Growing Ontologies

Paul R. Cohen
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
cohen@cs.umass.edu
Tracking number: A378
Content areas: Philosophical foundations, ontologies, cognitive modeling, commonsense, unsupervised learning
Submitted to AAAI 98

Abstract

Conceptual structures or ontologies are usually built by hand by skilled knowledge engineers. This paper presents a theory of how conceptual structure may be acquired by an intelligent agent interacting with its environment in an unsupervised way. Categories of activities are learned, then abstractions over these categories result in concepts. The entire conceptual structure is based on activities. The meanings of concepts and of conceptualizations of activities are discussed. Systems that implement aspects of the theory are presented, and their general characteristics described.

Introduction

The subject of this paper is the foundation of conceptual systems. In artificial intelligence, conceptual systems are sometimes called ontologies and they are usually built by highly-trained knowledge engineers. It's less clear how humans acquire conceptual systems, or rather, the scope of the innate endowment is unclear. In any case, if machines could acquire conceptual knowledge with the same facility as humans, then AI would be much better off. Many people think that our conceptual system is a prerequisite for all sorts of intelligent processes, from analogical reasoning to natural language understanding, from reasoning under uncertainty to computer-assisted cooperative work. Lenat and Feigenbaum (1987) promised us that a sufficiently large corpus of common sense knowledge would "go critical" and start learning, by reading, autonomously. That hasn't happened yet, but there's no denying the dream of a machine that knows roughly what we know, organized roughly as we organize it, with roughly the same values and motives as we have.

It makes sense, then, to ask how this knowledge is acquired by humans and how might it be acquired by machines. In particular and for various reasons, I want to focus on the origins of conceptual knowledge, the earliest distinctions and classes, the first efforts to carve the world at its joints. One reason is just the scientist's pleasure at getting to the bottom of, or in this case the beginning of, anything. Another reason is that the origin of conceptual systems is currently hotly debated: Some people think that neonates are born with moderately sophisticated conceptual systems (e.g., Baillargeon, 1994; Carey and Gelman, 1991; Spelke et al. 1992), others dispute this and seek an empiricist, or non-nativist, account of development (e.g., Mandler, 1988, 1992). I think one has to take a minimalist stance and avoid innate knowledge in one's explanations of the acquisition of later knowledge. Partly this reluctance comes from years of slogging in AI, where it seems we must always provide a lot of knowledge for our systems to do relatively little with. So I want to focus on the first concepts because unless I do, then I'll have to provide them by hand, which is a bore, and also makes me very uncertain about the explanatory power of what follows.

To elaborate this last point, it is a commonplace in AI that everything depends on representation---get the right set of attributes, or the right representation of a game board, or the right set of clinical terms, and problem solving is relatively easy, usually just search of some kind. But surely this means that the hard problem solving was done by us. If instead we ask how does the machine come to conceptualize the world this way rather than that, if we ask the machine to form its own conceptual system, then we don't have to do the hard work and our claims to understanding intelligence are that much stronger.

The principal claim of this paper is that concepts can be learned without supervision by abstracting over representations of activities. Piaget (1952) insisted that concepts must arise out of activities — because simple action schemas develop before conceptual thought — but the mechanisms he proposed to explain concept acquisition were vague by the standards of artificial intelligence. Some implemented mechanisms are presented later in the paper. Between here and there I discuss some questions raised by the claim that all concepts — even those that represent static objects — are acquired from representations of activity: What are concepts; how do they come to have meaning; why should agents learn them rather than get them ready-made from programmers; how are they learned; and what do these learning mechanisms have in common?

Concepts, Conceptualizations and Meaning

Definitions of *concept* differ but I believe all have in common that concepts are abstractions. Often, concepts are taken to be abstractions over features or attributes of objects; this is certainly the dominant view in machine

learning. I want to use the mechanisms of machine learning but I want to focus them not on the attributes of objects but on the dynamics of activities. Thus I define concepts as abstract representations of activities including the participants in and the entailments of the activities. Figure 1 shows sketches of two concepts. Both begin with one agent, A, running toward another, B. One of these interactions develops into an embrace, the other into a chase. The entailments of these concepts are different, too: the rush-to-embrace concept implies that A and B are happy whereas the frighten-and-chase concept implies that B is unhappy.

