Grounding the Unobservable in the Observable: The Role and Representation of Hidden State in Concept Formation and Refinement

Clayton T. Morrison
Experimental Knowledge Systems Lab
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003, USA
clayton@cs.umass.edu

Tim Oates
Artificial Intelligence Lab
Massachusetts Institute of Technology
545 Technology Square
Cambridge, MA 02139
oates@ai.mit.edu

Gary King
Experimental Knowledge Systems Lab
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003, USA
gwking@cs.umass.edu

Introduction

One of the great mysteries of human cognition is how we learn to discover meaningful and useful categories and concepts about the world based on the data flowing from our sensors. Why do very young children acquire concepts like *support* and *animate* (Leslie 1988) rather than *between three and six feet wide* or *blue with red and green dots*? One answer to this question is that categories are created, refined and maintained to support accurate prediction. Knowing that an entity is animate is generally much more useful for the purpose of predicting how it will behave than knowing that it is blue with red and green dots.

The idea of using predictability, or a lack thereof, as the driving force behind the creation and refinement of knowledge structures has been applied in a variety of contexts. Drescher (1991) and Shen (1993) used uncertainty in action outcomes to trigger refinement of action models, and McCallum (1995) and Whitehead and Ballard (1991) used uncertainty in predicted reward in a reinforcement learning setting to refine action policies.

Virtually all of the work in this vein is based on two key assumptions. First, an assumption is made that the world is in priciple deterministic; that given enough knowledge, outcomes can be predicted with certainty. Given this, an agent's failure to predict implies that it is either missing information or incorrectly representing the information that it has. Second, it is assumed that knowledge structures sufficient for the task can be created by combining raw perceptual information in various ways. That is, everything the agent needs to make accurate predictions is available in its percepts, and the problem facing the agent is to find the right combination of elements of its perceptual data for this task. (See (Drescher 1991) for an early and notable exception.)

Our position is that the first of these assumptions represents an exceedingly useful mechanism for driving unsupervised concept acquisition, whereas blind adherence to the second makes it difficult or impossible to discover some of the most fundamental concepts. To better understand this position, consider the child-as-scientist metaphor (Gopnik 1997). Generally speaking, scientists aim towards an understanding of the way the world works by developing theories that explain as many observable phenomena as compactly and accurately as possible. Predictiveness is central to this enterprise. If theory A makes either more accurate predictions or a larger set of verifiable predictions than theory B, then theory A is preferred. To explain observed phenomena, scientists often posit the existence of unobservable entities (Harré 1970; 1986). No one has ever seen gravity or black holes, but they explain such a wide range of observable phenomena so accurately that their existence goes virtually unchallenged. Scientific progress would come to a standstill if not for the ability to posit and collect evidence for the existence of causally efficacious entities that do not manifest themselves directly in our percepts in the same way that, say, the color blue does.

To bring the discussion back to concept acquisition in children, consider the following example. Humans clearly cannot perceive the mass of an object in the same way that they can perceive its color. Though it may seem as if we perceive mass directly from visual observation this is an illusion grounded in a vast corpus of knowledge about objects gathered over a lifetime of physical interaction with the world. Without this grounding we would be, like an infant, unable to make judgements about the masses of objects based solely on their visual appearance. Indeed, it is equivocal whether we would even have a concept of mass at all.

Representing the Unobservable

How might a child that cannot perceive mass directly ever hope to gain a concept of mass? Our answer is that the child posits the existence of a property that explains how objects behave and that they later learn that this property is named "mass". Objects with different masses yield different proprioceptive sensations when you lift them and different haptic sensations when you drop them on your foot, they require different amounts of force to get them moving at a given speed and they decelerate at different rates when that force is removed. Positing a hidden feature of objects that explains and is correlated with any of these observations suffices to predict (to some degree of accuracy) all of the others. That is, this hidden quantity makes a wider array of more accurate predictions than any directly observable feature such as color or size.

So far we have mentioned one of the fundamental triggers for positing new unobservable representational elements:

failure to predict. But an account of how the neonate human or machine might support the acquisition of knowledge of these hidden states requires a representational infrastructure that can

- accommodate recognition of failure to predict, and
- accommodate representations whose content is fundamentally reviseable and extendable.

