.
5. There is a bias to initiate processing at the highest level available.
6. If an information overload occurs (such as when movement is added or when the task involves forming a category), the optimal strategy is to fall back to a lower level of processing.

At the very least, long term development appears to play an important role in the perception of high-level concepts such as causality, regardless of the concept's origin. There are countless learning systems which model knowledge acquisition. But we know of no such model that conforms to the six propositions of developmental information processing described above. There are also very few computational models of infant cognitive development that differentiate between long-term learning and short-term habituation, let alone use one to determine the other. One example in the language development domain has recently reported by Schafer and Mareschal (2001).

Constructivist Learning Architecture

The Constructivist Learning Architecture (CLA) (Chaput, 2001) is a hierarchical self-organizing system designed to generate concepts at multiple levels of abstraction through the organization of sensory input. Using the six propositions of cognitive development listed above as a design specification, CLA was built to learn hierarchical knowledge structures through observation, and use those structures to produce the kind of short-term habituation effects that infants exhibit throughout their development.

The information processing approach to cognitive development described above does not mention any kind of corrective feedback, and thus suggests an unsupervised learning system. For this reason, CLA uses one such system, the Self-Organizing Map (Kohonen, 1997), as a building block. The Self-Organizing Map (SOM) is a two-dimensional matrix of nodes, each of which stores a feature vector. As stimuli are repeatedly presented to the SOM (as feature vectors), the SOM adjusts the feature vectors of its nodes to represent the observed stimuli. A trained SOM exhibits the following attributes: 1) the feature vectors of nodes in a SOM reflect the stimuli presented to the SOM (*environmental representation*); 2) nodes which are close to each other in the network have similar feature vectors (*similarity topography*); and 3) stimuli which occur more often will be represented by a larger number of nodes (*frequency effect*).

But although the SOM performs the kind of unsupervised category formation that appears to be at work in cognitive development, it does not by itself form the kind of hierarchical knowledge representation suggested by the information processing approach.

CLA achieves this hierarchical representation by connecting multiple SOMs into a hierarchy. Like a regular SOM, the lowest layer of CLA (Level 1) receives raw input from the environment. When a stimulus is introduced, each node in the Level 1 layer receives an activation, *A*, which is in proportion to how close the stimulus is to the nodes' representation. (This is determined using a Euclidean distance metric.) These activation values are then collected for the layer into a corresponding matrix of activation values, or an *activation matrix*. This activation matrix then becomes the input vector to the layer directly above. This process then repeats for as many layers as are contained in the whole system. For an illustration, see Figure 2.

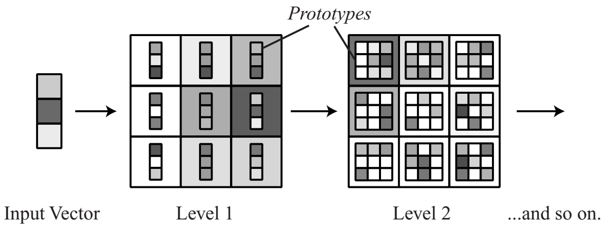


Figure 2: The first two layers of an example CLA. The darkness of each cell represents its level of activation.

Of course, two SOMs connected in this fashion can learn no more or less than a single SOM would. But the stratification of the SOMs allows for processing to occur at intermediate stages. There are three types of intermediate processing that are relevant to the present discussion: activation decay, activation blurring, and multi-modal layers. *Activation Decay* allows the activation of a given node to decay over time, rather than reset at each time step. This is useful as a rudimentary representation of time and sequence. *Activation Blurring* will "blur" the activation matrix by applying a Gaussian filter to the matrix. This has the effect of spreading the activation of a given node to surrounding nodes, which is particularly useful given the SOM's similarity topography feature. Finally, a layer can receive input from two or more lower-level layers. We call these layers *Multi-Modal Layers*.

The result is a system that will have the information processing properties listed above, which we address one by one. First, CLA is not designed for any particular domain, but is a domain-general learning architecture that can readily be applied to any number of areas in cognitive development. Second, when a stimulus is introduced to the system, it will create activation patterns at multiple layers, which are actually different levels of organization; processing of these patterns can occur at any of these layers. Third, because each layer is organizing activation patterns of the layer beneath it, higher-level representations can be defined in terms of the types of relations among lower-level representations.

