A Grounded Mental Model of Physical Systems: A Modular Connectionist Architecture

Amit Almor
Department of Cognitive and Linguistic Sciences
Brown University
almor@drew.cog.brown.edu

Abstract

Some basic characteristics of subjects' use of mental models of physical systems are discussed. Many representations for physical knowledge suggested so far, including qualitative-reasoning-based models, do not account for these experimental findings. This paper presents a connectionist architecture which suggests an explanation of these experimental results. Two simulation experiments are described which demonstrate how mental models of physical systems may evolve and why grounding symbols used by a mental model to a quantitative representation is necessary.

Mental Models of Physical Systems

Recent studies of physical knowledge acquisition have focused on the way a mental model of a physical system can be created from a set of elementary pieces of knowledge about the physical world. In this context, mental model means a structured representation of knowledge about a specific system. Norman (1983) observes the following facts. (1) Mental models evolve through interaction with the system they model. (2) Mental models are used to facilitate the interaction between the subject and the physical system, and are not accurate descriptions of the physical system. (3) Mental models are runnable, i.e. subjects can run their mental models and predict a particular future state of the system. (4) People are notoriously bad in running mental models through a large number of stages or for a long time. Also, people are often hesitant about the validity of their mental-model-based judgements. All these characteristics relate to the performatory aspect of mental models, that is to the actual behavior of subjects in experiments in which, presumably, they use their mental models. Most research in this domain has focused on the form of knowledge representation which gives rise to these behavioral patterns.

Qualitative Reasoning Theory

The qualitative reasoning theory (Weld & deKleer, 1990) evolved out of research into mental models of physical systems. Often, when subjects apply their physical knowledge, their behavior and reports are incompatible with any theoretical law of physics. Thus, an alternate, simpler "qualitative" physics theory has been formalized. A qualitative-reasoning based mental model is a list of qualitative equations describing the physical system. The qualitative equation is an expression describing the interaction between coarse-valued variables. Special qualitative arithmetic is defined to operate on these qualitative values. Expressed this way, some physical concepts e.g. "flow" can be expressed by specifying their interactions with other concepts such as height of liquid columns (deKleer & Brown, 1990).

Difficulties. Critics of the symbolic paradigm however, claim that a qualitative-reasoning-based mental model, is not a satisfactory model for any cognitive process since it gives rise to the symbol grounding problem (Harnad, 1990). Symbols cannot be arbitrary forms which are assigned meanings independently of the cognitive model. Rather, their form must be causally determined in a bottom up manner.

A further difficulty with symbolic knowledge representation is its artificial distinction between competence and performance. The theory of qualitative reasoning does not account for how mental models evolve through interaction, why mental models are runnable, and why subjects are so bad in running them over many stages. The symbolic framework excludes these confounds from any discussion about the knowledge representation form. An alternative framework, under which both competence and performance confounds will be explained by the postulated knowledge representation form should be preferred on the grounds of parsimony.

This paper presents a modular connectionist architecture for mental models of physical systems which allows the transition from quantitative to qualitative knowledge, and which avoids the problems described above. The architecture generates symbols which are assigned "real-world meanings" as a natural and necessary quality of the processes by which they evolve. Relations between the generated symbols constitute an alternative to the symbolic notion of compositional structure (Fodor and Pylyshyn, 1988).

The distinction between competence and performance is eliminated by using a connectionist

knowledge representation form. (1) Due to the connectionist training process, the representation gradually evolves through interaction with the environment. (2) By imprinting the behavior of the physical system in a connectionist network, a model can be "rerun" later in order to make predictions on the system's future state. (3) The statistical nature of the knowledge representation built makes it hard to run the model for many stages or for a long period as errors propagate and accumulate quickly.

I further describe two simulation experiments with the architecture, which lead to a number of interesting observations. In the current model, adequate symbol generation is possible only if the system has reached some level of familiarity with the real environment. Once this level of familiarity is attained, improving the system's knowledge of the environment is faster using the generated symbols than by increasing the system's familiarity with the environment. The question arises as to the computational status of symbols. First, "grounding" the symbols is no longer a mere philosophical requirement. Rather, it is a computational requirement in order for symbols to be functional. Second, the role of symbols might be conceived as efficient knowledge modifiers rather than arbitrary shapes used as building blocks for some compositional structure.

