Connectionist Synthetic Epistemology: Requirements for the Development of Objectivity

Ron Chrisley & Andy Holland
School of Cognitive & Computing Sciences
University of Sussex

November 29, 1994

Abstract

A connectionist system that is capable of learning about the spatial structure of a simple world is used for the purposes of *synthetic epistemology*: the creation and analysis of artificial systems in order to clarify philosophical issues that arise in the explanation of how agents, both natural and artificial, represent the world. In this case, the issues to be clarified focus on the content of representational states that exist prior to a fully objective understanding of a spatial domain. In particular, the criticisms of (Chrisley, 1993) that were raised in (Holland, 1994) are addressed: how can we determine that a system's spatial representations are more objective than before? And under what conditions (tasks, training regimes, environments) do such increases in objectivity occur? After analysing the results of experiments that attempt to shed light on these questions, the study concludes by comparing and contrasting this work with related research.

1 Synthetic epistemology: Philosophy and AI/ALife

Sometimes in order to clarify the theories and concepts one would like to use to explain a natural system, it can be of great assistance to try them out on a simple, artificial system, which allows greater control and clearer analysis. Just as one might more readily come to a clear understanding of the principles of aerodynamics by studying a simple, artificial glider than by studying the particularities of the feathers and muscles of sparrows, so one might also see more readily the general structure of a proper psychology of real systems by first attempting to apply it to a simple, artificial agent.

Thus, to clarify some new ideas being proposed for the explanation of natural intentional systems, it seems a promising idea to turn to *synthetic epistemology*: the creation and analysis of artificial systems in order to clarify philosophical issues that arise in the explanation of how agents, both natural and artificial, represent the world.

Synthesis can thus be justified as an approach to understanding epistemology in the same way that it can be justified as an approach to understanding intelligence (AI), or biology (ALife):

> Artificial systems which exhibit lifelike behaviors are worthy of investigation on their own rights, whether or not we think that the processes they mimic have played a role in the development or mechanics of life as *we* know it to be. Such systems ... expand our understanding of life as it *could* be. By allowing us to view the life that has evolved here on earth in the larger context of *possible* life, we may begin to derive

> a truly general theoretical biology capable of making universal statements about life wherever it may be found and whatever it may be made of. (Langton, 1989, p.*xvi*, original emphasis).

The specific epistemological issue which this research addresses is understanding the nature of, and mechanisms underlying, the transition from heavily perspective-dependent to more objective modes of representation. Some ways of representing the world are objective or near-objective, some are not. A way of representing some aspect of the world is objective if, e.g., it presents that aspect of the world as something that could exist while unperceived. Pre-objectivity involves representing the world, but not *as* the world, not as something that is or can be independent of the subject. Another important aspect of ways of representing is that they can be *more or less* objective.

There are several reasons (Chrisley, 1993; Cussins, 1990) for thinking that a connectionist architecture is much more suited than traditional symbolic architectures to the investigation of the development from less objective to more objective cognition. Furthermore, the acquisition of more and more sophisticated navigational abilities is plausibly seen as a paradigmatic case of the move from perspective-dependent to more perspective-independent ways of representing (Cussins, 1990). Thus, the Connectionist Navigational Map (*CNM*; (Chrisley, 1990)) was developed for these purposes.

2 The Connectionist navigational map

This section reviews the CNM architecture, environment, and learning regime; those already familiar with these details may skip to section 3.

2.1 CNM architecture

The *Connectionist Navigational Map* is a computational architecture being developed with the aim of providing an autonomous robot with the ability to learn and use spatial maps for navigation. One component of this architecture, the *predictive map*, allows the robot to predict what sensations it would have if it were to move in a particular ego-centrically specified manner (e.g "rotate $\pi/4$ radians to the right", "move forward 10 feet"). Of course, this requires the robot to have some kind of representation of its current location, since, in general, the mapping from actions to sensations is dependent upon where one is in the world. That is, the mapping from sensations and actions to sensations is one-to-many, since more than one place can have any given sensory signature. Thus, the spatial environment, and therefore a model of it, can be seen instead as a function from current location and current action to predicted sensations. The input consists of a state representation, or *location code*, corresponding to the current location a of the robot, and an action representation representing the move m being made. The output of the network is a vector that is supposed to be equal to the sensation vector the robot would receive from its senses if it were actually at the place that is reached by making the move m at location a.

Of course, there is more structure to space than a simple, direct mapping from locations and actions to sensations indicates (see figure 1).