Fourth, learning in CLA involves progressing to higher and higher levels. When CLA is learning, each layer is trying to build categories out of the activation patterns of the layer beneath it. While one layer is organizing, all the layers above it cannot form stable categories because the underlying activation patterns are unorganized. Only when a layer "settles" into a coherent organization can the layers above it organize. The result is a progression from level to level.

Propositions 5 and 6 involve the resulting behavior of the system, rather than its organization. This paper does not address these propositions directly, but we do consider their ramifications in the discussion section.

Experiment: Learning a Causal Event

We conducted an experiment designed to show whether CLA could model the data produced by human infant subjects in the Cohen and Amsel (1998) study, given the same experiment design and starting set of assumptions. Specifically, we are looking to see if our model exhibits the part-to-whole progression demonstrated in infants between 4 and 6.25 months.

Design

Cohen and Amsel (1998) posit that infants are able to perceive the spatial and temporal components of a launching event before they perceive its causality. For this reason, we present the launching events to the learning system by means of two input vectors. The first input vector (the "movement vector") reported the magnitude of the movement of each of the two balls with two rows of nodes. By ignoring the position of each ball, this layer would use activation decay to represent the temporal information of the launching event and exclude the spatial information. The movement vector represented the movement of each ball, with each ball represented in its own row. The element of the row corresponding to the amount of absolute change in position, scaled to the width of the input vector, was set to 1.0. So, elements ~~to~~ on the right side of the vector represent rapid movement, while the left most element would represent no movement. These values decay over time.

Figure 3, for example, shows the state where the first ball (represented in the top row) is now stationary, but was recently moving; while the second ball (represented in the bottom row) is now moving, but was recently stationary. This state occurs in a direct launching event shortly after a collision.

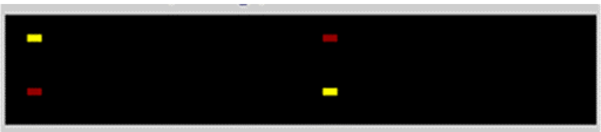


Figure 3: The movement input vector. The decay is shown by a decrease in brightness (or change in color from yellow to red).

The second input vector (the "position vector") reported the position of the balls on the table. The position was represented by a 20 element vector. (The vector only had one row of nodes because the collisions presented were always horizontal.) The positions of the balls on the table were scaled to the 20 element vector, and the element nearest to each ball's position was set to 1.0. Complimentary to the movement vector, the position vector reports the spatial information and excludes the temporal information.

Figure 4: The position input vector.

In Figure four, we see the state where the two balls are close, but not touching. Like Figure 3, this state occurs in a launching event shortly after a collision.

Receiving each input vector was a 5-by-5 node layer to observe it. The "movement layer" organized activation patterns in the movement vector, while the "position layer" organized activation patterns in the position vector. These Level 1 layers would learn the independent spatial and temporal features of the launching event.

Finally, there was a Level 2 layer (the "Top Layer"), which observed *both* bottom-level layers. This layer, 6-by-6 nodes, would learn the combination of activation patterns in the Level 1 layers. Thus, it should discover the combination of spatio-temporal events that comprise each of the events.

The movement vector had an activation decay of 0.25, so that the movement layer could learn the temporal attributes of the ball movements. All other layers had an activation decay of 1.0 (instant decay). Similarly, the position vector had its blurring set to 0.6 to let it see proximal ball positions as similar (that is, having a ball at position 3 is very similar to having a ball at position 4, but very different from having a ball at position 17). All other layers had their blurring set to 0.0 (that is, turned off).

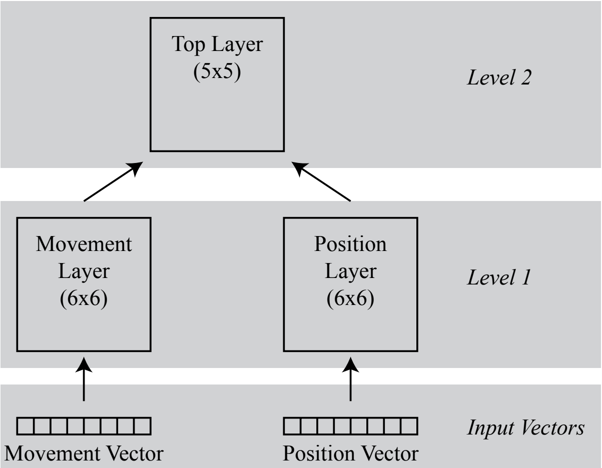


Figure 5: A schematic of the CLA used in the causality experiment.

