Modeling the Creation and Development of Cause-Effect Pairs for Explanation Generation in a Cognitive Architecture

John Licato[1], Nick Marton[2,3], Boning Dong[2,3], Ron Sun[3], and Selmer Bringsjord[2,3]

Analogical Constructivism and Reasoning Lab (ACoRL)[1]
Indiana University/Purdue University - Fort Wayne
Fort Wayne, IN, USA

Rensselaer AI and Reasoning (RAIR) Lab[2]
Rensselaer Polytechnic Institute (RPI)[3]
Troy NY USA

Abstract. The ability to generate explanations of perceived events and of one's own actions is of central importance to how we make sense of the world. When modeling explanation generation, one common tactic used by cognitive systems is to construct a linkage of previously created cause-effect pairs. But where do such cause-effect pairs come from in the first place, and how can they be created automatically by cognitive systems? In this paper, we discuss the development of causal representations in children, by analyzing the literature surrounding a Piagetian experiment, and show how the conditions making cause-effect pair creation possible can start to be modeled using a combination of feature-extraction techniques and the structured knowledge representation in the hybrid cognitive architecture CLARION. We create a task in PAGI World for learning causality, and make this task available for download.

Keywords: Explanation, Cognitive Architecture, CLARION, Analogy, Causality

1 Introduction

Faced with some unfamiliar event, an agent[1] will attempt to make sense of it by constructing an explanation, even if the explanation that ultimately gets accepted is not entirely coherent. Generating explanations is also important to artificial cognitive systems, particularly those that need to communicate with other humans, for example, to present rationales for its own actions.

[1] In this paper, 'agent' will refer to any actor (artificial or natural) capable of cognitive thought, 'cognitive system' will refer to any system that attempts to model cognitive phenomena, and 'cognitive architecture' will refer to full cognitive systems (such as CLARION) satisfying the definition of cognitive systems in [20].

Previous work (e.g., [6, 9, 14]) modeled the generation of explanations by using structured representations of cause-effect pairs. In an extremely simple case, explaining some explanandum e involves finding a cause-effect pair (c, e), where c is either believed to be true by the reasoner or plausible to the reasoner in some sense. More complicated explanations can be generated by collecting a sequence of cause-effect pairs and lining them up to produce a causal chain [14], by drawing from multiple source analogs simultaneously [8, 9], or a number of other possible approaches. But these approaches all seem to presuppose the existence of cause-effect pairs, and little is done in the way of actually modeling how the initial cause-effect pairings are initially created.

In this paper, we attempt to understand how the sort of cause-effect pairs that are used in explanation generation can be created in an autonomous agent, in a psychologically plausible way. Section 2 reviews some literature on the emergence of causality in children, focusing on a classical Piagetian experiment we will call the *floating task*. We then describe a task, implemented in the simulation environment PAGI World, for testing abilities that underly the autonomous creation of cause-effect pairs, along with an algorithm to perform this task, implemented in the cognitive architecture CLARION (Section 3). Section 4 discusses future work and concludes.

2 The Development of Causality

If we are to understand how cause-effect pairs can be created automatically by a cognitive system, it would be very helpful to understand how the ability to reason causally develops in humans. We will start with a particularly relevant Piagetian experiment.

2.1 The Piagetian Floating Task

In one of Jean Piaget's early works, *The Child's Conception of Physical Causality*, Piaget introduced a task to elicit clues from children as to how they generate explanations. In what we will refer to here as the *floating task*, Piaget presents a series of objects to a child (e.g., a wooden boat, a pin, a pebble, and so on) and asks the child to predict whether or not the object (the candidate floating object) would float. The child makes his prediction, explaining his or her reasoning when possible, and then the object is placed in the water. The child watches whether or not his prediction was correct, and then is asked to explain why the object did or did not float.

Piaget found that the responses given by children seemed to be roughly categorizable into four stages. These stages are to be seen as continuously changing behavioral phenomena, meant to describe general trends noticed in subjects' explanations. In the first stage, explanations are characterized by "animistic and moral reasons," e.g. a boat will float "because they must always lie on the water," or a piece of glass will sink "because it's not allowed to put glass on the water" [17, p.136]. Piaget described stage-1 explanations as moral because they seemed to him to imply a sense of social obligation on the part of the inanimate objects, as opposed to adherence to some natural law.