Specifically, location and action determine a new location, which itself determines the sensations of the robot. Thus, it might be easier for a robot to learn (or a theorist to analyse) a predictive map if its structure reflects this regularity of the spatial environment. The predictive map of the CNM is a composition of two mappings: a topological mapping T (from locations and actions to

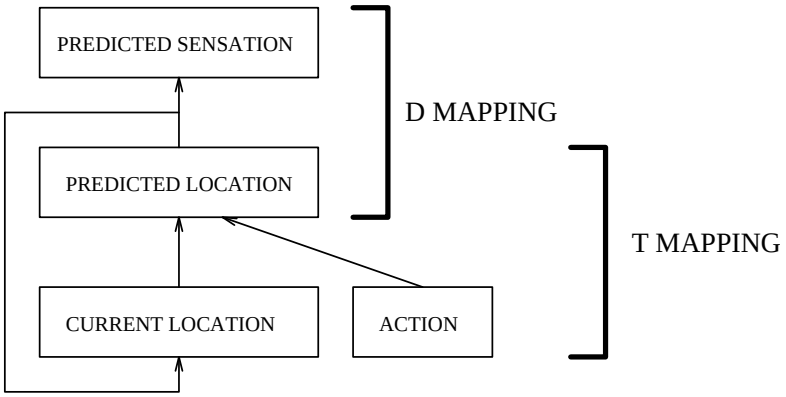

Figure 1: The PDP architecture of the predictive map ($locations \times actions \mapsto sensations$) formed by composing a topological mapping T ($locations \times actions \mapsto locations$) with a descriptive mapping D ($locations \mapsto sensations$). Arrows indicate directed, full inter-connection between layers of units.

states) and a descriptive mapping D (from locations to sensations). In actual use, the location output of the T mapping, after a given action, is used as the location input to the T mapping for the next action.

Thus, if a constantly north-facing robot considers moving forward and then moving right, it can use the map to predict what sensations it would have after those moves by calculating $D(T(T(a,$**move-north**$),$ **move-east**$))$, where a is a location representation corresponding to the robot's initial location before the actions, and **move-north** and **move-east** are action representations with the intuitive interpretation.

Given the iterative nature of the T mapping, the predictive map must be a recurrent network; in the experiments discussed here, it is implemented as a *simple* recurrent network (Elman, 1990).

2.2 The experimental setup

The experimental situation used here (roughly the same as the one used in (Chrisley, 1993) and (Holland, 1994)) is a deliberately impoverished one: a developing (learning) agent moving through a simulated "grid world"; the part of the world simulated has only 81 cells or locations (9 by 9). Each location has a 4-bit vector associated with it, which can be understood to be the sensations the agent has when at that location (see figure 4, in section 7 below). As in its normal use, the CNM is to provide a means for this agent to improve its navigation of its space (and thus increase the objectivity of its ways of representing that space) through sensory prediction.

The agent has eight actions available at any location, those of moving into each of the adjacent locations (orientation is not modeled: the agent can be thought of as always facing north).

It is assumed that the developing agent has somehow managed to conduct a journey that starts and ends at some privileged location, called *home*. The developing agent stores the sequence of actions taken and the sensations that result. Note that sometimes the same action yields different sensations, and that different actions sometimes result in the same sensations.

2.3 Learning regime

When the agent returns home, it iteratively learns the route, not by actually moving, but by reviewing the remembered route in the following manner:

- First, generate a training set:
    1. Assume some arbitrary representation or code for the initial location ("home"). Store this code, and the code for the first action taken, as an input pattern; store the sensations that were observed after taking that action as the target output for that pattern.
    2. Propagate the current input pattern into the T mapping.
    3. Use the output of the T mapping, along with the next action on the remembered route, as the next current input pattern. Store this pattern into the training list as an input pattern, as well as storing the next remembered sensation as the target output pattern for that input.
    4. Go to 2 until finished with the remembered route.
- Then, learn with the current training set: adjust the weights of the T and D mappings according to the gradient of the error (difference between target and actual outputs); i.e., use the backpropagation learning algorithm (Rumelhart et al., 1986).
- After some period of learning with one training set (in the simulations described, the time was 6 epochs) , a new training set is created, in the same manner as before (steps 1-4), and the agent trains with the new set.

In the simulation used for the experiments which follow, this 6 epoch cycle was repeated until the network had learned the route. That is, starting with the code for home and the initial action taken, the T mapping would produce a new code that not only yielded the correct predicted sensations via the D mapping, but which also, in conjunction with the representation for the next action taken, produced a code via the T mapping which could itself yield both the right sensation vector and location code, and so on, iteratively.

