Building A Baby

Paul R. Cohen, Tim Oates, Marc S. Atkin
Department of Computer Science
Carole R. Beal
Department of Psychology
University of Massachusetts, Amherst, MA 01003
cohen@cs.umass.edu

Abstract

We show how an agent can acquire conceptual knowledge by sensorimotor interaction with its environment. The method has much in common with the notion of image-schemas, which are central to Mandler's theory of conceptual development. We show that Mandler's approach is feasible in an artificial agent.

Introduction

It is a great mystery how adult concepts can develop from infant sensorimotor activity. The title of this paper borrows from two papers by Jean Mandler in which she discusses how this development might work. Central to her theory is *perceptual analysis by image-schemas* (Mandler, 1988; Mandler, 1992; Johnson, 1987; Lakoff and Johnson, 1980; Lakoff, 1984): "I propose that perceptual analysis results in redescriptions of spatial structure in the form of image-schemas. These redescriptions constitute the meanings that infants use to create concepts of objects, such as animate and inanimate things, and relational concepts, such as containment and support." (Mandler, 1992, p. 587). As far as we know, nobody has demonstrated by means of a running computer program that image-schematic redescriptions can produce conceptual structures from sensorimotor interactions. This is the purpose of the research reported here. A second aim of the research is to show that very little prior structure is sufficient to acquire conceptual structures. This result causes us to think that the "models" of the physical world attributed to infants by Spelke, Carey, Baillargeon, and others, might be learned, not innate.

What are image-schemas? Think of them as pattern detectors or filters that map sensory streams onto redescriptions or partial representations. For example, when we see a cat walk across the room, our ANIMATE MOTION image-schema produces a partial representation of the scene; similarly when we see a person walking down the road. But when we see a car zoom past, the ANIMATE MOTION image-schema stays quiet.

As minimalists we are leery of image-schematic theories of conceptual development because they are unconstrained. How many image-schemas are required? Are they all innate or are some learned? Which aspects of sensory experience should image-schemas analyze? One purpose of this paper is to propose a very few, well-motivated image schemas, and show that an artificial agent can use them to learn a lot about the states, objects and processes in its environment.

We have implemented a simulated agent called Neo that learns representations of objects, states and activities, and is poised to learn categories, by a process of perceptual analysis, using several versions of a single learning rule that has many of the properties claimed for image schemas. As it happens, Neo's analysis is of temporal, not spatial, structure. Even so, we present Neo as evidence that Mandler's theory of infant conceptual development is, in its broad outlines, sufficient for an artificial agent to learn concepts.

Baby World

Neo lives in a simulated environment called BabyWorld, which implements Neo's sensations, mental representations, mental and physical activities, and the behavior of objects and other agents that interact with Neo. BabyWorld has two parts: one, Neo, implements everything that Neo does, including learning, moving, mouthing, looking, crying, and so on. The other part, called StreamsWorld, represents Neo's environment, and it implements events that happen around Neo and in response to Neo's actions.

Neo senses its environment through a collection of *streams*, which are divided into discrete time steps. In each time step a stream holds a *token*. Tokens represent *sensations* or processed percepts. For example, one token is **rattle-shape** and it is placed in the appropriate stream whenever Neo's eyes point at an object that is shaped like a rattle. The streams that represent Neo's internal sensations include an affect stream that contains tokens such as happy and sad, a pain stream, a hunger stream, and somatic and haptic streams that are active when Neo moves and grasps.

The Babyworld simulator is simple and probabilistic. For example, Neo gets hungry some time after eating, it cries when it is unhappy or in pain; when Neo cries, Mommy usually visits, unless she is angry at Neo for crying, in which case she stays away. Neo falls asleep intermittently; it can move its arm and head, and grasp several objects, including three rattles, a bottle, a mobile, keys and a knife. The latter causes pain. The rattles make noise when shaken. Currently, Neo is incapable of anything we would call volition. If Neo's eyes alight on a rattle then Neo will grasp the rattle with some probability. However goal-directed this might appear, Neo's mind contains nothing that could be interpreted as a goal.