To generate the launching events, we used a simulated "pool table" written for this experiment. This pool table is used for both the long-term training of the learning system as well as the short-term habituation studies. A simple algorithm was used to represent the state of the pool table through the two input vectors.

For long-term training (meant to represent an infant's long-term experience with the world), the learning system was presented with 2500 complete launching events. The type of launching event presented to the learning system was chosen using a probability meant to approximate roughly the nature of the real world: a direct launching event had a 0.85 probability of being chosen, while delay, gap, and delay+gap events each had a probability of 0.05. The presentation of a complete launching event constituted a single training cycle. The learning rates and neighborhoods of each SOM in the CLA system were decreased as training progressed. (Learning rate decreased from 0.1 to 0.0 over 1000 cycles, and the neighborhood from 1.0 to 0.0 in the same time frame.) This is the customary training procedure when using SOMs.

In order to simulate the changes during the short-term habituation trials, a "familiarity" variable *F* was associated with each node. Remember that we associate an activation *A* with each node. Familiarity for each node always approached the node's activation by some rate using the formula *F*=*F*+*r*(*A*-*F*), where *r* is the rate of approach. We then determined an output *O* for each node using the formula *O*=*A*^(1.0-(*A*-*F*)). Recalling that both *F* and *A* are between 0.0 and 1.0, the result is that output levels are amplified, relative to raw activation, as activation differs from the familiarized level. (See Figure 6 for a graphical representation of these values.) The output for each node in a layer was summed to create a Layer Output. This was then averaged across the duration of the event, giving us a Mean Layer Output. Dishabituation was measured as *change* in Mean Layer Output.

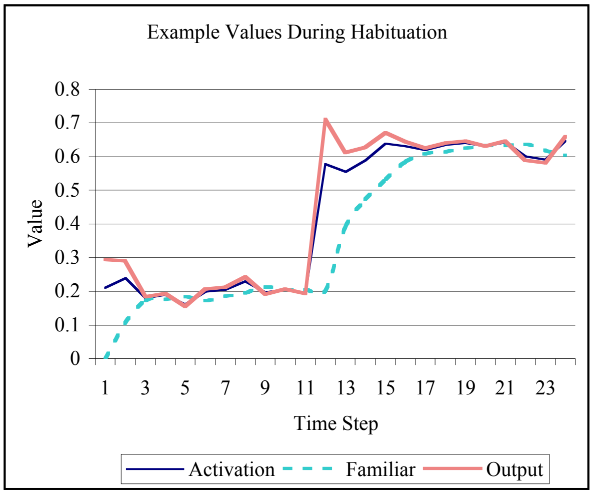


Figure 6: Example values of Activation, Familiarization and Output for a single node during the habituation trials. An event change at time step 12 is represented by a jump in activation.

For each habituation trial, the network was exposed to five repetitions of the habituation event, and then exposed to the test event. A complete habituation experiment consists of 16 parts: four habituation events by four test events. Familiarity levels were cleared for all nodes before each part of the habituation experiment. In all, 12 "simulated infants" were fully trained and tested.

Results

Because of the nature of the events we are dealing with, the difference between a delay event and a gap event should be the same as the difference between a direct event and a delay+gap event. This is because both pairs involve equal changes along the spatial *and* temporal axes. (See Figure 7.)

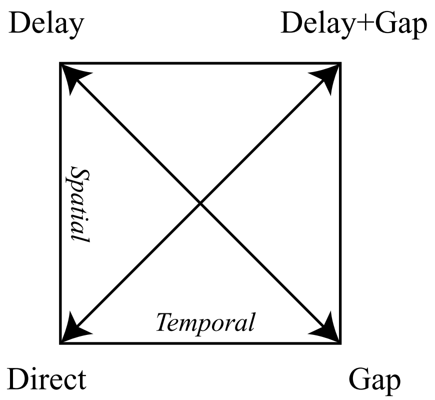


Figure 7: Launching events along spatial and temporal axes. Events at opposite corners involve both a spatial and temporal change of equal amounts and, thus, should be equivalent given a component model.