The Modular Architecture: From Quantitative to Qualitative

The proposed architecture consists of two inter-related modules. The first interacts with the environment to construct a non-symbolic mental model of it. The second uses the internal analog representations built by the first module and associates qualitative symbols with these representations (see Figure 1). The entire model is then able to make qualitative statements and predictions about the state of the environment, given any qualitative specification of an initial scenario.

The Quantitative Module

The first module is a feed-forward three-layered network which is exposed to a representation of the environment (The exact form will be discussed in the next section). The input consists of a representation of the state of the environment at time *t*. The expected output is a representation of the state of the environment at time *t+1* (See Figure 2). The module is trained using the error back-propagation rule (Rumelhart, Hinton, & Williams, 1986).

There are two important facts regarding the coupling of this module with the environment. First, even though (technically speaking,) back-propagation is a supervised training scheme, in this case, with the environment supplying both the input and the correct

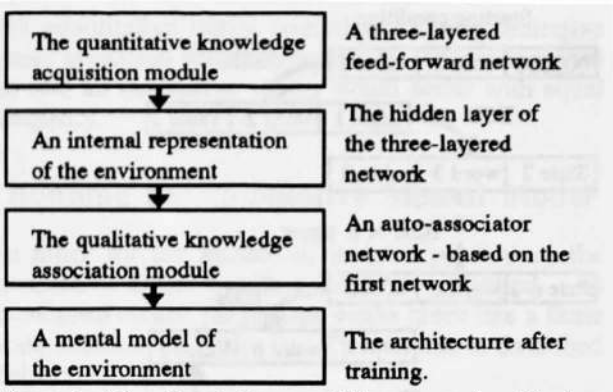

Figure 1: *A functional diagram of the proposed architecture*

output, the training is teacher-less and thus psychologically plausible. Second, since the forms of the input and the output of this network are identical, the network should be viewed as a recurrent network with one input/output layer and one hidden layer. For computational simplicity, the network is trained as a three-layers feed-forward network.

The motivation for using this particular architecture for the first module is twofold. First, a three layered feed-forward network trained by error back-propagation is capable of learning complex interactions in the environment. Second, it allows for the generation of an internal representation of the environment over the hidden layer which already encompasses some information about what the next state of the environment will be. This internal representation is then available for further processing.

The Qualitative Module

The second module auto-associates verbal labels and qualitative values with activation patterns over the hidden layer of the first model. As demonstrated in the next section, the labels and qualitative values do not have to correspond to explicit representations in the input for the first module. This module consists of a recurrent network trained using the Widrow-Hoff (1960) learning rule. An expansion of the "Brain State in a Box" (BSB) algorithm (Anderson, Silverstein, Ritz and

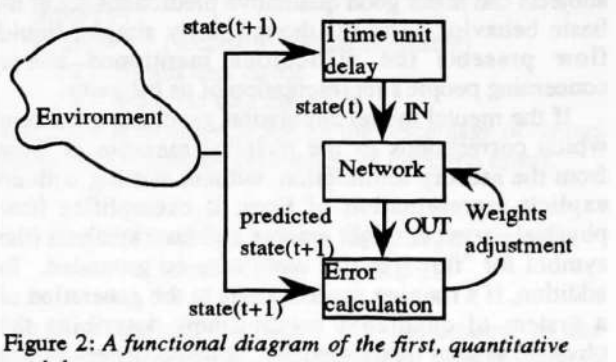

Figure 2: *A functional diagram of the first, quantitative module.*

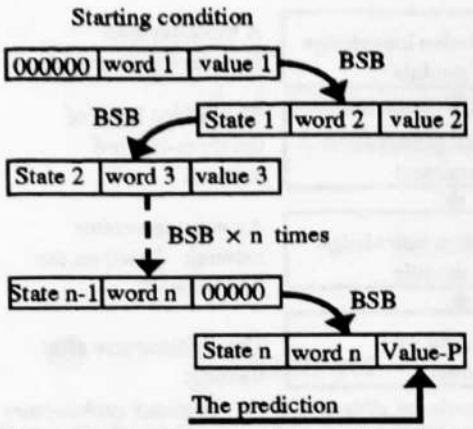

Figure 3: *A diagram of the full activation cycle of the proposed architecture. The first component of the state vector is used as a "state" descriptor or a memory cell. The second and third components have the word and a qualitative value representations. The flow prediction is eventually extracted from the last 40 units of the last state vector.*