In the second stage, starting at about 5 years of age, we see the appearance of dynamism, or the invocation of an abstract force in explanations. Children explain that boats float because they're heavy, big, or because the "water is strong." However, they apply their explanations in inconsistent or contradictory ways. Compare this to the third stage (starting at about 5 or 6 years), where children instead tend to use the explanation that boats are *light*, rather than heavy. The difference here, according to Piaget, is subtle but important: floating is no longer explained by an appeal to a simple property of the candidate floating object. Rather, the lake "produces an upward-flowing current which sustains the lighter [floating] body." In other words, floating is understood to be a property that emerges out of an interaction necessitated by both properties of the lake and properties of the candidate floating object.

Finally, in the fourth stage (starting at about age 9, but parts of which are seen as early as ages 6–8), we start to see reasoning taking into account multiple properties of an object simultaneously. By referring to the hollow-ness of the boat, for example, children relate the boat's volume to its weight. Furthermore, whereas in stage 3 properties of the candidate floating object like light-ness or heavy-ness are no longer regarded by the child to be absolute, internal properties. Instead, they are seen as properties that only hold relative to something else (in this case a corresponding volume of water).

2.2 Why Piaget?

Piaget's work is extremely voluminous, spanning almost 60 years, and careful scholars have noted evolutions in Piaget's thought that at times puts the younger Piaget against the older [3]. In part because Piaget's writings are so spread out over so many books, many of his concepts, which he refined in his later years, are subject to misinterpretations of the highest order. For some corrections of misunderstandings of Piagetian concepts, see [4, 15, 12].

For example, the description of stages that we reiterated in Section 2.1 is exemplary of the type of stage-based development that critics are quick to claim is virtually useless, since the scientific consensus is that "cognitive changes are gradual and cumulative" [1]. Contrary to such claims, however, Piaget was very aware of the limitations of using stages in describing children's behavior:

> "[S]tages must of course be taken only for what they are worth. It is convenient for the purposes of exposition to divide the children up in age-classes or stages, but the facts present themselves as a continuum which cannot be cut up into sections. This continuum, moreover, is not linear in character, and its general direction can only be observed by schematizing the material and ignoring the minor oscillations which render it infinitely complicated in detail" [18, p.17].

That being said, it is not the goal of this paper to mount a full-scale defense of the Piagetian body of literature. Although it cannot be denied that some of

Piaget's theories are incompatible with, and need to be refined by, more recent work in developmental psychology, let it suffice to point out that the critics of Piaget are overzealous in indiscriminately discarding the entirety of his work, especially the almost 60 years of qualitative observations of children's behavior. Even if one were to ignore all of Piaget's proposed explanations for developmental mechanisms, his observations remain a fertile ground for cognitive modelers, as they provide at the very least a set of expected behaviors of children of different ages when faced with very specific tasks. We described some of these behaviors in Section 2.1, and the current paper intends to model them.

3 Modeling the Development of Cause-Effect Representation in CLARION

The CLARION cognitive architecture [19] is divided into four subsystems: the action-centered, non-action-centered, meta-cognitive, and motivational subystems. Each of these is split into explicit and implicit components, thus enabling the deliberative processes associated with localist representations to work in parallel with the automatic processes associated with distributed representations. This dual-process approach to modeling cognition has been shown to be capable of modeling a variety of behavioral phenomena in psychologically plausible ways. For example, [22] implemented similarity-based and rule-based reasoning in the non-action-centered subsystem (NACS for short). Building on these processes, [13] showed that structured knowledge, and thus primitive deductive and analogical reasoning, can also be modeled in the NACS. And building on the structures of [13], the authors demonstrated a high-level approach to generating explanations of varying quality in [14].

The present paper can be considered another in that sequence. As mentioned earlier, the previous model of explanation [14] used cause-effect pairs, implemented as *template structures* (a particular type of organization of localist chunks in the explicit level of the NACS). But where do these cause-effect pairs come from? One way, suggested by the performance of the younger children in Piaget's floating task, is known as *feature selection.* Given a set of features of the object under consideration, the child will somehow select some subset of these features (in the case of stage-1 children, a subset consisting of a single feature) and hypothesize that the presence of this particular feature is the cause of the phenomena under observation (floating or sinking). In CLARION, feature selection comes naturally out of the operations of a backpropagation network built into CLARION [21].

