Prince, C. G. (accepted). Theory Grounding in Embodied Artificially Intelligent Systems. To be presented at *The First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, to be held Sept 17-18, 2001, in Lund, Sweden.

Theory Grounding in Embodied Artificially Intelligent Systems

Christopher G. Prince
Department of Computer Science, University of Minnesota Duluth,
Duluth, MN, USA, chris@cprince.com

Abstract
Theory grounding is suggested as a way to address the unresolved cognitive science issues of systematicity and productivity. Theory grounding involves grounding the theory skills and knowledge of an embodied artificially intelligent (AI) system by developing theory skills and knowledge from the bottom up. It is proposed that theory grounded AI systems should be patterned after the psychological developmental stages that infants and young children go through in acquiring naïve theories. Systematicity and productivity are properties of certain representational systems indicating the range of representations the systems can form. Systematicity and productivity are likely outcomes of theory grounded AI systems because systematicity and productivity are *theoretical concepts*. Theory grounded systems should be well oriented to acquire and develop these theoretical concepts.

Introduction

Theories are used by humans, and to some extent, perhaps by other animal species, because of their practicality. Informal theories enable us to make predictions about social interactions and physical situations and inform our actions. Formal theories are constructed by certain adults – notably various academic researchers. These individuals are involved in disciplined and culturally organized approaches to the development of theory. It seems evident that young children also develop informal or naïve theories. For example, research into *theory of mind* holds that children cognitively develop skills enabling them to understand the behaviors of others in mental process terms (Flavell, 2000; Wimmer & Perner, 1983; Wellman, 1990).

Enabling computer systems to have skills with theories would be useful from a practical standpoint. Just as humans find theories practically useful, theories can be useful to artificially intelligent (AI) systems. Theories constrain the possible set of hypotheses an AI system considers in its search space (i.e., they act as biases, Mitchell, 1980) and also improve the quality and increase the acquisition rate of hypotheses (Mitchell, Keller, & Kedar-Cabelli, 1986; Mooney, 1993). In short, theories help both AI systems and humans make generalizations. Theories used in combination with data enable principled generalizations to be formed.

This paper introduces *theory grounding* as a new conceptual approach to imbuing computers with theory skills and knowledge. Theory grounding differs in two ways from traditional embedding of theory in AI systems. First, this approach to theory incorporation is *grounded*. This is an extension of Harnad's (1990) symbol grounding and Brooks (1999) physical grounding. Instead of directly programming theoretical biases into an AI system from the top-down, we propose that theories should be causally connected to the world via sensory-motor systems from the bottom-up. Theories do not exist in isolation from the world. Rather, they formulate concepts about regularities and irregularities that exist in the world. Therefore, theories need to be grounded. Second, in a manner similar to the emergent behavior concept of behavior-based robotics, we further propose that theories should not only be grounded, but should also be semi-autonomously learned or *developed* by the embodied AI system. We make this second proposal as a direct extension to the rationale for grounding theories, in general. Theories are not just temporally isolated, static representations and behaviors. On the contrary, skills and understandings related to theories change over time because of additional information, data, examples etc. If the theories acquired by an AI system are developed or learned, then these theories should, as a natural function of that learning and development, change over time, and hence naturally incorporate additional information, data, examples etc.

This paper focuses on two issues in relation to theory grounding: (1) a rationale for theory grounding as a method of achieving a core goal in cognitive science—variously referred to as productivity, systematicity, and conceptual representation, and (2) a route to achieving theory grounding, in a fully developmental manner, by modeling the cognitive development of theory in children. After this presentation, we provide a few notes about computational mechanisms that may help in theory grounding. The last section of the paper closes with discussion and conclusions.