Redescription of Sensations

When Neo starts to run its experience is "blooming, buzzing confusion," with no apparent structure. (We know this is probably not true of neonates, but we don't want to assume prior mental models of the physical world if these might be acquired through interaction, as we believe they can. Thus we start with very minimal prior structure.) Neo goes through

five levels of redescription of its experience: 1) Changes in token values: Tokens in streams are augmented by noticing when they change value. 2) Scopes: Neo finds pairs of correlated streams called scopes. 3) Base fluents: Neo finds common token-value pairs within scopes. 4) Context fluents: Neo finds base fluents that tend to follow each other in time. 5) Chains: These temporal dependencies are combined into temporal chains, which represent activities. Chains are used for activity-based categorization.

Each level of redescription produces an intermediate representation. Representations are interned in Neo's memory when they have accrued sufficient evidence from Neo's experience. The examples in this paper are from a single run of Neo, lasting 30,000 time steps, during which it created thousands of intermediate representations. A single counting mechanism is responsible for deciding when to intern a representation. Neo engages in all five levels of redescription as soon as it is able, but because each instance of intermediate representation requires statistical support, instances of deeper levels of representation are often created before instances of shallower levels.

Noticing When Token Values Change

A stream σ_i is said to change state at time t, denoted $\Delta(i,t)$, when $\sigma_{i,t-1} \neq \sigma_{i,t}$; that is, σ_i changes state at time t when it contains a different token at time t than it did at time $t-1$. Conversely, $\overline{\Delta}(i,t)$ means the stream doesn't change state: $\sigma_{i,t-1} = \sigma_{i,t}$. At this level of redescription, we don't care what the token values actually are, we only care whether they change. As it happens, this reduction in information serves to reduce the combinatorial space of representations at deeper levels (i.e., scopes, base fluents, context fluents and chains) and so is an essential first level of redescription.

Scopes

Neo learns a *scope*, s_{ij}, when streams σ_i and σ_j change together often. Said differently, Neo learns s_{ij} when the joint event $\Delta(i,t)$ & $\Delta(j,t)$ occurs frequently relative to the joint events $\Delta(i,t)$ & $\overline{\Delta}(j,t)$ and $\overline{\Delta}(i,t)$ & $\Delta(j,t)$. To assess the relative frequencies of these events, Neo uses contingency tables like this one:

	Δ(sight-color,t)	$\overline{\Delta}$(sight-color,t)	*total*
Δ(sight-shape,t)	2996	945	3941
$\overline{\Delta}$(sight-shape,t)	826	25232	26058
total	3822	26177	29999

This says that the streams **sight-shape** and **sight-color** changed state simultaneously 2996 times, and one changed when the other didn't $945 + 826 = 1771$ times. To assess the strength of association between **sight-shape** and **sight-color** Neo squares the frequency in the first cell of the contingency table (2996) and divides by the product of the first row and first column margins (3941 and 3822, respectively). The maximum value for this statistic is 1.0, and for the table above it is $2996^2/(3941 \times 3822) = .596$. [1]

[1] Neo could use other statistics, such as χ^2 and G, provided the contingency table is scaled to a constant total, preserving the proportions. (Scaling is necessary because χ^2 and G are not independent of sample size.) In practice, Neo learns the same scopes, and ranks them similarly, irrespective of how it measures association in its contingency tables.

Scopes provide a mechanism for cross-modal perception: the same contingency table mechanism will detect cooccurrences in the visual and tactile streams, for example Rose (1990) and Spelke (1987).

Fluents

Fluents represent things that don't change, or that change in highly regular, predictable ways. The sound made by a rattle is a fluent, so is the sensation of holding the rattle, and so are the visual sensations of the shape and color of the rattle. Of course, the concept "rattle" has all these components, so fluents for the color, shape, sound and texture of a rattle must be linked up in a single fluent.

Although the simplest fluents represent sensations, fluents are not identical with sensations. This is an important point, because the distinction between streams and fluents is how we implement the distinction between sensory experience and cognitive experience. Neo can experience the sensations associated with looking at a red rattle without saying, mentally, "Ah, I recognize a red rattle." The sensory experience is implemented as tokens for red and rattle-shaped in the appropriate streams, whereas the cognitive experience of a red rattle involves activating a fluent that represents the red rattle.