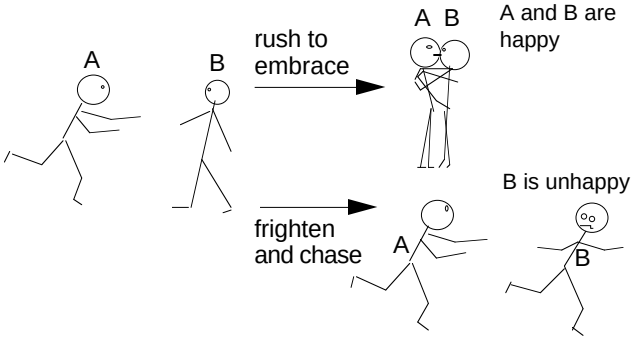

Figure 1. Illustrations of concepts for two activities, rush-to-embrace and frighten-and-chase. The participants or roles are identified (A and B) as are some entailments (e.g., A and B are happy).

Conceptualizations are instantiated concepts, formed by binding the entities in an activity to the roles in a concept. Suppose you are in an arrival lounge at the airport and you see a child running at an adult getting off the airplane. By binding the child to A and the adult to B in the rush-to-embrace concept, one conceptualizes the activity. Were the same child to run at another, younger child, the conceptualization and its entailments would be different: Interactions that are conceptualized as frighten-and-chase often end in tears.

Activities mean nothing in and of themselves. Once conceptualized, the meanings of activities are given by the entailments of the concept. We may infer that the child is happy to see the adult getting off the airplane because we have conceptualized his activity as a rush-to-embrace. (I would rather give these concepts gensyms for names, to emphasize that their meanings are entirely in their entailments, but I will keep the descriptive names for simplicity of exposition.)

This is just one of several possible accounts of meaning. Other popular accounts include: Meaning is a formal relationship between an assertion about the world and the state of the world; the meaning of a concept is given by its relationships to other concepts; and meaning is given by reducing an assertion to semantic primitives. Because I want to build agents that learn concepts by interacting with the world, I am most comfortable with the idea that the meaning of a conceptualization of an interaction is the inferences one can make about the interaction —the entailments of the interaction — given the conceptualization. (I will discuss the meaning of concepts, as opposed to conceptualizations, shortly.) Among the entailments are predictions about how an interaction will unfold. If you conceptualize reading an article (not this one, I hope) as a chore, then you can anticipate discomfort, tedium, and a growing desire for milk and cookies or something stronger. This is what it means for reading an article to be a chore. Another conceptualization — say reading-as-intellectual-exploration — makes different predictions.

If activities are meaningless until they are conceptualized, are concepts meaningless until their roles are bound to entities in activities? That is, are concepts meaningless until they are grounded in actual experience? Apparently not, for I am trusting that you and I understand the meaning of rushing-to-embrace without actually observing or participating in the activity. I am trusting that your concept rushing-to-embrace has more or less the same roles and entailments as mine. Perhaps I am counting on you to imagine (i.e., conceptualize) an interaction, but I think we can both contemplate the concept without instantiating it physically or even mentally through imagination.

Doubtless some conceptualizations *feel* more meaningful than others — contrast embracing someone, observing an embrace, imagining yourself embracing, watching an embrace on television, reading about it in a novel with good character development, and reading about it here. These conceptualizations feel different partly because some are richer than others in terms of the roles they specify and partly because of the different affective responses to doing something, reading about it, watching it, or imagining it. Should the meanings of concepts and conceptualizations include affective responses? In my current formulation, concepts and conceptualizations may *describe* affect in their entailments (e.g., A and B are happy), but affect *itself* is not part of their meanings. This seems forced: The meaning of embracing my wife resides in the statement that we are happy, not the happiness itself! Eventually I hope to offer a less disembodied account, in which the meaning of conceptualizations (if not concepts) includes not only inferences but also affective responses.