The T mapping was realized by a network which comprised 9 inputs (5 for the location code, and 4 for the action code; action codes were the 4 unit vectors for north, south, east and west, or the sum of the relevant unit vectors for the other 4 directions: NE, SE, SW and NW), 6 hidden units, and 5 outputs (a location code). The D mapping was implemented by a network with 5 inputs for a location code (in practice, this was the same as the output layer of the T mapping), and 4 outputs for a sensation vector at that location. The network was simply recurrent; backpropagation through time was not necessary.

3 Simplicity as a virtue: arguments for and against the CNM approach

It might be thought that this grid "world" is too impoverished to be of much interest. In particular, it might be thought that there are too few states, and that the "sensory properties" of each place are too coarse-grained, for this work to be of any relevance. There are several reasons why we disagree.

First, this work can be seen as an extension of the research in learning finite state machines or formal grammars, e.g. (Cleeremans, 1992) and (Dienes, 1994). The finite state machines in that

work typically involve *fewer* states with only one or two transitions possible from or to a state, and have no notion of the sensory properties of a state that may be shared with another state, and be sensed by the agent. We believe that by adding the complexity found in the CNM world, one begins to justify talk of learning spatial representations, instead of mere arbitrary grammars. But even if that assumption is illicit, the CNM paradigm should still be valuable, at least within the finite state machine learning paradigm.

Furthermore, coarse-grained sensations actually support the intended spatial interpretation of the CNM's activity. Since there are so few (i.e., 16) types of sensory properties a location might have, the CNM cannot rely, in achieving its predictive aims, on merely recording the superficial sensory contingencies, but rather is forced to learn the more abstract spatial structure of its environment. To exaggerate this effect, we did not even let the CNM use the current sensations as an input to its predictive map, but rather forced it to use only its own representations.

It could still be objected that this, too, is unlike human cognition. It could be claimed that the way that humans and other animals achieve most of their navigation is by learning associations between actual detailed sensations, and not by developing some more abstract topological representation. That is, organisms predict what comes next by *looking* and seeing where they are.[1]

It would be a mistake to think that because we are interested in understanding how cognizers are able to make transitions from less objective to more objective ways of representing the world, that we somehow think that the majority of cognition involves representations that are at the extreme objective end of this scale. In fact, we agree that there are many kinds of cognitive interactions with the world that *require* relatively unsystematic, pre-objective ways of representing, if they involve any representation at all. Furthermore, it may be impossible for any embodied, finite system to ever achieve total objectivity or total systematicity. Nevertheless, we do think that there are interactions for which the ability to increase systematicity is a cognitive virtue, and spatial navigation is one of these.

In order to pump your intuitions concerning these matters, consider the kinds of mistakes we (and other animals) do and do not make in navigating. Suppose that I leave the lecture theatre between talks at a conference in Skövde. While I am out of the room for a few minutes, the rest of you redecorate the lecture theatre so that it very closely resembles another one, with which I am familiar, in Brighton. You then hide (so as to give no clues about the true location of the theatre), and watch what I will do from behind doors, desks, etc. You might expect to have a good laugh upon seeing my puzzled expression, (the lengths to which some people will go for a gag!), but you would *not* expect me to actually think I have somehow travelled hundreds of miles back to Brighton! Behind my puzzled expression is the thought "Why does this place look so much like the lecture theatre in Brighton all of a sudden?", not the thought "Whoa! How did I cross the North Sea all of a sudden?!"

The objector might not find this story relevant, since what is being denied is that two different locations ever could, other than in the laboratory, ever yield *identical* sensations. On this view, the reason why I am not tricked into thinking I am in Brighton is due to my (perhaps sub-conscious) ability to make very fine sensory discriminations (e.g., I can see that the walls in Skövde have been recently painted to be that beige colour, whereas the paint is much older in the Brighton room).

But surely this is unlikely. It would imply that we would have difficulty recognizing the lecture theatre we are in now as the one that we were in before the break given that, e.g., the overhead projector has been moved slightly. We would be like the mnemonist S., whose eidetic memory made it difficult for him to recognize a face as the same as one seen earlier if the face's expression

[1] Thanks to David Rumelhart for pointing out this objection.

was different (Luria, 1968, reported in Glass & Holyoak, 1986, p. 330). But we are not typically like that. Usually, one can recognize a place as being the same, in virtue of its relational properties, even though its intrinsic (sensory) properties have changed considerably since one's last visit. This cannot be explained by a model that does not allow for some topological, spatial representation, in addition to sensory association.