Base Fluents

Neo's smallest fluents, called *base fluents*, represent cooccurring tokens within scopes. Suppose stream σ_i contains a at time $t-1$ and b at time t. Then we say token a *stops* in stream i at time $t-1$, denoted ⊣(i,a,t-1), and token b *starts* in stream i at time t, denoted ⊢(i,b,t). Now suppose Neo turns its head and its eyes alight on a red rattle. Neo will detect two simultaneous events, ⊢(sight-color,red,t) and ⊢(sight-shape,rattle-shape,t). Sometime later, Neo might look somewhere else, which will generate two simultaneous stop events, ⊣(sight-color,red,v) and ⊣(sight-shape, rattle-shape,v). Simultaneous start events and stop events are evidence that a single object—in this case a red rattle—or a single activity, is making its presence felt in two streams. Of course, two *unrelated* events could occur simultaneously in two streams, but this sort of coincidence is less likely than the coincidence of related events. Contingency tables like the one described earlier count the cooccurrences of start and stop events, and assess whether start and stop events happen simultaneously significantly often. Significant associations create base fluents.

Some of the base fluents discovered by Neo are shown in table 1. They make sense, given what we know about the Neo simulator. The first block of fluents in table 1 deals with mouthing: When Neo is mouthing, its arm is resting (**(mouth mouthing) (arm resting)**) and its voice is quiet (**(mouth mouthing) (voice quiet)**). When it isn't mouthing, Neo can make noises—crying, gurgling and screaming—and its arm can move.

The next block of base fluents in table 1 represents objects in Neo's environment, including the green rattle, the green mobile, the metallic keys, the knife, and so on. Not all the objects have been learned because Neo ran for only 30,000 time steps. The last fluent in this block (**(sight-color dark) (sight-shape none)**) represents what happens when Neo closes its eyes.

It is apparent to us, though not to Neo, that base fluents collectively have structure. Note that the block of ((**sound**

((mouth mouthing) (arm resting))	((mouth mouthing) (voice quiet))
((mouth not-mouthing) (arm move-lf))	((mouth not-mouthing) (voice cry))
((mouth not-mouthing) (voice gurgle))	((mouth not-mouthing) (voice scream))
((sight-color green) (sight-shape rattle-like))	((sight-color green) (sight-shape mobile-like))
((sight-color metallic) (sight-shape blob-like))	((sight-color metallic) (sight-shape knife-like))
((sight-color orange) (sight-shape blob-like))	((sight-color orange) (sight-shape rattle-like))
((sight-color red) (sight-shape rattle-like))	((sight-color white) (sight-shape rattle-like))
((sight-color white) (sight-shape crib-like))	((sight-color dark) (sight-shape none))
((sound cry) (mouth not-mouthing))	((sound cry) (tactile-mouth none))
((sound cry) (voice cry))	
((sound gurgle) (mouth not-mouthing))	((sound gurgle) (tactile-mouth none))
((sound gurgle) (voice gurgle))	
((sound quiet) (arm resting))	((sound quiet) (arm-speed resting))
((sound quiet) (mouth mouthing))	((sound quiet) (tactile-mouth skin))
((sound quiet) (tactile-mouth plastic))	((sound quiet) (voice quiet))
((tactile-hand none) (hand open))	((tactile-hand plastic) (hand close))
((tactile-hand wood) (hand close))	

Table 1: **Some of Neo's Base Fluents**

cry)...) base fluents has exactly the same structure as the block of **((sound gurgle)...)** fluents: Neo is not mouthing, it has no tactile sensations in its mouth, and it is doing something with its voice. Regularities like this are the basis for categorization, as we describe below. The **((tactile hand)...)** fluents illustrate similar regularities.

Context Fluents

Suppose Neo is holding a rattle, and then it starts to mouth the rattle. While it is holding the rattle, the fluent **((tactile-hand wood)(hand close))** is active, and when it starts mouthing, the fluent **((tactile-mouth wood)(do-mouth mouth))** will become active. The latter fluent starts in the context of the former. If this happens significantly often then Neo will form the context fluent,

(CONTEXT ((tactile-hand wood)(hand close))
((tactile-mouth wood)(do-mouth mouth))) .