Concepts may refer to other concepts in their roles. For example, the frighten-and-chase concept may specify that entity A is a male-toddler-with-a-discipline-problem. Suppose you see a male toddler thrashing and screaming at the supermarket. You conceptualize this activity as a tantrum and you conceptualize the boy as a male-toddler-with-a-discipline-problem. What is the meaning of this concept? As before, meaning comes from entailments, which in this case come from the activities in which the concept participates. Thus, the meaning of male-toddler-with-a-discipline-problem is that he can have tantrums and

he can chase and frighten other kids.

Why Should Agents Learn Concepts?

What are concepts and ontologies for and why should they be learned rather than constructed by knowlege engineers? Imagine yourself to be a mobile robot wandering around the lab. What concepts do you need, and what will you do with them? Presumably you have some reactive routines to keep you out of trouble. These recognize aspects of your internal state and the world state, and generate actions in response; for example, when you detect an obstacle in your path, you change direction or stop. Suppose you have a function `obstacle-in-path` that takes as input some sensory information such as sonar readings. As described, the concept "obstacle" simply doesn't exist for you, the robot. It exists for the person who wrote the function `obstacle-in-path`, but not for you. You do not know that obstacles may sometimes be pushed aside, although you may have a function to push obstacles aside; you do not know that obstacles impede paths, although obstacles impede your paths; you do not even know what a path is, although you trace one whenever you move. You have no conceptual structure whatsoever, just a bunch of routines to keep out of trouble. And why *should* you know anything, if these routines work? Why do you need concepts like "obstacle" and "impede" and "path" at all?

The function `obstacle-in-path` was written by a person who thought about all the situations you might encounter and realized that some of them involve an "obstacle" in your "path." Because this person conceptualized your experiences, you have an appropriate response to a common situation. Your behavior is organized around your programmer's conceptual structure. This structure — of which you are innocent — serves you well. This is what concepts are for: They give you interpretations of your sensors, a basis for judgments that situations are similar, and the distinctions on which you decide what to do. Concepts are necessary even when they are not explicit. We must stop deluding ourselves that a robot or any other agent is capable of intelligent behavior without a conceptual structure. If `obstacle-in-path` produces intelligent behavior, the intelligence must be attributed to the programmer who dreamed up the concepts "obstacle" and "path." It's delusional to say, "Our robot achieved so much with nothing more than a handful of reactive routines"; the reactive routines are just the business end of an entirely hidden but sophisticated conceptual system.

Categories and Concepts

To this point I have argued that repertoires of even very simple activities are based on conceptual systems. Now I will argue that conceptual systems can arise from activities. This is not circular: Concepts and activities bootstrap each other. So how do we get concepts out of activity? An agent does things, like wave its arms, or traverse internet links, or manipulate blocks, or chew on a frog, and somehow concepts emerge. How does this happen?

An intermediate step in the acquisition of concepts may be the collation of *categories*. A category is a collection of experiences that have something in common. For instance, the child in Figure 2 spends a lot of time reaching, grasping and mouthing objects, so she may form a category of reach-grasp-mouth activities and also a category of objects that are reached for, grasped, and mouthed. Following Rosch (1975) concepts are "typical" category members, although a concept may match no category member exactly. Moreover, a concept has entailments, whereas category members do not (reference removed for blind review).

Figure 2. An infant grasps and mouths a frog

We have to be careful about how concepts are abstracted from categories. One account goes something like this: Concepts are lists of necessary and sufficient, objective conditions for something to be an instance of a concept, categories are just the extensions of concepts —the things that satisfy the concept definition —and the meanings of a concept is just a list of other concepts to which this one is related by relationships such as ISA and PART-OF. Indeed, this view of how to get concepts from categories is still dominant in supervised machine learning, where training instances are categorized —given a class label — a priori, and the task is to find necessary and sufficient conditions for instances to be members of classes.

Interactionism

There are many problems with this view. As Lakoff (1984) points out, human categories are not well described by lists of necessary and sufficient conditions. Perhaps a simple example will illustrate the problem: There are three toys in Figure 2 but one would be hard pressed to define "toy" based on their attributes. The frog is shiny, the

others are furry. All are quadrapeds, and all have eyes, but these aren't necessary or sufficient conditions for being a toy. Most cogently, the fact that we characterize these things as toys doesn't mean the infant does. In fact, she interacts with these objects quite differently. She always chews on the frog and very rarely chews on the others. If the infant formed categories based on how she interacted with objects, then the frog would belong to a different category than the other objects.