In CLARION's NACS, localist chunks corresponding to outputs can be placed on the top level, and microfeatures corresponding to inputs and hidden nodes can be placed on the bottom level. In Section 3.2, we set up the NACS in this way, and apply a feature selection algorithm to the floating task. First, we turn to a description of our computational simulation of the floating task.

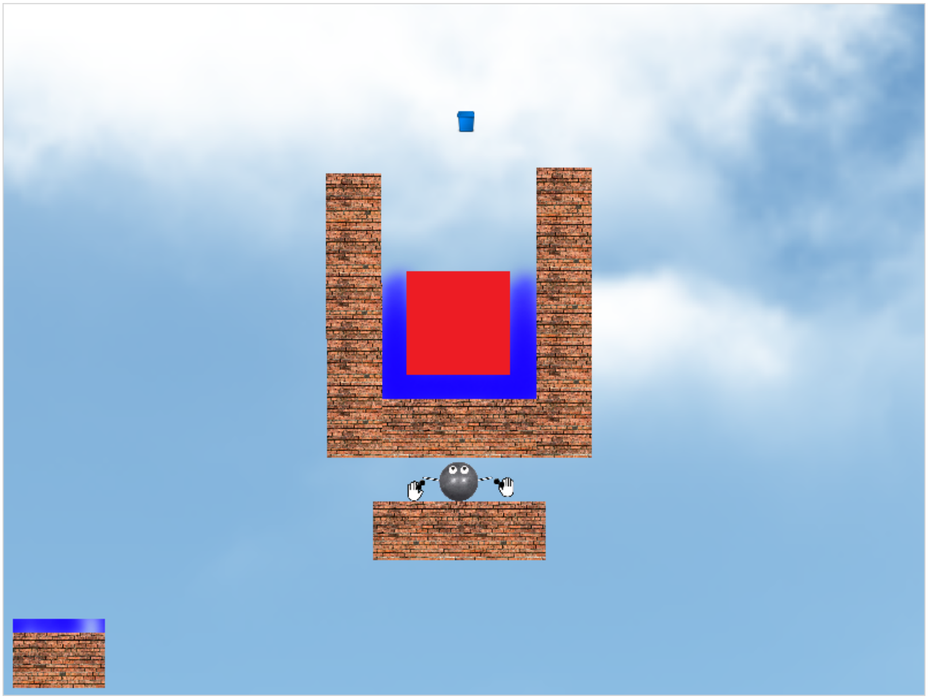

Fig. 1. The Floating Task

3.1 A Task in PAGI World

PAGI World [2, 16, 12] is a simulation environment for the evaluation and development of AGI and cognitive systems. PAGI World is built in Unity, allowing for execution on all major operating systems. It is built on Unity's 2D physics engine, so that mass, volume, velocity, texture, temperature, etc., can be experienced by the AI actor in a realistic way. The AI actor (a ball-shaped creature with two hands, who we sometimes refer to as 'PAGI guy') is controlled by a script that can be written by researchers in any programming language that supports TCP/IP. The information sent between the controller script and PAGI World is mostly low-level: PAGI World sends information from its visual, tactile, and other sensors (including some medium-level data such as object names), while the controller script can send commands to apply a force vector to PAGI guy's body and hands to control it.

PAGI World is easy to learn and use, thanks to design choices that we hope will encourage researchers to make use of PAGI World. Because it can be run on almost any operating system and controlled using almost any programming language, PAGI World provides a platform for cognitive architectures of all types (particularly those which claim to be general-purpose) to compare their performance on the exact same tasks.

Piagetian experiments are somewhat difficult to model computationally for two important reasons: First, they often rely on objects that need to move in a physically realistic way, and it is nontrivial for researchers to program sufficiently realistic simulations for every model they create; second, assessing agents in Piagetian experiments makes heavy use of explanatory dialogue, that is, the experimenter must be able to ask questions about the task and the subject must be able to answer them. Although this second difficulty is one that is still beyond the reach of AI researchers, the first difficulty is handled quite nicely by PAGI World, since PAGI World has the ability to simulate water and create objects that float or do not float in it.