Randal, 1976) is used as the network activation scheme. The proposed architecture employs the basins of attraction of the BSB model to achieve qualitative linguistic judgements. (See Hopfield, 1982; Anderson, Silverstein, Ritz and Randal 1976; and Golden, 1986; for a formal analysis of the effect of basins of attraction).

The manner in which the entire architecture functions is similar to a finite-state-automaton where the "internal representation" component of the activation vector functions as the "state." A word, and possibly a qualitative value, is the input which allows transition from the current state to the next state using the BSB dynamics (see Figure 3). The following sections describe a low-scale implementation of the architecture for modelling the generation of mental models of physical systems.

Simulation Experiments

The architecture was used to construct a mental model of liquid flow between reservoirs (See Figure 4). Liquid flow was chosen because: (1) it is familiar, subjects can make good qualitative predictions about its basic behavior; and (2) though fairly simple, liquid flow presents the difficulties mentioned above concerning people's representation of its behavior.

If the mental model simulation generates a concept which corresponds to the physical measure of flow from the sensory information, without starting with an explicit representation of flow, it exemplifies how physical concepts might emerge and how symbols (the symbol for "flow" in this case) may be grounded. In addition, if a training process leads to the generation of a system of qualitative relationships describing the physical system being modeled, it demonstrates how a qualitative-reasoning based mental model could arise.

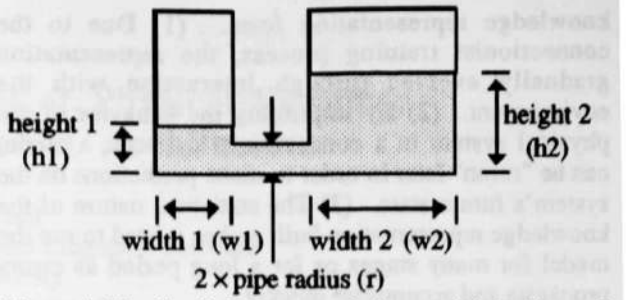

Figure 4: *The liquid flow system being modeled. The real, continuous time dimension across which the process of flow occurs is divided into digital units at which the state of the physical system is sampled.*
Q_t-The flow-rate at time t, is given by Torriceli's law:

$$Q_t = 2\pi r^2 \sqrt{2g|h1_t - h2_t|}$$

where h1 and h2 are the heights of the liquid column in the two reservoirs and r is the radius of the pipe connecting them. Therefore, the heights of the liquid level after a single time unit will be:

if h1>h2:	*if h1<h2:*	*if h1=h2:*
$h1_{t+1}=h1_t-Q_t/\pi w1$	$h1_{t+1}=h1_t+Q_t/\pi w1$	$h1_{t+1}=h1_t$
$h2_{t+1}=h2_t+Q_t/\pi w2$	$h2_{t+1}=h2_t-Q_t/\pi w2$	$h2_{t+1}=h2_t$

where w1 and w2 are the widths of the two reservoirs.

Input and Output Representation and Training Order

The first module consisted of a three layered network with 200 input units, 80 hidden layer units and 80 output units. The second module consisted of 160 fully connected units (See Figure 5).