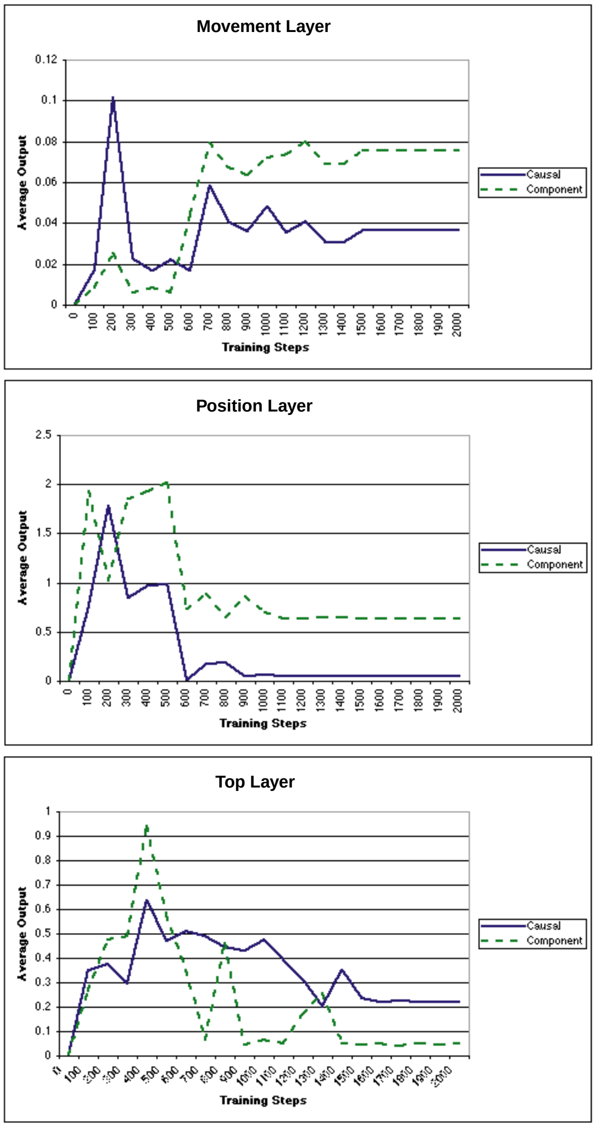
We averaged the dishabituation levels for delay-to-gap trials with gap-to-delay trials, and compared them to the average of direct-to-delay+gap trials and delay+gap-to-direct trials. We did an analysis of variance on both of the Level 1 layers and found that, in fact, both showed a significantly greater response to delay-gap changes, $F(1,11) = 243.3$, $p < .0001$ and $F(1,11) = 34.4$, $p < .0001$ for the movement and position layers, respectively.

This odd disparity comes about because we have designed each of the Level 1 layers to train on one component exclusively, to the exclusion of the other. For example, the movement layer, which is sensitive to temporal differences, is *insensitive* to spatial differences, so that a direct event and a gap event look nearly identical. For this reason, we would expect this layer to see the direct event and the delay+gap event as similar, since the direct event looks like a gap event, and the delay+gap event also has a gap. The converse is true for the position layer. Thus, we would expect these two layers to respond more to the delay-to-gap change.

We verified that this difference was the result of the exclusivity of the input by comparing the average of trials where there was strictly a spatial change to the average of trials where there was strictly a temporal change. There was a significantly greater response to temporal changes than spatial changes in the movement layer, $F(2,22) = 4.1$, $p < .05$. And, conversely, there was a significantly greater response to spatial changes in the position layer, $F(2,22) = 123.2$, $p < .0001$.

Having verified that our Level 1 layers were operating according to our expectations, we then wanted to see if the Top Layer was responding to the components of the events or to their causality. We ran an analysis of variance on the same two event types as above: direct-delay+gap and delay-gap. Without the component exclusivity present in the Level 1 layers, the difference between these two conditions should be the same. However, there was a significantly greater response to the direct-delay+gap change than to the delay–gap change, $F(2,22) = 15.3$, $p < .0001$. This shows a clear preference on the basis of causality rather than just the independent components.