Thus, for all of the reasons discussed above, PAGI World is an ideal choice for hosting the Piagetian floating task. In our implementation, PAGI guy is positioned below a tank of water. An object with a randomly generated volume and weight is created, and appears in the middle of the tank, where it then either floats to the top, sinks to the bottom, or stays relatively motionless (Figure 1). After a few seconds, this object disappears and the process repeats. This allows PAGI guy to collect data about what it observes, so that we can later ask questions.

Algorithm 1 The Feature Selection Algorithm Used in Each Experiment of Section 3.2

```
Require: Set of features $F = \{f_i\}$
Require: Set of training examples $EX = \{ex_1, ..., ex_n\}$
Require: Number of epochs $e$
Require: Number of iterations $it$
  for $i$ iterations do
    for all $f_i \in F$ do
      Build neural network $n_i$, using only $f_i$ as sole input
      for doe epochs
        for do$ex_i \in EX$
          Execute network $n_i$ with $ex_i$
          Update weights of $n_i$ w/backpropagation
        end for
      end for
      for do$ex_i \in EX$
        Execute $n_i$ with $ex_i$; compare prediction to actual result
      end for
      Determine final error rate by averaging over all $ex_i \in EX$
    end for
    Select the feature that had the highest accuracy rate
  end for
  Return the number of times each feature was selected as having the highest accuracy
  rate
```

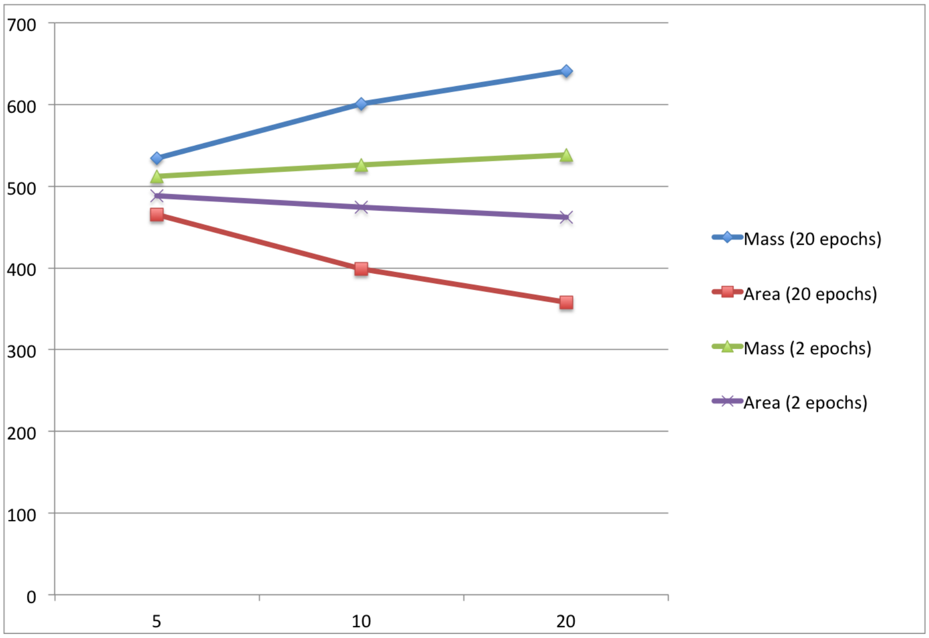

Fig. 2. A plot of the number of training examples n (x-axis) vs. the amount of times each feature was chosen for having the lowest error rate (y-axis). Note that only mass and area are shown here, since the other color-related features showed less than once per thousand iterations.

3.2 Bottom-Up Feature Selection in CLARION

In this section we demonstrate that a simple feature selection algorithm can be implemented in CLARION, by using a network that takes in low-level microfeatures and outputs a prediction as to whether an object will float, sink, or remain stationary. Feature selection is an inherent property of backpropagation, in the sense that as backpropagation updates weights, certain nodes (which can correspond to features) will have higher weights connected to them than others.

CLARION is designed to work with low-level distributed networks that can be trained with backpropagation. We started by creating a network consisting of five inputs, all microfeatures in the bottom level of CLARION's NACS: mass, volume, and three microfeatures for color (red, green, blue). Each input can be activated by a value between 1 and 255. Three outputs are created, each of them implemented as a chunk in the top level of the NACS: float, sink, and stationary. We also create five additional microfeatures $h_1, ..., h_5$, to serve as the hidden layer of the network.