The contingency table mechanism that learns scopes and base fluents also learns context fluents. Specifically, when fluent F_2 starts at time $t + i$, Neo checks to see whether fluent F_1 is active, and if so, it updates the first cell of the contingency table, ($\vdash F_1, t$ & $\vdash F_2, t + i$). If F_2 starts and F_1 isn't active, then Neo updates the third cell of the table, ($\overline{\vdash F_1, t}$ & $\vdash F_2, t + i$). If F_1 is active but F_2 doesn't start within a window of i time steps, then Neo increments the second cell of the table, ($\vdash F_1, t$ & $\overline{\vdash F_2, t + i}$).

What's missing from this account is what it means for a fluent to be "active." In fact, we finessed this problem earlier, when we described how Neo learns base fluents, implying that every start and stop event within a scope is "active," that is, contributes to the contingency table for some base fluent. A more psychologically plausible mechanism might include some sort of selective attention, so not all scopes are monitored for start and stop events all the time. The problem of attention is even clearer when we contemplate building context fluents from other fluents, because fluents are representations in memory. The start event $\vdash F_1$ really means, "something happens in the streams, and as a result, the fluent F_1 is retrieved from memory." We haven't explained exactly how the event $\vdash F_1$ is implemented. It is really too simple: When all the token values in a scope change simultaneously, Neo compares the new values to the base fluents it has learned, and if it finds a match, it "activates" the associated base fluent. It activates a context fluent whenever one of its component fluents is activated. As soon as a fluent is activated, its "level of activation" begins to decline, and after a period of time it becomes inactive even if the sensory events that activated it are still present. This is how we implement a crude form of habituation. We are unable to model complex patterns of habituation and dishabituation. Neo's attentional mechanism is the focus of ongoing work.

Some of Neo's context fluents are illustrated in table 2. The first block of fluents begins with the fluent **((sight-movement resting) (arm resting))**. In this context, Neo very often observes the start of the fluents **((do-hand close) (hand close))**, **((tactile-hand plastic) (hand close))**, and **((tactile-hand wood) (hand close))**. That is, Neo has learned three activities that begin when its arm is resting. The first is, "experience the intention to close the hand and the sensation of the closed hand"; the second and third are, "experience the tactile sensation of wood/plastic in the hand and the sensation of the hand closed." The next two context fluents in table 2 have the same endpoints—the hand closing, and tactile sensations in the hand—but they begin with Neo crying and not mouthing. The final two context fluents begin with the tactile sensation in the hand and end with quiet sound and a tactile sensation in the mouth, and with the tactile sensation and mouthing, respectively.

```
(CONTEXT    ((sight-movement resting) (arm resting))
            ((do-hand close) (hand close)))
(CONTEXT    ((sight-movement resting) (arm resting))
            ((tactile-hand plastic) (hand close)))
(CONTEXT    ((sight-movement resting) (arm resting))
            ((tactile-hand wood) (hand close)))

(CONTEXT    ((sound cry) (mouth not-mouthing))
            ((do-hand close) (hand close)))
(CONTEXT    ((sound cry) (mouth not-mouthing))
            ((tactile-hand plastic) (hand close)))

(CONTEXT    ((tactile-hand plastic) (hand close))
            ((sound quiet) (tactile-mouth plastic)))
(CONTEXT    ((tactile-hand plastic) (hand close))
            ((tactile-mouth plastic) (mouth mouthing)))
```

Table 2: **Some of Neo's Context Fluents**

Chains and Classification

Neo aggregates context fluents into *chains*. For example, given the context fluents,

```
(CONTEXT    ((tactile-mouth none) (voice cry))
            ((tactile-hand wood) (hand close))
(CONTEXT    ((tactile-hand wood)(hand close))
            ((tactile-mouth wood)(do-mouth mouth))),
```

Neo forms the chain,

```
(CHAIN      ((tactile-mouth none) (voice cry))
            ((tactile-hand wood) (hand close))
            ((tactile-mouth wood)(do-mouth mouth)) ) .
```

Here is another, very similar chain that Neo learned:

```
(CHAIN      ((tactile-mouth none) (voice cry))
            ((tactile-hand plastic) (hand close))
            ((tactile-mouth plastic)(do-mouth mouth)) ) .
```

The only difference between these chains is the object that Neo grabs and mouths: in the first case it is wooden, in the second, plastic. We may form a *class* of things that Neo can grab and mouth. The chains don't say exactly which objects are in the class, but we know they are either wood or plastic, and they are graspable, and mouthable.