We can see, too, that this difference in processing has a developmental or long-term experiential component. Figure 8a and 8b shows the preference for a component model of causality settling in the two Level 1 layers after about 800 training cycles. Figure 8c shows that the Top Layer does not settle on a causal model until about 1500 cycles. As mentioned earlier, this is because of the nature of CLA: the lower levels must settle before the higher levels can.



Figures 8a, 8b and 8c: The acquisition of a component model vs. a causal model in different layers over time.

Discussion

Our CLA system has created a hierarchical knowledge structure that can produce habituation results compatible with those from similar studies with human infants. These results are consistent with those of Cohen and Amsel (1998) as well as Leslie (1986). Contrary to Leslie's conclusions, however, our model does not rely on a causality module.

That is not to say, of course, that our model has *nothing* built in to it. CLA is not *tabula rasa.* Unlike a modularist view, though, the innate attributes of CLA are domain general information processing principles. More generally, CLA has innate processes, rather than specific innate knowledge of abstract concepts.

CLA also relies on the vast majority of direct events in the world compared to non-direct events. We believe that infants also rely on this arrangement. CLA is guided by the nature of the environment to develop a causal model because there are simply more direct events in the environment. We can imagine an alternate universe which contained more delay events than any other kind, and our model would predict that development in this kind of environment would result in a drastically different world view.

One might ask what the point of having a stratified representation of causality might be, when it might be possible to achieve this same learning with a monolithic system. As previously stated, our hierarchical approach has the effect of producing the stages in development that we see in infants. But more than just fitting the experimental data, a hierarchical representation makes it possible to address the last two information processing principles described above. Cohen and Cashon (2000), and others, have observed hierarchical knowledge processing in infants, both in terms of perceptual preference and in handling cognitive overload. CLA's hierarchical design makes such processing possible, where a monolithic system would not. We intend to use CLA for robotic control, and we feel that principles five and six can be used with CLA's knowledge hierarchy to give certain layers priority over others. Also, we plan to test proposition six by overloading our system and seeing if it produces the "fall back" phenomenon that has been demonstrated in infants.

Conclusion

Although there are several connectionist models of infant development, CLA is the first to use hierarchical representation and differentiate between long-term and short-experience. These are important factors in cognitive development, and are often not given much weight even in *real* infant habituation experiments. The information processing approach to cognitive development has been applied to infant cognition with considerable success. We feel that a computational model which uses this approach holds promise for modeling the acquisition of a variety of domains within infant, child, and even adult cognition.

Acknowledgments

This research was supported in part by NIH grant HD-23397 to the first author from the National Institute of Child Health and Human Development.

We would like to thank Risto Miikkulainen for his advice and guidance at various stages of this project.

References

Chaput, H. H. (2001). Post-Piagetian constructivism for grounded knowledge acquisition. To appear in *Proceedings of the AAAI Spring Symposium on Learning Grounded Representations*, March 2001, Palo Alto, CA.

Cohen, L. B. (1998). An information processing approach to infant perception and cognition. In G. Butterworth and F. Simion (Eds.), *Development of Sensory, Motor, and Cognitive Capacities in Early Infancy: From Sensation to Cognition.* Sussex: Erlbaum (UK) Taylor & Francis.

Cohen, L. B. & Amsel, G. (1998). Precursors to infants' perception of the causality of a simple event. *Infant Behavior and Development 21 (4)*, 713-732.

Cohen, L. B., Amsel, G., Redford, M. A. & Casasola, M. (1998). The development of infant causal perception. In A. Slator (Ed.), *Perceptual development: Visual, auditory and speech perception in infancy.* London: UCL Press (Univ. College London) and Taylor and Francis.

Cohen, L. B. & Cashon, C. H. (2000). Infant object segregation implies information integration. *Journal of Experimental Child Psychology*, (in press).

Cohen, L. B. & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology,* Vol. 29, No. 3, 421-433.

Kohonen, T. (1997). *Self-Organizing Maps*. Berlin: Springer-Verlag.

Leslie, A. M. (1986). Spatiotemporal continuity and the perception of causality in infants. *Perception, 13,* 287-305.

Michotte, A. (1963). *The Perception of Causality.* New York: Basic Books.

Schafer, G. & Mareschal, D. (2001). Modeling infant speech sound discrimination using simple associative networks. *Infancy 2*, 7-28.