I. Grounded Symbols vs. Grounded Theories in Embodied AI Systems

The concept of symbol grounding (Harnad, 1990) or physical grounding (Brooks, 1999) has led to advancement in AI and robotics. Behavior-based robotics has achieved behavioral abstractions, computed from the bottom-up, which while not necessarily meeting some criteria for what all would call symbol-type abstractions (e.g., MacDorman, 1999), have at least met some mid-way criteria between a level of raw sensory-motor configurations and symbol-type abstractions. For example, the *steered prowling* behavior abstraction of the six-legged robot Genghis is the 8th layer of the control-system architecture for the robot and is grounded with eight layers of processing modules connected to sensors and motors. Steered prowling is causally connected to the sensory-motor system of the robot, and is computed from the bottom-up (Brooks, 1989).

Notably lacking in this behavior-based approach is conceptual representation (Kirsh, 1991) and productivity and systematicity (MacDorman, 1999). Productivity and systematicity are properties of certain representational systems. Productivity refers to a system being able to encode indefinitely many propositions (Fodor & Pylyshyn, 1988). Systematicity occurs when representing a relation *a***R***b* implies the system can also represent *b***R***a* (Fodor & Pylyshyn, 1988). In our view, this "lack" of conceptual representation, productivity, and systematicity is not surprising if we assume a starting theoretical basis of symbol grounding. Not only is there a great deal of variation in what is considered to be a "symbol" in the cognitive science literature, but also there is little in the way of offered method to achieve conceptual representation, productivity, and systematicity in the ideas of symbol grounding. In this author's view, the key contribution of symbol grounding is noticing that AI systems should not just have their internal program symbols connected to other internal program symbols. Rather, they should have their symbols connected to the real world. Such a step in our understanding of AI systems is necessary, but not sufficient to achieve other psychological features oftentimes associated with symbols and symbol-use, such as conceptual representation, productivity, and systematicity.

A few examples of contemporary use of the term "symbol" should illustrate the variation in use of this term. Gallistel (2001; Gallistel & Gibbon, 2000) uses the term symbol to mean "an objectively specifiable aspect of [an] animal's experience—for example, the duration of a conditioned stimulus—[which] enters into information processing operations, such as the combinatorial operations of arithmetic (addition, subtraction, multiplication, division, and ordination)" (Gallistel, 2001, abstract). Gallistel argues that a purely associationistic view of conditioning phenomena in nonhuman animals is untenable, and he makes this argument in terms of information processing or symbolic models. Of course, the area of animal language research is also replete with symbolic processing claims (Herman, Richards, & Wolz, 1984; Prince, 1993; Savage-Rumbaugh, Murphy, Sevcik, Brakke, Williams, & Rumbaugh, 1993). Animals including rats (Macuda & Roberts, 1995), parrots (Pepperberg, 1992), gorillas and orangutans (Byrne & Russon, 1998) have also been found to have limited skills often associated with symbolic processing – recursive or hierarchical processing (Touretzky & Pomerleau, 1994). In the area of computational modeling various forms of recursive or hierarchical processing and systematicity have been demonstrated (e.g., Elman, 1991; Pollack, 1990).

This range of the contemporary scientific concept of symbolic skills or processing might be thought of as scientific discourse that has not yet settled on a precise characterization. Alternatively, and the view held here, it can be the case that the very concept of symbol and symbol processing (like the term "representation") is open to interpretation and thus has a great deal of variation associated with it. In terms of AI systems, and particularly physically grounded AI systems, one can ask the question: What have we achieved with symbol grounding? It seems apparent that we have not achieved conceptual representation, productivity, and systematicity. Why? The concept of symbol and the associated framework of terms and concepts (e.g., see Fodor & Pylyshyn, 1988) do not guide us towards an understanding of these ideas that naturally leads us to realizing conceptual representation, productivity, and systematicity in our AI systems.