This of course is the liberating insight of Lakoff (1984) and Johnson (1987). In their *interactionist* view, category distinctions are based on activity, so the frog belongs to a category of things to grasp and chew, whereas the furry toys belong to a category of things to wave about and rub on one's face. The fact that we consider the frog a toy, and a spoon a utensil doesn't matter to *this* baby, who for the longest time considered the spoon to be just another thing to grasp and mouth. On the interactionist account, only when she *uses* the spoon to eat food (Fig. 3) will she differentiate it from the frog, and only then will she form the category that we adults call "utensil."

Figure 3. An infant tries to feed herself

In the interactionist view, concept acquisition is unsupervised: Categories are not defined exogenously — as so often happens with machine learning — but by activities. Concepts are abstractions of activities, their participants and their entailments.

Left unexplained, however, is exactly *how* categories of activities are extracted from ongoing streams of actions. The problem for the infant child or intelligent agent has an adult analog in basketball. I watched the game for months before I was able to recognize the pick-and-roll, the give-and-go, the box-out, and so on. And my learning was supervised in no small part by the accompanying commentary. Imagine trying to learn these categories of activities *without* the commentary. The infant has an easier job, perhaps, because she is often a participant in activities, and they aren't as complex. Even so, she must find recurrent patterns in complex streams of actions without supervision. Formulated this way, however, one can see how she might succeed with any number of simple techniques for finding patterns in time series. The following section will describe some implemented mechanisms.

From Activities to Concepts

Here is the problem to be solved: An agent such as a robot is "born" with a small set of physical activities but no conceptual system. As it interacts with its environment, it forms categories, and by induction over category instances, concepts. How does this work? What innate structures are required? This section presents some implemented methods that solve parts of the problem and collectively might solve the whole problem once they are integrated into a single system. Following the synopses of these methods, I will describe in general terms what they have in common.

Categories of Activities: Baby and the Roles Problem

The first challenge for the infant child or agent is to form categories of activities. Keep in mind that these categories are being learned while the activities themselves are being learned. One formulation of the problem characterizes experience as a trajectory through a state space of very high dimension, where each dimension is a sensor such as a sonar or a strain gauge. Given this formulation, I thought, naively, that the problem of finding categories of activities becomes the problem of finding recurring time series of sensations. This idea is easily illustrated in the Baby system (reference removed for blind review).

Baby is an agent that experiences a simulated world through 26 sensor *streams*, including those encoding the color and shape of what it is looking at, the position of its hand, the angle of regard, hunger, boredom, and so on. Baby learns *fluents*, which are regions in the streams that don't change, or that change in regular ways. Fluents can be of any length. Baby attends to the beginnings and ends of fluents, and not to the interval in between, so it can find long patterns in the streams without exponential search.

Baby learns fluents incrementally, further reducing its computational burden. First it finds all pairs of streams for which a change in one predicted a change in the other. Then it searches in these pairs of streams for specific token values that predict each other (e.g., turning its head to its limiting angle predicts white in the sight-color stream because white is the color of the crib bars, which are seen when Baby turns its head to its full extent.) Then Baby learns longer chains of these predictive rules. Here are two:

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

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

The chains are obviously very similar: Each represents an activity in which Baby has nothing in its mouth and is crying, then it has something in its hand (wooden or plastic) and its hand is closed, then it has something (wooden or plastic) in its mouth and it is mouthing. These fluents would appear to be the sort of thing I called members of a category of activity — representations of activity that are very similar. Following the presciptions of earlier sections, the next step would be to form a category prototype,

```
(CHAIN
 ((tactile-mouth none) (voice cry))
 ((tactile-hand *) (hand close))
 ((tactile-mouth *)(do-mouth mouth))),
```

and then a concept, which might look something like this:

```
(concept: self comfort
  activity:
    ((tactile-mouth none) (voice cry))
    ((tactile-hand *) (hand close))
    ((tactile-mouth *)(do-mouth mouth)))
  entailments:
    ((tactile-mouth none) (voice cry)) -->
      ((tactile-hand *) (hand close))
    ((tactile-hand *) (hand close)) -->
      ((tactile-mouth *)(do-mouth mouth)) ... )
```