GRASPABLE and MOUTHABLE are *interactional* properties (Johnson, 1987) that characterize Neo's activities in its environment. Unlike TEXTURE—wood or plastic—they are in a sense *subjective*: What's graspable by one agent isn't necessarily graspable by another. Whereas TEXTURE is an inherent property of an object, GRASPABLE is a property of the object *and the agent* who may try to grasp it. Interactional properties like GRASPABLE are the basis for categories in Lakoff and Johnson's theory of categorization (Johnson, 1987; Lakoff, 1984; Lakoff and Johnson, 1980) and also in Mandler's theory of conceptual development (Mandler, 1992). However, we believe categories are best defined in terms of *activities*, and the attractiveness of interactional features is due to them describing activities better than objective features such as texture (Cohen, Oates and Atkin, 1996).

In fact, although Neo learns activities, represented as chains, we are responsible for using these chains to identify features and form classes (Cohen, Oates and Atkin, 1996). The first step is to match up chains that have the same stream names in the same order, creating an *abstract chain* of scopes. For example, the chains above are both described by the abstract chain **(tactile-mouth voice)** → **(tactile-hand hand)** → **(tactile-mouth mouth)**. Now when we look at the token values that can instantiate this abstract chain, we find that the **tactile-hand** and **tactile-mouth** streams contains either **none**, **wood** or **plastic**. In other words, the abstract chain identifies an activity in which Neo has nothing in its mouth and is crying, and then has something wood or plastic in its hand and its mouth. We know, because we built the Neo simulator, that the wood or plastic objects include Neo's rattles and bottles, but not the mobile, Mommy, or Neo's own hand. Currently, this class is "implicit." Neo doesn't have an ontology in its head, nor declarative definitions of categories. Still, there is an implicit class of objects that can participate in the abstract chain. (See Cohen, Oates and Atkin, 1996, for further examples.)

Discussion

We have illustrated five levels of redescription of Neo's sensory experience, and suggested how the regularities in these redescriptions can be the basis of classification. Each level of redescription provides the opportunity to learn representations, and one simple mechanism produces all these representations but chains. The mechanism maintains contingency tables for pairs of start (⊢) or stop (⊣) events, and when a measure of association for the table achieves significance, Neo interns a representation. The only thing that changes, from one level of representation to the next, is *what* starts and stops, and whether a lag is allowed between these events.

Each version of this basic contingency-table mechanism can be viewed as an image-schema, in the sense Mandler (1992) intends: It produces a redescription and an intermediate level representation of raw sensory experience. It happens that all Neo's redescription takes place in the temporal domain, but this is appropriate for an agent that is biased to learn a predictive model of events in its environment. Although Mandler presents image-schemas as processors of spatial information, they are equally well described in temporal terms. Mandler cites, for instance, SELF MOTION, ANIMATE MOTION, CAUSED MOTION, and AGENCY as image schemas. We would argue that Neo has learned some of these predicates. For example, the abstract chain **(do-arm arm)** → **(sight-movement arm-speed)** defines a class of actions in which, in the context of an arm movement, Neo sees the arm moving fast. Arguably, this is a SELF MOTION image-schema.

This example raises the question of how many image-schemas is a baby born with, and whether more are learned. Mandler lists many image schemas in her paper; we suggest one for scopes, one for base fluents, and one for context fluents (applied recursively to produce chains). Another mechanism is required to produce abstract chains. Still, we take a distinctly minimalist position: If an image-schema such as SELF MOTION can be learned as described above, we prefer not to assume babies are born with it. We think it is very valuable to implement agents such as Neo to find out how much or little is required in the way of innate structure, especially as image-schemas and the kinds of "models" discussed by Leslie (1988), Spelke (1988), Carey and Spelke (1994), Keil (1994),

and others are informal (lacking interpretation as data structures and processes) and their interactions with memory and attention are largely unspecified.