The first property highlights that these representations must provide some way for the agent to recognize that something has gone wrong: its current representational repertoire fails to be useful in predicting some aspect of the environment.

The second property is a little more complicated. First, it highlights that, in its intial, naked form, the proposed *new* representational element serves simply as a "place-marker" associated with the context of the failure to anticipate some aspect of the world. At this point, the agent does not even know whether the new element represents a hidden state, some unobservable property, or some unseen process. Next, given the new element, the agent must *now* search for conditions under which it can reliably predict its state value. Once these conditions are discovered, the state value of this representational element can then be used to make predictions of the the previously unpredictable aspects of the world. These conditions are themselves states of the world that are directly observable or can be accurately anticipated by the agent. They are associated with the newly constructed representational element, serving as conditions for determining its state value. In this way, the *content* (the meaning) of the proposed new representation of an unobservable aspect of the world is fundamentally revisable and extendable.

Interestingly, a common theme is shared among several computational models of representation learning along these lines (e.g., (Drescher 1991; Kwok 1995)). Namely, the base data structure for a representational unit in these discovery systems is a triple involving preconditions, some action, and postconditions; following Drescher's (1991) terminology, we will refer to this data structure as a *schema* (see Figure 1).

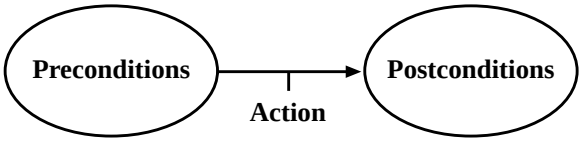

Figure 1: Base structure of representation - preconditions, action, postconditions.

The preconditions and postconditions initially consist of sets of predicates whose values represent directly observable states of the world (e.g., colors, shapes, relative positions of visual objects in the retinal field), and later may include newly learned states (e.g., states that are not directly observable, including new representational elements whose value-determining conditions have already been learned). The actions can be simple physical-motor actions (e.g., opening one's mouth), or more abstracted activities that themselves consist of a structured set of composed actions (e.g., carrying out an experiment). Irrespective of where along the continuum from atomic/concrete to compound/abstract these actions are specified, they all share in common the key property of affecting, in some way, what future input the agent will receive from the world. In this view, there is no such thing as a system that is purely disembodied or passive with respect to the environment it is learning about: the world affects the agent and the agent affects the world (see (Bickhard 1993)). Furthermore, note that even "passive" actions that don't directly causally impinge on states of the world, such as passively observing a scene, still involve an "action" that affects the subsequent input – e.g., passive observation specifically involves sitting still and not altering states of the environment while there is a passage of time.

Constructing a New Predicate

Within this representational framework, it is now possible to create (discover) new predicates that may correspond to unobservable states or properties of the world. These new predicates are then useful for deterministically predicting previously unpredictable behavior of the world. Following the base representational form of the schema, and using a "just so" story involving the dicovery of a mass-like concept, such creation and discovery works as follows:

Initially, the agent does not know how to predict the outcomes of actions or interactions involving objects and such diverse outcomes as proprioreceptive feedback when lifting an object, haptic feedback when the object is dropped on one's foot, the force required to achieve some velocity in moving the object, and the rate of deceleration of the object when one stops pushing. Instead, the different outcomes of these events are simply experienced as they happen. Figure 2 depicts a pair of schemas representing two different experiences of lifting objects that are otherwise perceptually identical (also in conditions that are essentially identical).

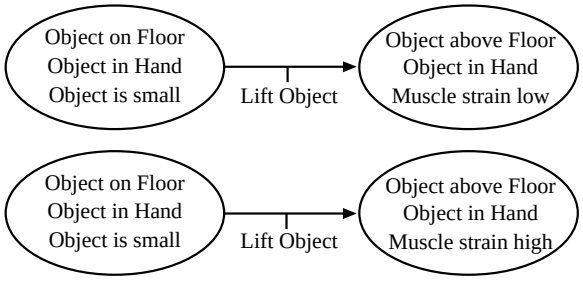

Figure 2: Initial set of schemas with outcomes that are not predictable based on the observable preconditions.