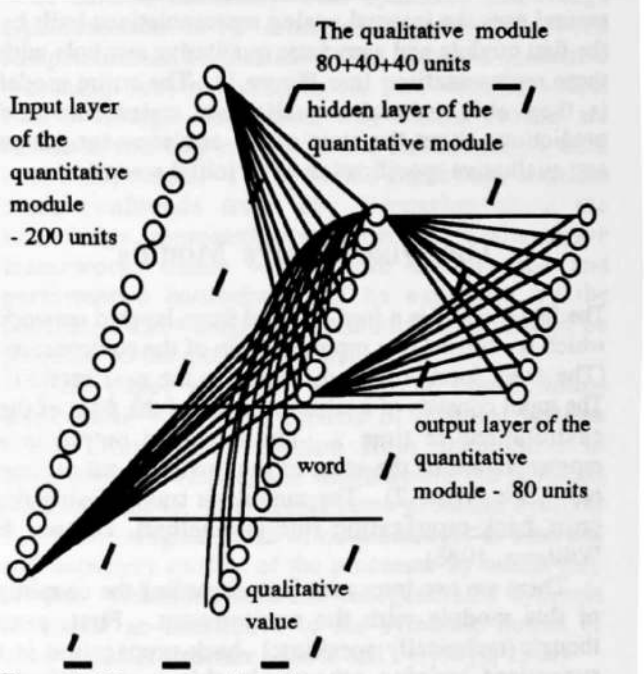

Figure 5: *The proposed architecture*

The input to the first module consisted of quantitative representations of the heights of the water columns in the two reservoirs, the width of the two reservoirs and the width of the pipe connecting them. Each measure was represented in an area of 40 units in which a sliding bar of five units indicated the value. For example, a water height of 50 in the first reservoir is represented by:

00000000000000000000000000000000011111

The total dimensionality of the input layer was therefore 5x40=200 units. For reasons of simplicity and computational feasibility, the output layer was only 80 units long and had only the two heights of the two water columns. It should be noted that bar code representations have some biological appeal since they resemble brain maps that have been described in different areas of the brains of many species. It is therefore plausible to assume bar-code representations as a general quantitative representation form.

Given any quantitative initial condition, this module predicts the quantitative condition at the next time unit. The training phase included exposing this module to 2000 different flow scenarios; each scenario begins by choosing random values for the five input measures and then uses Toricelli's law to calculate the states of the system in the following time units until equilibrium is achieved.

The input to the second module consisted of the internal representation of the physical system's state generated over the hidden layer of the first, quantitative, module, plus an arbitrary representation for a word and a qualitative value. Since the word and the qualitative value are arbitrary symbols (grounded by training this module), they were represented by arbitrary representations that maximize orthogonality. The hidden layer of the first model was 80 units long. The word and qualitative value were each represented by 40 units. Therefore, this module consisted of 160 fully connected units.

The words represented were: "HEIGHT-1", "HEIGHT-2", "WIDTH-1", "WIDTH-2", "WIDTH-PIPE" and "FLOW". The qualitative values represented for all the words but "FLOW", were "HIGH", "MEDIUM" and "LOW". For the word "FLOW", the value field was segmented into two parts, the first representing the qualitative strength of the flow: "HIGH", "MEDIUM" or "LOW" and the second representing the direction of the flow: "FROM-1-TO-2" or "FROM-2-TO-1" or "NONE". For flow value of "NONE", the magnitude field was ignored.

Training the second module started by presenting each one of the 2000 initial states used to train the first module to the first module. Then, the internal representation generated over the hidden layer of the first module was auto-associated with each of the six words and with the qualitative value corresponding to the quantitative initial condition. The quantitative range associated with each qualitative value was chosen so that all qualitative values would occur with equal frequency.

Running the Qualitative Mental Model

In order for the system to make a prediction, the qualitative initial condition must be specified. As mentioned earlier, the process works much like a finite state automaton (Figure 3). The process is described below:

1) Zero the internal representation component of the second, qualitative module.
2) Load the first word and qualitative value of the initial condition into the appropriate areas.
3) Run the BSB activation scheme until saturation.
4) Leave the internal representation area as is (since it functions like the state in a FSA) and load the next word and qualitative value specifying the initial conditions.
5) Repeat 3-5 until all the qualitative initial measures have been specified.
6) Zero the qualitative value.
7) Load the representation for the word "FLOW" to the word area.
8) Run the BSB activation scheme until saturation.
9) Take the prediction for the flow direction and magnitude from the qualitative value area (More precisely, the closest qualitative value to whatever is taken out of the qualitative value area in terms of vector cosine).

The evaluation of a prediction takes into account the fact that when an initial state is specified qualitatively, more than one qualitative prediction can be correct. Based on the probability of each prediction given any initial condition, the following evaluation scheme is used:

1) Test the network for predictions for all possible initial condition.
2) Categorize each prediction as either correct (and most probable), second best choice, third best choice or direction error.
3) Assign a grade to the overall performance of the model by:
   Grade = 4 * (%correct - %direction error)
   - 2 * %third best errors
   - 1 * %second best errors.