A theoretical proposal of the present paper is that conceptual representation, productivity, and systematicity arise as a consequence of the theoretical structure of a system. That is, as a consequence of the skills and knowledge that a system has for acquiring, processing, and representing theory. Theoretical structures enable the processing and representation of structures related to infinite competence (Chomsky, 1968). It is not that a system can process infinite size or infinite duration structures. *Any* finite system will have finite performance limits. However, a system can have *concepts* of infinite size objects and *concepts* of infinite duration events. These are fundamentally theoretical skills and knowledge. It is for these reasons that arguments about limited depth hierarchical or recursive processing in nonhuman animals or computational systems falter. Limited depth arguments suggest that humans have a performance depth capacity of some number *M* and some other animal species has a performance depth capacity of some number *N*, where *M* is greater than *N* (e.g., Byrne & Russon, 1998). This, however, misses a crucial point. It is far more relevant that humans have *concepts relating to infinity* as

opposed to being able to practically produce or understand sentences of some particular limited depth. For other animals or machines, it is more interesting to ask if they have concepts of infinity rather than practically performing to particular limited depths of (e.g., sentence) processing.

II. Theory Understanding and Skills

Theory grounded approaches show theoretical promise for resolving issues in embodied AI systems. The approach also offers some more specific guidance towards these unresolved issues. Young children acquire relatively broad capacities with theory understanding and attain understandings of various specific theoretical topics. The guidance offered by this approach is that of methods and results from child development and especially that of early infant development. Theories and data from cognitive development can provide us with constraints on designing systems that develop their own theories. We suggest that a way to approach theory grounding is by modeling the cognitive development of theory understanding and skills in young children.

Before we turn to an example of one area of naïve theory development in children, two issues deserve addressing. The first is that of development itself—we assume here a generally non-nativist position towards psychological development. The second is the specificity of theory acquisition. Both general and specific theory skills can be acquired.

Why Development?

While some theorize that theory-building skills in children have highly abstract innate or biological components (Gopnik & Meltzoff, 1997; Meltzoff, 1999), the view taken here is that a series of cognitive developments must take place in children to allow them to think and behave theoretically. This seems reasonable because various theory-related skills become available to children at various ages. For example, it is not until about 5-years-of-age that children can utilize a representational view of others' mental states. It is at this age that children start to be able to utilize distinctions related to appearance-reality, false-belief, and perspective taking (Flavell, 2000). This view, while also taken by others (e.g., Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996), is particularly useful here. A view of an ongoing period of development in acquiring theory should enable an AI system to more robustly respond to exceptions to the formed theory.

General vs. Specific Theories

There are two overall ways in which embodied AI systems can develop theories. First, the system can acquire skills and knowledge regarding a particular theory. For example, a theory of objects could be acquired covering topics such as friction, gravity, and agents (including self) interacting with objects. Second, the system could acquire the skills and knowledge needed to develop skills with particular theories. Acquiring general skills and knowledge needed for acquiring particular theories has been referred to as development of a *theory theory* (Gopnik & Meltzoff, 1997). That is, acquisition of a theory about theories. It is this latter development that we suggest should be of most interest to researchers interested in theory grounding. If we can imbue our embodied AI systems in such a way that they can develop their own theory or theories about theories, then they should be able to acquire their own particular theories about the world.

Of course, it is a question as to how far this approach can be taken. It may be most efficient to start off early and specific in terms of the type of theory being developed. For example, some developmental psychologists posit that the perception of relatively sophisticated social concepts occurs early in human infancy (e.g., Woodward, 1999).

Theory of Mind

We now turn to a specific example of naïve theory development in human children. We advocate theory grounding utilizing behavioral data and tasks from the experimental study of child development to constrain the theory-grounding problem.

As we grow and develop from infancy through childhood to adulthood, humans develop skills with theories and come to acquire specific theories. As scientists, we are used to requiring that theories generally are testable. The specific theories developed and acquired by children and adults (outside academic contexts) appear to have some properties related to testable, scientific theories. One area in which we as humans develop such theories is the area of social understanding, known as theory-of-mind. Wellman (1990) suggests "that our naive understanding of mind, our mentalistic psychology, is a theory. It is a naive theory but not unlike a scientific theory" (p. 2). The area of theory-of-mind seems particularly relevant to us in the context of theory and symbol grounding. One aspect of symbol use often taken for granted is the referential nature of symbols. Symbols refer or are about something. As humans, we use symbols referentially in communication and this referential use typically involves utilizing an understanding of the intended receiver of the communication. That is, your theory of mind is involved in generating language utterances that you believe the other person will understand. Hence the referential nature of symbols has a great deal to do with social understanding and relates to theory-of-mind.