One might wonder why we are demanding that the CNM *learn* its spatial representations. Since the structure of space does not change within a creature's lifetime, surely the system of spatial representation could be innate. This may be, but there are two reasons that remain for using the CNM. First, in order to naturalize our systems of mental representation, not only do we have to have a (synchronic) understanding how they can be realized in our current physical structure; we must also have a (diachronic) understanding of how such abilities could be the product of a natural selection process (Cussins, 1992). So if there is no "development of spatial objectivity" story to be told for any individual, then there must be some such story to be told for cognizing *species*. Second (but more concessively), there might be good design reasons for having an adaptive spatial representation system, even if its parameters are initialized at birth to some near-objective value (Chrisley, 1991).

Also, it should be re-emphasized (cf section 1) that the CNM is for *synthetic* epistemology, and is obviously not meant to be a detailed model of the actual mechanisms of spatial learning in any natural system. Rather, it is meant to explore and illustrate some general principles and phenomena that are relevant under certain conditions (e.g., those in which local sensory information is not sufficient to guide navigation).

Nevertheless, we do plan to improve the CNM's "world" in several ways, including making the space continuous; using unit-free, routine-based actions; making the environment dynamic; and eventually using a real, non-simulated robot in a real-world environment. It is hoped that after some more general observations like the ones expressed in this paper, these added degrees of realism will allow us to address more specific issues, in addition to further testing conclusions already drawn on the basis of these simpler "grid world" experiments.

4 Objectivity as a by-product of maximizing predictive success

Because of the feedback inherent in simple recurrent nets, the CNM's representations (location codes) change over time: at any time, a code may be used to represent a location or set of locations different from the location(s) that it is used to represent at a different time; and at any time, a given place may be represented by code(s) that are different from the ones used to represent that place at other times. The idea behind this research is that in some cases this dynamic process may be seen as a developmental one, in which the CNM's codes become more and more objective, conceptual and perspective-independent.

In general, the CNM uses a different location code after each move, even when moving to a place that it has been before. That is the CNM typically uses different location codes on different occasions for the same objective location. Thus, typically, CNM representations are *non-systematic*. For our purposes, systematic [2] representation can be defined as follows:

[2] The terms "systematic" and "systematicity" have already been used in the connectionist literature; we are using it here as a technical term only. That is, we think it is *related* to the other notions of systematicity that have been used, but we are not yet clear on what exactly that connection is. However, this connection need not be made clear in order for the notion to do its work here as a measure of the degree of objectivity which a set of representations exhibit.

Definition: A system represents a location l systematically if there is a representation a such that:

1. whenever the system uses a, or a representation very functionally similar to a, it does so to represent l and not some other location l'; and
2. whenever the system needs to represent l, it is capable of using a, or a representation very functionally similar to a, to do so.

For the case at hand, these requirements boil down to:

> The CNM represents a location l systematically if there is a location code a such that, normally, a is active on the "current location" units if and only if the agent is currently at l.

Often, when speaking about the CNM's representations, we use expressions like "the same representation" or "different representations", when, strictly speaking, there is no such relevant issue of representational *identity*, but rather only representation *similarity*, in particular functional similarity. Thus, the above requirement is that normally all the codes a that the CNM has active on the "current location" units when at A are functionally very similar, and the CNM never has a code b, that is functionally very similar to one of the a, active on the "current location" units when the CNM is at a place other than A. Thus, there are at least three ways in which systematicity is a matter of degree:

1. the greater the number of different ways of getting to the place A that yield a code functionally equivalent to a, the greater the systematicity;
2. the greater the number of ways of getting to places other than A that yield a code functionally equivalent to a, the less the systematicity; and
3. the degree of systematicity will vary with the degree of functional equivalence in the above two conditions.

Therefore, the CNM, like a typical connectionist system, uses analog representations, rather than digital representations which can be *exactly* functionally equivalent. A consequence is that it seems unlikely that the CNM could ever achieve 100% systematicity, since it is a non-linear system, and even slight differences in location codes will, through iteration in the T mapping, most likely result in a large divergence at some point. Even if this is the case, it is not necessarily an argument against the CNM as a cognitive model, since it is not clear that, without external symbol systems, humans can be completely systematic either.