Unfortunately, chains do not have enough structure to solve what I call the *roles problem*. Roughly speaking, the roles problem is to extract from a representation of activity "who did what to whom." The people who built Baby (reference removed for blind review) report that the sensation of an object in the hand (wood or plastic) is caused by the same object that causes a sensation in the mouth, and this object is the same one Baby is mouthing. But the chain above doesn't say these things. It describes a common sequence of sensations, not a causal story involving Baby, crying, and grasping and mouthing an object. The chain representation contains too little information to recover the roles played by agents, objects and sensations in activities. For instance, it cannot assert preconditions, such as "one must be holding something in order to mouth it."

One manifestation of this problem is that Baby learned a lot of "junk" fluents in addition to those above. These are statistically significant associations of sensations that correspond to no coherent causal story. For example, one component of a chain would describe a tactile sensation, another would note that the lights are off, and a third would report the angle of the head. While these sensations are indubitably associated in a significant proportion of Baby's activities, the chain doesn't describe those activities.

The roles problem is typically finessed by providing innate roles or an activity template that biases what agents learn (e.g., Drescher, 1991). As a minimalist I am reluctant to do this. I don't know of any work in AI on the roles problem, yet it gets in the way of all our attempts to make an agent learn concepts through interaction, and I suspect it will prove a stumbling block to anyone who tries.

In sum, Baby could find recurrent patterns of sensations without supervision, but one cannot claim that it found *activities* in streams, because the patterns it found had none of the causal structure — the participants, what they did, and what happened as a result — that we expect of representations of activities. This is unfortunate, because Baby is an extremely simple, robust, incremental algorithm that produces easily generalized fluents (as shown earlier). If these fluents specified roles, then Baby would be a good candidate to learn categories of activities, and then concepts.

From Categories to Concepts

Here I will describe several examples of learning concepts given categories of activities, and two examples of learning both categories and concepts, albeit finessing the roles problem.

One study (reference removed for blind review) describes a simulation of two agents, each of which adopts one of nine behaviors, including crash, avoid, kiss, and so on. For instance, A might try to *avoid* B while B tries to *crash* A. An *interaction map* was learned for each of the 81 pairs of behaviors. Interaction maps have two axes, one representing the distance between A and B, the other representing the derivative of distance. An interaction between A and B thus traces a trajectory through an interaction map. Each point in an interaction map indexes a probability distribution. In particular, from each point in a map, the agents have different probabilities of making contact, escaping, and engaging in a perpetual chase. One map was learned for each of the 81 pairs of behaviors. These maps are concepts according to my earlier definition: They are abstractions of activities, they have entailments (i.e., they make predictions about outcomes of interactions) and they identify roles. The roles information is provided a priori in the dimensions of the map, which implicitly identify two participants in the metric "distance from A to B." Interaction maps are learned in a supervised manner: For each map, thousands of interactions of that type were simulated, and the probabilities of outcomes at each point in the map were learned. Thus, the learning algorithm already knew the category of a behavior; it just had to learn the corresponding concept, which it did handily.

The same authors (reference removed for blind review) developed an unsupervised version, where the system *clusters* training trajectories without knowing which behaviors generated them. With very little training, the system comes up with six clusters. Three represent types of interaction where A escapes B. In the first kind of escape, B never gets close to A. In the second, B nearly reaches A, but A slips away. In the third, B's momentum causes it to overshoot A, which escapes. The fourth cluster represents cases where B catches A. The fifth and sixth clusters represent versions of perpetual chasing. Each cluster corresponds to a category of activities, as I defined the term earlier. The categories yield concepts by simply averaging the trajectories within a cluster. The six resulting average trajectories have the properties ascribed to concepts. They are abstractions of activities, they support inferences about the activities (i.e., whether A and B will crash, escape one another, or enter a perpetual chase) and they specify roles, albeit in the same unsatisfactory way I described earlier.

What is remarkable about these results is that the system was not instructed to cluster trajectories by their outcomes. But time series representations of activities are so redundant that clustering by dynamics produces clusters that have qualitatively different outcomes.