Some features of theory development in children are fairly well established, while other aspects are still being filled in by empirical research. In a child's theory-of-mind one well established feature is that there is a striking change between three and five years of age on

children's skills on tasks such as false-belief and appearance-reality (Flavell, 2000). In the false-belief task the question of interest is: Can a child represent someone else as having a false belief about the world? A false belief can arise if a situation changes unbeknownst to the person. For example, if I think my truck has a full tank of gas, but actually the tank has developed a leak, and the tank is empty, then I have a false belief about the fullness state of my gas tank. Three-year-olds typically perform poorly on tests of false belief, whereas five-year-olds can understand that someone can have a false belief about the world. In appearance-reality tasks the issue is: Can the child realize that something may look like one thing and actually be another? For example, a fake rock can look like a rock but actually be a sponge. Again, the younger children typically perform poorly on these tasks, but the older children understand this distinction. Wellman (1990) has characterized this change from three years to five years as a change from desire (or simple desire) psychology to belief-desire psychology. Three-year-olds are conceived of as being able to mentally represent the goals and desires of others, but not being able to represent the beliefs of others. These children think in terms of object-specific desires, which while not propositional, do involve reasoning about others as having internal longings for or attitudes about objects. Three-year-olds do not yet reason about others as having attitudes about propositions. The belief-desire reasoning of five-year-old children does, however, involve attitudes about propositions. That is, the desires reasoned about by five-year-olds are desires about propositions or represented states of affairs (Wellman, 1990).

What is particularly important to a theory grounded approach to embodied AI is the development of theory skills from early infancy. In order to properly ground the theory, the theory skills of our artificial systems must be grounded from the bottom-up. An area of early infant behavior that holds promise is early causal learning. Watson (1972) and Rovee-Collier (1990) provide examples of this type of infant behavior. For example, Watson (1972) demonstrated that 2-3 month old infants rapidly learn that a mobile above their head is activated (by a concealed switch) when they turn their head. Quickly, they learn to utilize this contingency and cause the mobile to move and smile and coo while doing so. At this young age, the infants appear to have a learning focus on temporally contingent and structurally parallel features, and by 8-10 months of age, they start to appreciate the role of spatial contact in causality (see Gopnik & Meltzoff, 1997, chapter 5). These developmental phases may provide a fruitful modeling source for theory grounded approaches. Algorithms need to be developed that model such developments and do so in a parsimonious way. Of course, some innate primitives need to be assumed, and some environmental constraints need to be provided. Important aspects provided by the infant data include the kinds of perceptual features (e.g., temporally contingent and structural parallelism) that the infants principally attend to. A vital question in modeling is just how do we construct our developmental computations so that we do not have to pre-design much if any of the developmental sequence? How much of the developmental sequence can emerge as a result of the consequences of the very learning done by the child, and how much is a result of a relatively pre-programmed developmental sequence?