However, in (Chrisley, 1993), it was suggested that if the CNM could develop a *more* systematic representation for a place (i.e., use the same location code, across different contexts, for the same location, and only that location) then the objectivity of its way of representing would be increased dramatically. That ability was observed to be present; in many cases the CNM did develop functionally equivalent codes for the same place as encountered at two different stages along a route (i.e., at a place where the route crossed itself). For a full explanation of the connection between systematicity, generalization, and objectivity, the reader is referred to section 4.4 of that paper; however, a brief explanation can be given here. Specifically, the systematicity that was observed yielded a kind of generalization. Since the codes were functionally equivalent, the predictions/associations the CNM learned for either one of the two codes were "inherited" by the other code. Thus, the CNM would successfully predict what would happen if it were to take

a path it had never taken before. That is, form of spatial generalization occurred; the CNM was shown to be more than just a means of memorising a list of action/sensation sequences.

Such cases of the emergence of systematic codes suggested that the CNM was able to make transitions from less objective to more objective ways of interacting with the world, which, along with the counterpart transitions from more objective to less objective ways of representing, are the operations we take to be at the heart of cognition. However, we have since realized that those results need to be qualified in two important ways.

4.1 Sameness of location vs. mere sameness of sensation

First, recent work (Holland, 1994) has pointed out that one must take care in inferring an increase in objectivity on the basis of the kind of evidence presented in (Chrisley, 1993). Holland reports that the phenomenon of convergence of codes that correspond to the same place can be reproduced very reliably. But he makes an important observation: *the convergence also occurs for location codes that do not correspond to the same place, but merely to places that have, e.g., the same sensory properties.*

This is a consequence of the CNM's non-symbolic form of representation. In (Chrisley, 1993) it was pointed out that a key difference between symbolic and non-symbolic architectures is that in the former, associating a representation a with another representation d does not constrain the class of representations that the system can associate with a different representation, b. However, in non-symbolic architectures like the CNM, mapping a set of sensations to a via the D mapping *does* constrain what D can map to b. For example, if a and b are very similar (but still, say, functionally distinct), it is very difficult for D to map different sensation vectors to a and b. Looked at the other way, if it is a constraint on the codes that the CNM develops that D must map a and b to the same outputs, then the CNM will tend to develop similar codes for a and b.

Thus, it appears that the CNM's observed tendency to develop common codes for the same place encountered in different contexts can, at least in some cases, be explained as just a fortuitous by-product of a more pervasive, and less impressive tendency: to develop common codes for places that have the same descriptive mapping. If so, then this calls into question the appropriateness of the CNM for studying the development of objective representations.

4.2 The functional equivalence of hidden representations

Also, the earlier study places too much emphasis on the actual Euclidean similarity/identity of two location codes. Objectivity does not require that the location codes used to represent the same place in different contexts be themselves the same, or even similar; rather, they need only be *functionally equivalent* [3]. Conversely, the fact that two location codes are, e.g., clustered together in a cluster analysis, does not guarantee that the codes will play the same, or even similar, causal roles in the network. Two codes a and b are said to have a (second-order[4]) functional equivalence of $F_d(a,b) = -\frac{F_2(a,b)+F_1(a,b)}{2}$, where:

[3] The notion of functional equivalence here focuses on the similarity of the *effects* of two codes. If, in addition, one paid attention to the similarity of the *causes* of the two codes, then one might not be able to distinguish identity and functional equivalence. We think (for reasons which we cannot elaborate on here) that it is best *not* to include causal origins in a characterization of the functionality of a representation, so we are therefore compelled to acknowledge the difference between brute vector similarity and functional equivalence.

[4] Obviously, this definition of functional equivalence can be generalized via a recursive definition to nth order functional equivalence (i.e., the negative of the distance between predictions made for locations up to n moves away) for arbitrary n. For our purposes it is sufficient to use only these first few terms for such a generalization, since they dominate the results.

- $F_2(x,y) = \frac{\sum_{m=1}^{A} ||D(T(x,action_m)) - D(T(y,action_m))||}{A}$;
- $F_1(x,y) = ||D(x) - D(y)||$;
- A is the number (in our case 8) of actions available, and $action_m$ is the mth element of the list (N, NE, E, SE, S, SW, W, NW).

This value can be thought of as the negative average distance between corresponding sensory predictions (corresponding to the current and eight surrounding locations) that a and b give rise to.