The grade is mostly affected by the percentage of correct predictions versus the percentage of direction errors.

Experiment 1: The Importance of Grounding Symbols

The first experiment tested the importance of the "grounding knowledge" to the overall performance of the system. While the amount of training put into the

second module (the symbol associator) was kept constant, different types of quantitative internal representations were used.

1) In order to test the importance of the internal representation component in the dynamics of the symbol associator, random internal representations were used rather then real activation patterns over the hidden layer of the trained first module. The weights connecting the units of the first module were assigned randomly, yielding random activation patterns over the hidden layer. The same 2000 exemplars were used for training the second module. If this network is trainable, then the real knowledge of the specific system must have no importance; the model would work due to there being an internal representation component that "anchors" the symbols arbitrarily, regardless of their contents. Such a result would suggest that this model does not offer any advantage over symbolic architecture because it is indifferent to the meaning assigned to the symbols.

2) Even if the random internal representations proves unsuccessful, it is still possible that the system can function with the real internal representations due to the general structure present in these representations rather than their particular contents. That is, the presence of structure might be sufficient to "anchor" the symbols. In order to rule out this possibility, another version of the first module was used. As before, the first module was not trained but was assigned weights in a structured pattern yielding structured and systematic activation patterns over the hidden layer which were still non related to the real flow system[1]. Again, if this network is trainable, then the real knowledge of the specific system has no importance.

3) Finally, it needs to be shown that given the appropriate training, the system can work. To clarify the difference between the previous cases and the case of the real training, I used the two arbitrary weights setups (the random and the structured) as initial weights for the first module, and trained it using one introduction of each of the 2000 flow scenarios. The training of the qualitative module started only after the quantitative knowledge was generated. This training was done by associating the appropriate words and values with the 2000 initial states. Each of the associations was used 10 times on average during the training.

4) To further explore the importance of the qualitative grounding knowledge, the previous cases were replicated with the exception that the first module was further trained by using the same training set once more before starting the training of the second module.

Over all there were two control systems in which there was no grounding knowledge, two systems in which there was a certain amount of grounding knowledge, and two other systems in which there was more grounding knowledge. The training of the qualitative module was identical for the six systems. The qualitative performance of each system was then evaluated by the scheme described above and assigned a grade.

Results of Experiment 1: The results are shown in Table 1. Although in no cases were the predictions made by the system perfect, the results still suggest the following points: (1) Real grounding knowledge is necessary for better qualitative performance. (2) Arbitrary structure of the internal representations is not sufficient for qualitative knowledge generation. (3) Better grounding knowledge consistently yields better qualitative results. (4) Symbols (such as the word "FLOW" and its associated qualitative values) can serve to generate novel concepts from internal representations of "sensory inputs." The relation between the novel concept and the concepts associated with the "sensory input" is the alternative this framework offers to the notion of "compositionality" in the symbolic framework.

Initial weights	Control groups	Single quantitative training cycle	Two quantitative training cycles
random	-105	-3	2
structured	-105	-29	8

Table 1: *Results of experiment 1. The untrained control groups did the worst. Two quantitative training cycles improved the performance meaning that "stronger grounding", improved the overall results.*

Experiment 2: The Importance of Symbols for Teaching

This experiment examines how the system's predictions can be improved. One method is to further train the first module; i.e., let the system "watch" more flow scenarios. An alternative method would be to retrain the second, qualitative module if it failed to make the correct prediction about scenarios which were part of its training set; i.e., "tell" the system more about how flow behaves qualitatively. In this method, the first module is not retrained. The last method I consider is to only correct the qualitative errors the system does in the evaluation test. This is much like the manner in which a teacher would qualitatively test a student on novel situations and correct his/her errors.

I also wanted to inspect the effect of "grounding knowledge" on the ability to improve the predictions made by the system, by making qualitative corrections. Twelve systems were compared in this experiment. Four were the systems from the previous experiment which were grounded to the real physical system. For each of these four systems, the two qualitative

[1] The weights were setup according to a Gausian formula to ensure that systematic changes in the input values would yield systematic changes in the activation patterns over the hidden layer.

correcting procedures were applied separately. The twelve networks were evaluated as before.