A related question is why particular relationships are image-schemas. We offer two kinds of explanation. Scopes serve to reduce the combinatorial space of potential fluents, so perhaps some image-schemas have evolved for computational reasons. Base fluents are learned when tokens start and stop simultaneously. Simultaneity is rare among independent events, so an image-schema that detects simultaneity is ideal for associating parts of a whole. (Thus it may not be necessary to posit an innate and sophisticated understanding of the physical world, e.g., Spelke, 1988.) Context fluents are learned when one fluent follows another more often than would be expected by chance, which is a necessary though not sufficient condition to infer cause (Suppes, 1970; Cohen, 1995). So Neo has the image-schemas it has because they help Neo identify states, objects and potentially causal sequences.

Conclusion

Let us review what Neo learned: It learned that most of the regularity in its environment takes place in 30 pairs of streams, less than 10% of the $(26 \times 25)/2 = 325$ pairs of streams that it might have focused on. It learned base fluents corresponding to the shape and color of most objects in its environment. It learned the permanent locations of the green mobile (directly overhead) and the crib bars (to the extreme left and right of its field of view). It learned activities, such as grasping an object and mouthing it, or moving its arm and seeing its arm move. It almost learned conditions. For example, it learned a chain that includes **...((do-hand open)(hand open))((tactile-mouth skin)(mouth mouthing))**, but it has no way to learn that the first fluent is a condition for the second—that the hand *must* be open to be mouthed. It learned chains from which we abstracted classes that make sense in Neo's environment, such as the class of objects that can be grasped and mouthed, and the class of activities that end in seeing the arm moving fast.

Keep in mind that Neo's actions are largely random: when it grabs an object it *can* mouth it, but it's just as likely to drop it, or move its head. The only structure in Neo's actions is provided by conditions (e.g., it cannot mouth an object it hasn't grasped, and it cannot mouth its hand unless the hand is open) and by a handful of simple behavioral dependencies built into the simulator (e.g., it sometimes grabs what it looks at, and it cries if it gets hungry). Keeping in mind also that Neo ran for only 30,000 time steps, it seems to us that it learned quite a lot.

In conclusion, Neo provides preliminary evidence that image-schematic redescription of raw sensations is probably sufficient to form implicit, activity-based categories. Very few image-schemas are required; more may be learned.

References

Susan Carey and Elizabeth Spelke. (1994). Domain-specific knowledge and conceptual change. In Lawrence A. Hirschfeld and Susan A.Gelman, editors, *Mapping the Mind*. Cambridge University Press.

Paul R. Cohen. (1995) *Empirical Methods for Artificial Intelligence*. MIT Press.

Paul R. Cohen, Tim Oates, and Marc S. Atkin. (1996). Preliminary evidence that conceptual structure can be learned by interacting with an environment.

Mark Johnson. (1987). *The Body in the Mind*. University of Chicago Press.

Frank Keil. (1994). The birth and nurturance of concepts by domains. In Lawrence A. Hirschfeld and Susan A.Gelman, editors, *Mapping the Mind*. Cambridge University Press.

George Lakoff. (1984). *Women, Fire, and Dangerous Things*. University of Chicago Press.

George Lakoff and Mark Johnson. (1980). *Metaphors We Live By*. University of Chicago Press.

Alan M. Leslie. (1988). The necessity of illusion: perception and thought in infancy. In L. Weiskrantz, editor, *Thought Without Language*. Oxford University Press (Clarendon).

Jean M. Mandler. (1988). How to build a baby: On the development of an accessible representational system. *Cognitive Development*, 3:113–136.

Jean M. Mandler. (1992). How to build a baby: II. conceptual primitives. *Psychological Review*, 99(4):587–604.

S. A. Rose. (1990). Cross-modal transfer in infants: What is being transferred? In Adele Diamond, editor, *The Development and Neural Bases of Higher Cognitive Functions*. New York Academy of Sciences.

Elizabeth S. Spelke. (1987). The development of intermodal perception. In P. Salapatek and L. Cohen, editors, *The Handbook of Infant Perception*. Academic Press.

Elizabeth S. Spelke. (1988). The origins of physical knowledge. In L. Weiskrantz, editor, *Thought Without Language*. Oxford University Press (Clarendon).

Patrick Suppes. (1970). *A Probabilistic Theory of Causality*. North-Holland Amsterdam.