There are two other ways of measuring functional equivalence that we considered: the percentage F_p of neighbouring output *pattern* predictions that are the same, and the percentage F_b of neighbouring output *bits* that are the same. That is:

- $F_p(x,y) = 100\frac{\sum_{m=1}^{A} P[D(T[x,action_m]),D(T[y,action_m])]}{A}$; and
- $F_b(x,y) = 100\frac{\sum_{m=1}^{A} B[D(T[x,action_m]),D(T[y,action_m)]}{A}$;

where:

- $P(d_{xm}, d_{ym})$ is 1 if the (thresholded) sensory predictions d_{xm} and d_{ym} are equal, and 0 otherwise; and
- $B(d_{xm}, d_{ym})$ is the hamming distance between sensory predictions d_{xm} and d_{ym}.

The former measure is more strict that the latter; two codes a and b may be such that $F_b(a,b) = 75\%$, yet $F_p(a,b) = 0\%$ (i.e., they may always disagree on, say, bit 1 of the 4 possible output pattern bits, but agree on all others). Its advantage is that it better avoids apparent functional equivalences that are actually spurious in that they depend on some accidental similarities (e.g., those that are a product of the contingent distribution of sensory properties in the environment) between the output that a is producing and the output that b is producing. But the latter may be a better measure for some purposes, since agreeing *somewhat*, if not perfectly, on neighboring predictions, indicates some degree of functional equivalence that may be of some explanatory use (especially if the non-sensory properties of a place – presence of food, danger, etc. – are reliably correlated with its sensory properties). In the 75%/0% case, above, it seems that b and a have *some* degree of functional equivalence, even if it is systematically distorted. Thus, all three measures were used in reporting the results of the experiments below.

It may sometimes be useful to acknowledge the fact that two location codes are functionally similar with respect to the actions that the network actually made at those locations, while being functionally divergent with respect to the the remaining actions, which are, in effect, "don't care" values as far as the training regime is concerned. This notion of *relevant* functional divergence is denoted by F^*, and is calculated by only using actions that have actually been taken by the agent in the involved location(s) when summing and normalizing in the first equation above. It is desirable, of course, that $F(x,y)$ be low for any co-referring x and y, but such a situation is not strictly required for the correspondning $F^*(x,y)$ to be low, which would in itself constitute a degree of systematicity. In the experiments reported here, we ignored this weaker notion of functional equivalence, since one of our main interests is in the generalization from what has been explicitly experienced to what has not.

Note that functional equivalence of any of these three kinds is independent of *correctness*: two codes may give rise to the same predictions (and thus have $F_d = 0$ and $F_p = 100\% = F_b$), yet both

may be completely wrong in those predictions. The connection with correctness will be captured in two ways in the experiments that follow: the criterion that the net learn until it correctly predicts all sensations on its route; and the generalization that will naturally result in the cases of high systematicity.

In the experiments that follow, we give an example of the cases that justify the introduction of these functional equivalence measures: cases in which Euclidean distance/clustering would suggest a functional equivalence that is not present, and cases in which Euclidean distance/clustering would suggest functional divergence that is not present. This aspect of the research, then, can have a relatively broad application, even if one is not interested in synthetic epistemology, connectionist navigation or the development of objectivity.

5 A hypothesis concerning the requirements for the development of systematic representation in the CNM

In order to address these issues concerning the requirements for the development of objectivity, a hypothesis was formed concerning the conditions under which this style of representation will arise in the CNM, and experiments have been conducted to test this hypothesis.

Given the definition of systematic representation in section 4, the central hypothesis of this paper can be stated thus:

> **Hypothesis:** The CNM will only develop a systematic representation of a location l if its encounters with l, and with locations that resemble l, are so structured as to make such a form of representation a useful means of minimizing the error of its predictions.

The plausibility of the hypothesis is a consequence of the CNM's non-symbolic form of representation, as discussed in section 4.1. The holistic, as opposed to atomistic, nature of representation in the CNM implies that systematic representation will not be the default. Since what *primarily* determines whether the two location codes used at two different points in a route are similar is the similarity of the sensory predictions that such codes are required to produce (and not the identity of the two locations in question), the CNM will tend to violate the first of the two requirements for systematicity. Thus, it is only *likely* to satisfy the first requirement if its routes through its environment which generate its training regime are structured in particular ways.

The hypothesis itself doesn't have much force without some specifics concerning what kinds of structure the CNM's encounters must have in order to make the hypothesis true. If one prefers, one can rephrase the hypothesis into a question: what kind of spatial behaviours, if any, compel the CNM to form systematic spatial representations?