III. A Note About Computational Mechanisms For Grounded Theories

Embodied computational models of theory understanding have no small feat to achieve. One necessary issue that must be addressed in these models is that of modeling changes over the course of development. Some approaches have been offered, and elaboration of these techniques is needed. For example, in a connectionist model of the classical Piagetian A-not-B task, Munakata (1998) represented different aged children using varied strength of recurrent model weights. The A-not-B phenomena occurs when an infant is first allowed to search for an object at location *A*, and then on subsequent trials with placement of the object at location *B*, the infant persists in searching at location *A* (see review in Newcombe & Huttenlocher, 2000). A more general connectionist approach to modeling development is provided by the cascade-correlation algorithm (Fahlman & Lebiere, 1990), which is a supervised method that recruits new hidden units, and in doing so fixes the input weights to the unit. Starting from a reduced network, the architecture is developed to fit the task. Some other research has constructed embodied models of infant performance in areas such as mother-infant interaction (Breazeal & Scassellati, 2000). A challenge posed by theory grounding is to combine techniques such as used by Breazeal and Scassellati (2000) with a developmental learning approach (e.g., Fahlman & Lebiere, 1990) in the context of behaviors and developments underlying theory development in infants. Some further approaches are reviewed in Schlesinger and Barto (1999) and contained in Weng and Stockman (2000).

IV. Conclusions and Discussion

Theory grounding proposes that the theories our computers build should be intimately connected with the world. If we imbue computers or robotic systems with theory understanding directly (i.e., through programming of that understanding), then the computers' theoretical understanding likely will not be well connected with the world. Consider what happens when exceptions to a programmed theory arise. As in all theories, there will be exceptions. Adaptation of this theory in the face of exceptions will likely be difficult

because of the lack of world connection. However, if we design algorithms that enable the AI systems to acquire their theories through processes of simulated cognitive development, then the theories they construct will be connected with the world because the process of cognitive development itself involves an extended process of world interaction. Exceptions to the theory will be blurred with the very process of theory acquisition and development itself. When an AI system is semi-autonomously acquiring its theory, what distinguishes an exception to accommodate into the theory from a new datum to be assimilated into the theory?

Theory grounded systems should be more efficient in terms of both training and performance. Training should be improved both because the system will semi-autonomously acquire its own initial theory about theories, and also because it will semi-autonomously acquire specific theories. It will be an explorer and an investigator. Searching for knowledge, not just food goals. The systems should also be more efficient from a performance stance. To the extent that the systems acquire their own theories, they will generate hypotheses more effectively. Part of the notion of a hypothesis being acquired more effectively is the hypothesis being more closely related to reality and hence providing the AI system a better gain in the task it is trying to perform. Perhaps fewer, better quality, hypotheses will need to be tested to arrive at a goal.

Another outcome, we suggest, is that theory grounded systems will take us further towards the goals of conceptual representation, systematicity, and productivity in our artificially intelligent systems. We propose here that the infinite concepts associated with systematicity are exactly that: theoretical concepts. It makes little sense to imbue our connectionist models, for example, with limited depth recursive capacities when the infinity in natural language or sets or programming languages (to pick a few domains), is theoretical in nature and thus best captured by a system that acquires *theories* of these concepts.

In closing it may be apparent that we have not addressed the issue of language and modeling language abilities. Ideas of systematicity and productivity are strongly associated with language (Fodor & Pylyshyn, 1988). It seems likely that both cognition and language will be necessary in a system that fully models the development of theoretical skills in a manner similar to that of humans. It is likely no mistake that theories are codified frequently in language. The approach of theory grounding is conceptually open to modeling skills from both cognition and language.

Acknowledgements

I thank my Ph.D. advisor, Daniel J. Povinelli, who provided me with opportunities to tap into the area of theory of mind. Collaboration with Istvan Berkeley on a related topic provided the genesis for some of these ideas. Many other colleagues, students, and friends have helped me shape these ideas as well. Eric Mislivec has provided valuable discussions. Darrin Bentivegna provided comments on this article.

References

Breazeal, C. & Scassellati, B. (2000). Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior*, *8*, 49-74.

Brooks, R. A. (1989). A robot that walks: Emergent behavior from a carefully evolved network. *Neural Computation*, *1*, 253-262.

Brooks, R. A. (1999). *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: MIT Press.

Byrne, R. W. & Russon, A. E. (1998). Learning by Imitation: a Hierarchical Approach, *Behavioral and Brain Sciences*, *16*, 667-721.