Feature selection proceeds as follows. We collected sensory data from instances of the floating task in PAGI World, where an object of randomized color, mass, and weight appears in the middle of the tank and floats, sinks, or remains stationary. Each instance of an object appearing in the floating task is recorded and called an *example*. The input features are then individually isolated; that is, we only activate one feature at a time, allow the activation to propagate up to the hidden microfeatures ($h_1, ..., h_5$), and further up to the output chunks, and the output chunk with the highest activation is taken to be the 'prediction' of this particular instance. We repeat this for n examples; weights are updated using backpropagation after every example. One successful run-through of all n examples is called an *epoch*. We then execute another epoch, runing through the same n examples again.

After e epochs, we evaluate the average error on the same n examples that the network was trained on. Note that this differs greatly from standard machine-learning practice: generally a test data set is used that is non-overlapping with the training data set. However, we are not necessarily interested in getting the correct prediction; we are interested in modeling the reasoning of the child in a way that is psychologically plausible. It is psychologically plausible that a child would use a limited set of examples from his memory to validate hypotheses or features, and it is less plausible that a child would run through a set of thousands of training examples first.

In any case, the evaluation of error on the n examples gives us an error rate for the feature that was isolated. We can then repeat this entire process with the other features, obtaining an error rate for each feature. The feature that had the lowest error rate is taken to be the winner of this iteration. (Originally, we also recorded the feature that had the second-lowest error rate, but because the results were so overwhelmingly in favor of mass and volume (a color-related feature was selected less than once per 1000 iterations), we only present the data here for the lowest error rate.) The feature-selection algorithm is laid out in a more convenient form in Algorithm 1.

The iterations were repeated 1000 times per experiment. We carried out this experiment six times, for three different values of n ($n \in \{5, 10, 20\}$) and two different values of e ($e \in \{2, 20\}$). Figure 2 shows the value of n on the x-axis, and the number of times (out of 1000 iterations) some particular feature was chosen as having the lowest error rate on the y-axis.

The values of n we chose for each experiment were intentionally very small. It seems implausible that children carrying out the floating experiment would actually be trained using hundreds of instances before they output their predictions. Therefore, we kept n very low in order to see what results emerged. As it turns out, the results match our intuitions: using our feature-selection algorithm settles extremely quickly on either the mass or volume features, and the only growth we see as n and e are increased is a slowly growing gap between the amount of times mass is chosen and the amount of times volume is chosen (a gap which was larger for 20 epochs than it was for 2 epochs).

The fact that even tiny values of n and e identify mass and volume as the most relevant features is consistent with the idea that, in line with Piaget's suspicions, the growth allowing the more complex explanations of stage-2 and later reasoning is a growth in the complexity of the representations themselves—that is, new nodes (corresponding to new concepts) might be created to represent abstract ideas such as density, water-current, and higher-level features constructed out of the lower-level ones used in our experiments.

4 Future Work and Conclusion

This paper presents a task designed to closely model the Piagetian floating task, and then shows how the behaviors of stage-1 children can be explained as feature selection over simple representations in the CLARION cognitive architecture. Future work will attempt to explain the sequence of behaviors shown by Piaget in the floating task. For example, the ability to consider multiple properties at once (which appears in stage-3 children) may be explained using a template structure designed to group properties together. Likewise, the shift from single-place predicates to relations seen in stage 4 might be explained by a stabilization of the property groupings and the emergence of two-place predicates (a similar strategy is used in the DORA model [5]).

Another series of tasks, highly relevant to the study of the development of causality in children, may be interesting to examine using the model developed in this paper. These are the series of "collision" tasks [10, 11], in which infants can identify when some basic notions of physical causality are violated. When shown two objects that are about to collide, but one of them unexpectedly changes direction or stops before the collision is supposed to have taken place, infants will stare at the anomalously behaving object longer than they would at objects colliding normally. We have already started creating this task in PAGI World and hope to show that the present model can match the performance of human children closely.

The backpropagation used in this paper for feature selection is one of many ways CLARION can select features. In the future, as we tackle more complex tasks, we can make use of, e.g., principal component analysis (PCA) or sparse autoencoders [7].

Causality, of course, is an immensely complex and well-studied topic, and early steps such as those taken in this paper can only hope to scratch the surface. Future work will expand the philosophical, psychological, and historical perspectives on the notion of causality and how it relates to explanation generation.[2]

5 Acknowledgements

This work was funded by Office of Naval Research (ONR) grant number N000141310342.