We attempted to answer this question by considering it for each of the two components of the working definition of systematic representation:

1. under what conditions does the CNM avoid allocating functionally equivalent codes to distinct locations (even though the locations, e.g., have the same description)?; and
2. under what conditions does the CNM succeed in using functionally equivalent codes for the same location in different contexts?

In trying to answer these questions, one major obstacle to systematic representation, already alluded to, must be understood. If the CNM needs, in two different contexts A and B, to produce

the same (or very similar) outputs on the D mapping, then there will be a tendency for it evolve weights such that the codes that are active in those two contexts, a and b, are functionally equivalent, even if the CNM is at different (albeit sensorily similar) locations in those two contexts. Thus there is a tendency to violate the first of the two requirements for systematic representation. In what situations, if any, can this tendency be overcome, such that systematic representations *are* developed?

But this is only one example of how the predictive demands placed on the CNM constrain the kinds of representations used. Another example is that making the same move at two different parts of the route will tend to produce similar codes for the location after those moves. The representational demands of a recurrent network are extremely holistic, with the "optimal" representation for the current situation being determined both by what it will give rise to in the arbitrarily distant future, and by what what gave rise to it in the arbitrarily distant past, in addition to the constraints of the present. Not only does the code that is used for the current location have to be mapped to the current sensations via the D mapping, but it needs to give rise to a code that can lead to the right predictions for the next step in the route, and it needs to be such that it can be the product of inputting the last code and action into the T mapping.

6 Principles & Predictions

To make substantive the hypothesis of the previous section, we used it to make some predictions concerning the conditions under which systematicity would and would not develop.

First, we noted four principles that we take to characterize the holistic interdependence of CNM representations (i.e., the aspects of the CNM that make it non-symbolic, as discussed in sections 4.1 and 5):

1. same inputs tend to produce same outputs
2. different inputs tend to produce different outputs
3. same outputs tend to require same inputs
4. different outputs tend to require different inputs

These are, of course, only rough guides and tendencies, which are defeasible. But in the context of the CNM, we appealed to the above principles to suggest some more concrete tendencies concerning the functional equivalence of location codes that the CNM develops.

We focussed on the case of codes that the CNM develops to represent *sensorily equivalent* places. This is because we are interested in two kinds of case: the divergence between codes that represent sensorily equivalent but spatially distinct places, and the equivalence between codes that represent the same place (which, obviously, must also be sensorily equivalent).

We used the principles to derive the following postulates [5], expectations concerning how the CNM's codes would develop (numbers in brackets indicate which of the principles were used to derive each postulate):

1. $D(a) = D(b) \rightarrow a = b$ [3];
2. $ma_{-1} = mb_{-1} \rightarrow a = b$[1]; $ma_{-1} \neq mb_{-1} \rightarrow a \neq b$ [2];

[5] At least one or two of these postulates seem to have an analogue in (Cleeremans, 1992, pp 64-65) (e.g., postulate 5).

3. $D(a_{-1}) = D(b_{-1}) \rightarrow a = b$[3]; $D(a_{-1}) \neq D(b_{-1}) \rightarrow a \neq b$ [4];
4. $ma = mb \rightarrow a \neq b$[4]; $ma \neq mb \rightarrow a = b$ [3]
5. $D(a_{+1}) = D(b_{+1}) \rightarrow a = b$[3]; $D(a_{+1}) \neq D(b_{+1}) \rightarrow a \neq b$ [4];

where:

- "=" means "similar" for movement and sensation vectors, but means "functionally equivalent" for location codes; and
- "$\rightarrow$" means "tends to make true".

Postulate 4 requires some explanation, since it does not hold unconditionally. In general, the similarity or difference of moves made from a and b has no implication in itself for the functional equivalence of the codes. But it does have implications when interacting with other contexts. In particular, if $D(a_{+1}) = D(b_{+1})$, then $ma \neq mb \rightarrow a \neq b$. This is because differences in a and b will be required in order to cancel out the differences in ma and mb in order to have a constant result.

Conversely, if $D(a_{+1}) \neq D(b_{+1})$, then $ma = mb \rightarrow a \neq b$, by principle 4. To see why, first note that principle 4 implies that $D(a_{+1}) \neq D(b_{+1}) \rightarrow a_{+1} \neq b_{+1}$. Next, note that there will be an even stronger push (via principle 4 again) for $a \neq b$ than there would be based on prediction 5 alone, since the similarity in the moves ma and mb must be compensated for by greater differences in a and b in order to achieve a comparable difference in a_{+1} and b_{+1}. There will be no special tendency produced by $ma \neq mb$.