As is clear from the schemas in Figure 2, the agent does not have enough information to properly anticipate what will happen when lifting the small object in the two different cases. Sometimes small objects require a high amount of muscle strain in order to be lifted, other times they do not. This inability to anticipate what will happen serves as the trigger for the agent to propose a new predicate, P_1, whose values will help distinguish between the two situations (similar to Drescher's (1991) synthetic item). P_1 currently has

no content (meaning) other than it will be used to determine the outcome of lifting small objects in the described context (Figure 3).

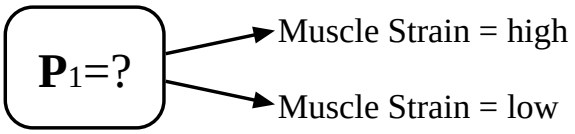

Figure 3: New predicate P_1 – has the function of determining outcome of lifting small objects (context), but currently has no associated conditions for determining its value.

Given this "empty" predicate, the agent now tries to find the conditions which would serve to consistently determine what the non-observable present value of the predicate is (which, in turn, will determine what the outcome of lifting small objects will be). (Note that the "empty" predicate currently has the function of driving the agent to search for the conditions that determine the predicate's state-value; such search could consist of a random walk through the action/state space, or be a result of other behavior of the agent dedicated to other tasks, or could result from specific strategies (learned or innate) for finding predicate value-determining conditions. We do not address such strategies here.) Suppose after some exploration, the agent discovers that if it pushes small objects before lifting them, outcomes of the pushing interactions serve to determine what the outcome of lifting those particular objects will be. And this increased predictability occurs consistently when interacting with the small objects (Figure 4).

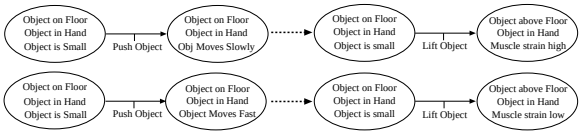

Figure 4: Discovery of consistent outcomes of pushing conditioning consistent outcomes of lifting.

Now the agent has found a set of schemas that can be used to determine under what conditions the proposed unobservable predicate would have particular state-values (Figure 5). That is, even though the initial situations are indistinguishable (small objects cannot be distinguished by simple observation), after carrying out a particular interaction (pushing) with a particular small object, then going on to lift that object will have a particular outcome. These precondition schemas now fill-out more of the content of P_1.

One might argue that at this point the agent can now jettison the original predicate, simply relying on the test of first pushing small objects before lifting them to see what outcome is expected of subsequent lifting. Removing the predicate, however, would make the immediate prior testing by pushing absolutely necessary. While consistent immediate prior testing may be desireable initially (e.g., to ensure that the suspected relationship between preconditions and the desired ability to anticipate the outcome of a future interaction does hold), we would like the agent to eventually only need to determine the prior predicate-value setting conditions once, thereafter knowing that the predicate has a certain value and all subsequent actions (e.g., lifting) will have particular outcomes. This allows the agent to keep track of a property of an object not directly observable, yet which holds over time. This is a powerful addition to the agent's representational repertoir. For example, if a particular small object is pushed and found to move slowly, the agent does not need to perform this test again (or every time prior to lifting) when the agent wants to anticipate what the outcome of lifting will be (Figure 6).

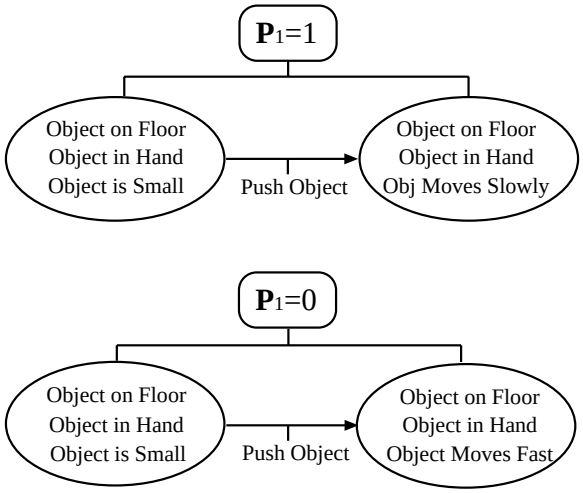

Figure 5: Schemas conditioning the state-value of the predicate P_1.