[2] The floating task presented in this paper is available for download, along with PAGI World, at the website:

`http://rair.cogsci.rpi.edu/projects/pagi-world/pagi-world-tasks/`

We encourage researchers to test their particular cognitive architectures or systems on this and other tasks, and report their results.

References

1. Anderson, J.R., Simon, H.A., Reder, L.M.: Radical Constructivism and Cognitive Psychology. In: Ravitch, D. (ed.) Brookings Papers on Education Policy. Brookings Institute Press, Washington, DC (1998)
2. Atkin, K., Licato, J., Bringsjord, S.: Modeling Interoperability Between a Reflex and Reasoning System in a Physical Simulation Environment. In: Proceedings of the 2015 Spring Simulation Multi-Conference (2015)
3. Beilin, H.: Piaget's Enduring Contribution to Developmental Psychology. Developmental Psychology 28(2), 191–204 (1992)
4. Chapman, M.: Constructive Evolution: Origins and Development of Piaget's Thought. Cambridge Univ Press (1988)
5. Doumas, L.A., Hummel, J.E., Sandhofer, C.: A Theory of the Discovery and Predication of Relational Concepts. Psychological Review 115(1), 1–43 (2008)
6. Friedman, S.E., Forbus, K.: An Integrated Systems Approach to Explanation-Based Conceptual Change. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, GA (2010)
7. Gregor, K., LeCun, Y.: Learning Fast Approximations of Sparse Coding. In: Proceedings of the 27th International Conference on Machine Learning. pp. 399–406 (2010)
8. Hummel, J.E., Landy, D.H.: From Analogy to Explanation: Relaxing the 1:1 Mapping Constraint...Very Carefully. In: Kokinov, B., Holyoak, K.J., Gentner, D. (eds.) New Frontiers in Analogy Research: Proceedings of the Second International Conference on Analogy. Sofia, Bulgaria (2009)
9. Hummel, J.E., Licato, J., Bringsjord, S.: Analogy, Explanation, and Proof. Frontiers in Human Neuroscience 8(867) (2014)
10. Leslie, A.M.: Spatiotemporal Continuity and the Perception of Causality in Infants. Perception 13, 287–305 (1984)
11. Leslie, A.M., Keeble, S.: Do Six-Month-Old Infants Perceive Causality? Cognition 25, 265–288 (1987)
12. Licato, J.: Analogical Constructivism: The Emergence of Reasoning Through Analogy and Action Schemas. Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY (May 2015)
13. Licato, J., Sun, R., Bringsjord, S.: Structural Representation and Reasoning in a Hybrid Cognitive Architecture. In: Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN) (2014)
14. Licato, J., Sun, R., Bringsjord, S.: Using Meta-Cognition for Regulating Explanatory Quality Through a Cognitive Architecture. In: Proceedings of the 2nd International Workshop on Artificial Intelligence and Cognition. Turin, Italy (2014)
15. Lourenço, O., Machado, A.: In Defense of Piaget's Theory: A Reply to 10 Common Criticisms. Psychological Review 103(1), 143–164 (1996)
16. Marton, N., Licato, J., Bringsjord, S.: Creating and Reasoning Over Scene Descriptions in a Physically Realistic Simulation. In: Proceedings of the 2015 Spring Simulation Multi-Conference (2015)
17. Piaget, J.: The Child's Conception of Physical Causality. Routledge (1930/1999)
18. Piaget, J.: The Moral Judgment of the Child (1960)
19. Sun, R.: Duality of the Mind: A Bottom Up Approach Toward Cognition. Lawrence Erlbaum Associates, Mahwah, NJ (2002)
20. Sun, R.: Desiderata for Cognitive Architectures. Philosophical Psychology 17(3), 341–373 (Sep 2004), `http://www.informaworld.com/openurl?`

`genre=article\&doi=10.1080/0951508042000286721\&magic=crossref|`
`|D404A21C5BB053405B1A640AFFD44AE3`

21. Sun, R., Peterson, T.: Autonomous Learning of Sequential Tasks: Experiments and Analyses. IEEE Transactions on Neural Networks 9(6), 1217–1234 (November 1998)
22. Sun, R., Zhang, X.: Accounting for Similarity-Based Reasoning within a Cognitive Architecture. In: Proceedings of the 26th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates (2004)