Chomsky, N. (1968). *Language and Mind*. New York: Harcourt, Brace and World.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195-225.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Fahlman, S. E. & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems*, *Vol 2* (pp. 524-532). San Mateo: Morgan Kaufmann.

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

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development*, *24*, 15-23.

Gallistel, C. R. (2001). *The Symbolic Foundations of Conditioned Behavior*. Colloquium given at University of Minnesota, Minneapolis, MN on May 11, 2001.

Gallistel, C. R. & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*, 289-344.

Gopnik, A. & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Herman, L. M., Richards, D. G., & Wolz, J. P. (1984). Comprehension of sentences by bottlenosed dolphins. *Cognition*, *16*, 129-219.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335-346.

Kirsh, D. (1991). Today the earwig, tomorrow man? *Artificial Intelligence*, *47*, 161-184.

MacDorman, K. F. (1999). Grounding symbols through sensorimotor integration. *Journal of the Robotics Society of Japan*, *17*, 20-24.

Macuda, T., & Roberts, W. A. (1995). Further evidence for hierarchical chunking in rat spatial memory. *Journal of Experimental Psychology: Animal Behavior Processses*, *21*, 20-32.

Meltzoff, A. N. (1999). Origins of theory of mind, cognition and communication. *Journal of Communication Disorders*, *32*, 251-269.

Mitchell, T. M. (1980). *The Need for Biases in Learning Generalizations*. Technical report CBM-TR-117, Computer Science Department, Rutgers University, New Brunswick, NJ.

Mitchell, T. M., Keller, R., & Kedar-Cabelli, S. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, *1*, 47-80.

Mooney, R. J. (1993). Integrating theory and data in category learning. In: G. Nakamura, R. Taraban, & D. L. Medin (Eds.), *Categorization by Humans and Machines: The Psycholology of Learning and Motivation*, Vol. 29 (pp. 189-218). Orlando, FL: Academic Press.

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the A $\overline{B}$ task. *Developmental Science*, *1*, 161-211.

Newcombe, N. S. & Huttenlocher, J. (2000). *Making Space: The Development of Spatial Reasoning*. Cambridge, MA: MIT Press.

Pepperberg, I. M. (1992). Proficient performance of a conjunctive, recursive task by an African gray parrot (*Psittacus erithacus*). *Journal of Comparative Psychology*, *106*, 295-305.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*, 77-105.

Prince, C. G. (1993). *Conjunctive Rule Comprehension in a Bottlenosed Dolphin*. Unpublished masters thesis, University of Hawaii.

Rovee-Collier, C. (1990). The "memory system" of prelinguistic infants. In A. Diamond (Ed.), *The Development and Neural Bases of Higher Cognitive Functions*. Annals of the New York Academy of Sciences (no. 608, pp. 517-542). New York: New York Academy of Sciences.

Savage-Rumbaugh, E. S., Murphy, J., Sevcik, R. A., Brakke, K. E., Williams, S. L., & Rumbaugh, D. M. (1993). Language comprehension in ape and child. *Monographs of the Society for Research in Child Development*, *58*, pp. v-221.

Schlesinger, M. & Barto, A. (1999). Optimal control methods for simulating the perception of causality in young infants. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society* (pp. 625-630). New Jersey: Erlbaum.

Schlesinger, M. & Parisi, D. (2001). The agent-based approach: A new direction for computational models of development. *Developmental Review*, *21*, 121-146.

Touretzky, D. S. & Pomerleau, D. A. (1994). Reconstructing physical symbol systems. *Cognitive Science*, *18*, 345-354.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103-128.

Watson, J. S. (1972). Smiling, cooing, and "The Game." *Merrill-Palmer Quarterly*, *18*, 323-339.

Wellman, H. M. (1990). *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Weng, J. & Stockman, I. (2000). *Workshop on Development and Learning* held at Michigan State University, Kellogg Center, East Lansing, MI, USA, April 5-7.

Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior & Development*, *22*, 145-160.