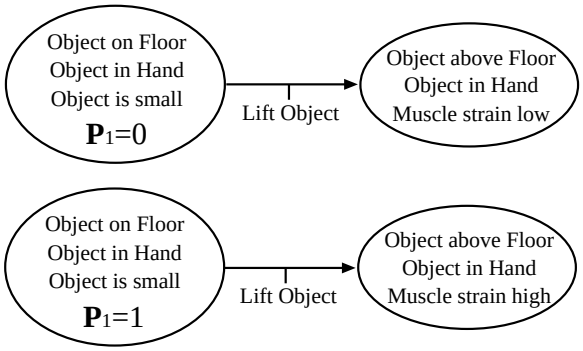

Figure 6: The "hidden state" predicate P_1 can now be used to distinguish outcomes of situations that are currently observationally identical.

It is also not a problem that there are some predicates whose values are not stable over time. Some predicates, for example, may only hold a particular value for a short amount of time; others may change their vlaues based on other conditions. The conditions under which these values change are conditions that the agent would need to further explore and discover in order to make accurate predictions. But they are not in principle unattainable within this framework.

In the above thought experiment, the agent discovered that the outcome of pushing could reliably condition the expected outcome of lifting. But the discovery did not need to happen in this order. For example the agent could have, and still can, discover that the outcome of lifting is a reliable predictor of outcomes of pushing. If this was an additional discovery, a new predicate (P_2) could be posited. However, because of the strong symmetry between pushing determining lifting and *vice versa*, the two predicates could be reduced to a single predicate whose value could be determined by either initial action tests (in turn determining what the other action outcome would be). Now, a single unified predicate, P_3, would have the value 1 when either lifting involved high muscle strain *or* pushing resulted in slow movement of the object, and a value of 1 would predict either outcome as well.

In fact, the use of this symmetry holds generally for any hidden properties, states or processes that may govern many different potential observable situations. As mentioned above, mass plays an important role in determining: (a) the outcome of muscle proprioreceptive feedback while lifting, (b) the amount of pushing force to bring an object on a surface to some velocity, (c) the kind of haptic feedback when objects are dropped on our feet, and (d) how long an object might move before coming to rest after a certain amount of force has been imparted to it. Because of the symmetries in the determining relationships between these different situations (as each is used to alternatively predict the others), they may all be unified into one general predicate, whose value could be determined by particular outcomes of any one of the individual schema action tests (Figure 7).

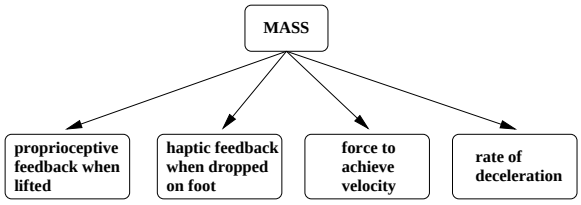

Figure 7: In the terminology of graphical models, the mass of an object renders the other observations conditionally independent.

It is here that we see the real power of keeping the proposed representational predicate: after learning of the connections between these different outcomes in different situations, a single outcome in one situation will immediately determine the outcome of all the others. This also demonstrates another feature of this model, again related to the base mutability of the content (the meaning or semantics) of the predicate: the agent's understanding of what the predicate represents slowly evolves as new relationships between initially independent outcomes are discovered. Certainly, the content of the predicate representation is not determined by the correspondence that it may in fact have with the true property of mass in the world. Also, we have been careful not to call any one of these predicates "mass", as though that predicate exhaustively captured all of the semantics of the term *mass* as scientifically understood today. Still, we can see a clear trajectory headed in that direction, as new mass-determined relationships are discovered and unified under the developing predicate's semantics.

This is a natural segue to considering the powerful role language plays in representational development – particularly predicate discovery of the kind described here. Language can provide labels for these theoretical entities and can serve a structuring role in helping the agent learn about other theoretical entities, as well as facilitate unification. For example, use of linguistic labels by other language users can help to focus attention on specific aspects of the environment requiring explanation; merely hearing the use of a term naturally leads one to seek an understanding of what that term refers to or how it is properly used. Or, having learned a term that the agent suspects refers to a property that its predicate represents, the agent may then test whether their use of the term following this grounding in their predicate representation is correct and/or complete. In general, language provides a proliferation of tools for positing, analyzing and evaluating theoretical entities – useful both for the agent by itself, and in its social interactions. Language greatly improves the speed and correctness of predicate discovery and refinement in children, and should for our would-be epistemic machines as well.