	Initial weights	Single qualitative training cycle	qualitative correction with the training set	qualitative correction with the complete evaluation
Single quantitative training cycle	random	-3	-105	-105
	structured	-29	-116	-116
Two quantitative training cycles	random	2	7	7
	structured	8	7	16

Table 2: *Results of experiment 2. Both methods of qualitative correction worsened the overall performance of the less grounded networks but slightly improved the overall performance of the more grounded networks.*

Results of Experiment 2: The results are shown in Table 2. There are two interesting observations: (1) qualitative correction does not improve performance for the less grounded systems. On the contrary, it reduces total performance[2]. Grounding is not only necessary for overall performance but also for making qualitative corrections. (2) In most cases, with some degree of grounding established, the more efficient qualitative correction methods enable further learning beyond that achieved by quantitative retraining.

General Discussion

The paper sketches a general connectionist architecture that remedies the symbol grounding problem without giving up the notion of symbols. The generation of the novel concept of flow, by associating symbols to an internal representation of simpler interacting factors, demonstrates how a compositional or hierarchical conceptual structure might evolve. The connectionist modelling techniques eliminate the artificial distinction between competence and performance that prevails in much of the research on mental models. The proposed architecture gives a unified account for both the form of the knowledge representation, and for the empirical evidence about how subjects perform tasks using this knowledge.

The generation of a mental model of a the liquid flow physical system demonstrates the two essential aspects of the symbols suggested in this paper. On the one hand, symbols need to be grounded to real-world meanings for the model to work. On the other hand, once this grounding condition is satisfied, using symbols has some evident advantages. These results suggest a wider interpretation of the symbol grounding problem. Not only do symbols need to be grounded to explain real-world meaning assignment, but their grounding is a necessary computational condition. The grounding is what causally determines the compositionality (Fodor and Pylyshyn, 1986) of the symbols. Obviously, this assumption is valid only in the framework sketched in this paper. The computational necessity for grounding, however, may contribute to the failure of the symbolic architecture to meet Turing's (1950) vision.

[2] Because the correction process is used only for wrong predictions, it can cause the network to forget predictions that it previously got right, therefore decreasing the overall grade.

References

Anderson, J. A.; Silverstein, J. W.; Ritz, S. A.; and Randal, J. S., 1976. Distinctive Features, Categorical Perception, and Probability Learning: Some Applications of a Neural Model. *Psychological Review* 84: 413-451.

Fodor, J. A., and Pylyshyn, Z. W., 1988. Connectionism and Cognitive Architecture: A Critical Analysis. In: *Connections and Symbols*. Pinker, S., and Mehler, J. eds. 3-72. Cambridge, MA: The MIT Press.

Golden, R. M., 1986. The 'Brain-State-in-a-Box' Neural Model is a Gradient Descent Algorithm. *Journal of Mathematical Psychology* 30:73-80.

Harnad, S., 1990. The Symbol Grounding Problem. *Physica D*. 42:335-346.

Hopfield, J. J., 1982. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of the National Academy of Sciences 79:2554-2558.

Norman, D. A., 1983. Some Observations on Mental Models. In: *Mental Models*. Gentner, D. & Stevens, A.S. eds. Lawrence Erlbaum Associates, Publishers.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J., 1986. Learning Internal Representations by Error Propagation. In: *PDP: Explorations in the MicroStructure of Cognition*, Vol I.:318-362. Cambridge, MA: The MIT Press.

Searle, J. R., 1980. Minds, Brains, and Programs. *The Behavioral and Brain Sciences* 3:417-424.

Turing, A. M., 1950. Computing Machinery and Intelligence. *Mind* 59:433-460.

Weld, D.S., and deKleer, J. eds. 1990. *Readings in Qualitative Reasoning about Physical Systems*. Morgan Kaufman Publishers, INC.

Widrow, B., and Hoff, M. E., 1960. Adaptive Switching Circuits. 1960 IRE WESCON Convention Record, New York: IRE. 96-104.