This lesson is repeated in two similar projects. Coelho and Grupen (1997) showed how the dynamics of grasping objects are sufficient to identify classes of objects. Very roughly speaking, a robot hand attempts to grasp a prismatic object by activating controllers that run to convergence, attempting to minimize force residuals and wrench residuals. The net effect of running these controllers is that the fingers of the hand migrate over the surface of the object until they achieve a stable grasp. These migrations describe trajectories in the two-dimensional error space. The union of all trajectories for a given object is called a preimage, and the shape of the preimage depends on the object geometry. Just as Rosenstein et al. learned dynamical map representations of the interactions of two agents, Coelho and Grupen learned maps of the interactions between fingers and prismatic blocks. The learning is supervised — the algorithm knows what kind of block it is trying to pick up — but it clearly produces concepts. In fact, when the hand conceptualizes an interaction with an unidentified block as, say, grasping a hexagonal prism, it can predict instabilities in the grasp (i.e., entailments) and modify its grasp. The robot can identify objects by feel as it attempts to gain a stable grasp around them. This is a very clear example of a conceptual activity — classifying objects —that arises out of a purely sensorimotor activity.

Elman (1995) learned grammatical categories in a roughly similar way. He trained a recurrent neural network to predict the next word in sentences, then clustered words by their hidden unit representations. Elman argues that because the network was recurrent, the hidden units represent the dynamics of sentences, that is, the transitions between words. Remarkably, the clusters of words appear to correspond to interesting grammatical categories such as direct object, verb, and so on. Elman's approach is supervised in the sense that the network gets feedback on what it is trying to learn, but unsupervised in the more important sense of inducing categories of words without being told which categories the words belong to.

Learning Concepts by Abstracting Over Dynamics

These approaches have something in common. Starting with a representation of the dynamics of an activity (or in Elman's case, the dynamics of word transitions), cluster trajectories with similar dynamics, then average the elements in a cluster to get a concept. Although this method is extremely simple, it appears sufficient to learn concepts given unclassified instances of activities (i.e., without supervision). The roles problem is still lurking — all the examples in this section finessed it — but abstraction over representations of the dynamics of activities appears to be a promising approach to growing a conceptual structure.

Conclusion

The principal claim of this paper is that agents may grow their own ontologies, unsupervised by us, by interacting with their environments, and this is a good thing for them to do. Success would also address a longstanding problem in human development, namely, how conceptual systems arise from the exercising of motor schemas. I have argued for the primacy of activities: Concepts may represent activities themselves or participants in the activities. The meanings of concepts (and conceptualizations) are just the inferences one can draw about activities and their participants. These entailments are often predictions about how activities will unfold. I showed how activities can be clustered by their dynamics in an unsupervised way. This provides abstract representations of activities which may be used to predict how activities will unfold, but these representations are concepts only because they finesse the problem of identifying the participants in the activities and the roles they play. I believe the roles problem is the only remaining impediment to a complete (and implementable) theory of how concepts are acquired through interaction.

References

R.Baillargeon. How do infants learn about the physical world? Current Directions in Psychological Science, 3:133--140, 1994.

S.Carey and R. Gelman. The epigenesis of mind: Essays on biology and cognition. Hillsdale NJ: Erlbaum, 1991.

Coelho, J.A., Grupen, R.A., ``A Control Basis for Learning Multifingered Grasps," Journal of Robotic Systems, 14 (7): 545 – 557. 1997. Wiley.

Gary L. Drescher. Made-Up Minds. MIT Press, 1991.

Elman, J. 1995. *Language as a dynamical system*. In *Mind As Motion*, R. F. Port and T. van Gelder, Eds. MIT Press. pp. 196 – 225.

Mark Johnson. The Body in the Mind. University of Chicago Press, 1987.

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

Lenat, D. and Feigenbaum, E. A. On the thresholds of knowledge. Proceedings of IJCAI -87. pp. 1173 – 1182.

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

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

Jean Piaget. The Origins of Intelligence in Childhood. International Universities Press, 1952.

E. Rosch and C. B. Mervis. Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7:573--605, 1975.

E. S. Spelke, K. Breinlinger, J. Macomber, and K. Jacobson. Origins of knowledge.. Psychological Review, 99:605--632, 1992.