(A final note: We are not making a positivist claim that all theoretical entities are to be understood as defined in terms of observables. Rather, our claim is that, in the limit case of no prior knowledge (such as neonates and our machines), this is the base mechanism by which theoretical entities are discovered, understood and grounded. We do need grounding in observables (derived from potentially ellaborate histories of interaction) in order to have some idea of *what* values those hidden states might currently have – this is no different than the idea of measurement in science. But the ultimate ontological status of the entities, processes or properties that these predicates might represent, while informed by, are not determined by the predicate values and how they're determined; the former are futher inferences made as part of the logic and structure of a scientific theory.)

Concluding Remarks

Science, of course, has developed an ellaborate logic for investigating theoretical entities and the conditions under which they are posited and evaluated. Also, given an already robust base of knowledge about how the world works, positing of theoretical entities may involve ellaborate use of analogy and metaphor, borrowing from already well-known environmental structure (this would involve whole structures being transfered to the new situation, not just single predicate invention; e.g., (Gentner 1989)). Nonetheless, we claim that the above account forms the foundation for grounding discovery and semantics of representation of theoretical entities in the limit of no prior knowledge; and we claim that this form of representation and discovery mechanism is likely present in all systems that make the leap from purely sensorimotor representation to ellaborated representation of the environment that includes unobservable causal structure.

At the core, this mechanism, by which an agent can posit and refine theoretical entities, is funded by a repesentational substrait that (a) accommodates recognition of failure to anticipate (predict) how the world will behave, and (b) accommodates new proposed predicate representations that have extensable and revisable content. As we have shown, the base schema data structure, with its integral relationship between preconditions, actions, and postconditions, is suitable for both of these requirements: (a') in the schema, the agent's own action, indexed by preconditions, entails outcome conditions (effects), allowing the agent to test for itself whether its own representations are true or false – that is, these repesentations bear a truth value that is detectable by the system itself, while also contingent on the environmental category (in this case, the potential unobserved state, process or property) that the predicate is intended to anticipate (see (Bickhard 1993)); and (b') the conditions under which the value for a predicate is determined is based on an evolving set of "triggering" conditions, in turn based on other schemas – these triggering conditions constitute the evolving content (meaning) of that predicate.

Acknowledgements

This research is supported by DARPA contract #DASG60-99-C-0074. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views an conclusions herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA or the U.S. Government.

References

Bickhard, M. H. 1993. Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence* 5:285-333.

Drescher, G. L. 1991. *Made-Up Minds*. MIT Press.

Gentner, D. 1989. The Mechanisms of Analogical Learning. In Vosniadou, S., and Ortony, A., eds., *Similarity and Analogical Reasoning* (pp.199-241). Cambridge University Press.

Gopnik, A., and Meltzoff, A. N. 1997. *Words, Thoughts, and Theories*. MIT Press.

Harré, R. 1970. *The Principles of Scientific Thinking*. Macmillian.

Harré, R. 1986. *Varieties of Realism: a Rationale for the Natural Sciences*. Blackwell.

Kwok, R. B. H. 1996. *Creating Theoretical Terms for Non-deterministic Actions*. In Foo, N. Y., and Goebel, R., eds., *PRICAI'96: Topics in Artificial Intelligence, 4th Pacific Rim International Conference on Artificial Intelligence*, Cairns, Australia, August 26-30, 1996, Proceedings. Lecture Notes in Computer Science, Vol. 1114, Springer.

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

McCallum, A. K. 1995. *Reinforcement Learning with Selective Perception and Hidden State*. Ph.D. Dissertation, University of Rochester, Rochester, NY.

Shen, W.-M. 1993. Discovery as autonomous learning from the environment. *Machine Learning* 12(1–3):143–165.

Whitehead, S. D., and Ballard, D. H. 1991. Learning to perceive and act by trial and error. *Machine Learning* 7:45–83.