In stating these tendencies, our use of "=" and "$\neq$" suggests that we are once again assuming either completely equivalent or maximally different description vectors. But in fact, the relevant description and movement vectors may be more or less similar or different. These differences should affect the functional equivalence of the relevant location codes accordingly, but given a random distribution on sensation vectors and moves, we believe these additional modifying factors can be ignored in our analysis.

In light of these postulates, we defined 7 (non-exhaustive) types of route, or scenarios, that we thought might generate a large variation in the degree of systematicity of the representations the the CNM develops for two locations that are sensorily equivalent. The situations are listed in figure 2.

Using the five principles, we predicted the following rough ordering of these situations with respect to the degree of systematicity that they impose on the CNM's representations for the two locations, from most systematic to least:

SIDO These scenarios should yield the best systematicity, since because functional divergence between the codes for different places is fostered by exploring the different sensory surround of the two locations, yet each of the two locations is entered via a constant approach, providing a basis for the development of very similar codes for the same place. Within this group SIDOD should be more systematic than SIDOS, since the differing ways in to the two locations will add the the divergence between their location codes.

DIDO This should be next best with respect to systematicity, because although the lack of a common approach to the locations will yield a divergence between the codes used for the same place, there will be a greater divergence between the codes used for the two different places, due to the exploration of their different sensory surround.

DIDO : Different ways in, different ways out. The route that the CNM takes approaches each places from several different directions, and leaves from each place in several different directions.

SISO : Same way in, same way out. There are four possible sub-cases:

- SS both the single direction in and the single direction out are the same for the two locations
- SD the single way in is the same, but the single directions out are different for the two locations
- DS the single ways in are different, but the single direction out is the same for the two locations
- DD both the single ways in and the single directions out are different for the two locations

DISO : Different ways in, same way out. For each location, the CNM's route approaches from several different directions, but always leaves by the same direction. There are two sub-cases:

- S the single way out is the same for both locations
- D the single way out is different.

SIDO : Same way in, different ways out. For each location, the CNM's route approaches from one direction only, but leaves by several different directions. There are two sub-cases:

- S one in which the single way in is the same for both locations, and
- D one in which the single way in is different.

Figure 2: The classification of routes used in the experiments.

SISO These should yield poor systematicity, due to the lack of exploration of the two locations' different sensory surrounds. However, SISODS and SISODD should be more systematic than SISOSS and SISOSD, since the single moves in are not the same between the two locations, thus causing *some* functional divergence between the codes for the two places. SISODS should be slightly more systematic than SISODD, and SISOSS more than SISOSD, for reasons similar to the ordering given within the SIDO category, above. All of these should be more systematic than the DISO scenarios, since at least the codes for a location are being produced by a common factor: the move in.

DISO DISOS should yield poor systematicity, but DISOD should be even worse, since in DISOS there is at least one basis for forcing a divergence between the codes that represent the two places: the different predictions required of their common move out of those places (the same code cannot produce both 1111 and 1001 when combined with the move North). In DISOD, the single ways out are different for the two locations, and thus there will be no need to develop different codes to accommodate the different predictions (the same code *can* produce 1111 when combined with North and 1001 when combined with E).

Of course, the variables used in calculating these predictions are not all of the ones that are relevant in determining the degree of systematicity of the two representation codes. In particular, we have said nothing about the distribution of description vectors for the places surrounding the two locations in question, yet this will typically have considerable effect. For example, if there were local duplications (if, e.g., the locations surrounding a and b had corresponding description vectors), then postulate 5 would suggest that exploring the sensory surround of a and b will push the two codes together, not cause them to diverge, as was assumed in the rationales for the above predictions. Nevertheless, if one assumes a uniform distribution of sensation vectors, the predictions that we have made will tend to hold, given that local duplications are highly unlikely.

1. **DIDO:** N; W; S; SE; E; N; N; W; S; E; S; S; NW; N; NW; E; N; S; E; SE; S; W; W; SW; NE; E; E; NW; W.
2. **SISOSS:** N; W; S; SE; SE; N; W; N; N; W; NE; SE; SE; S; SW; N; W; NE; W; N; W; SE; S; SE; N; W; W; NE; N; W; S; S; SE; E; N; W; N.
3. **SISODD:** N; E; S; S; W; NW; E; N; E; E; S; W; S; W; S; NW; NE; N; E; SW; E; S; W; N; N; E; S; S; W; N.
4. **DISOS:** N; W; SE; E; S; W; NW; N; E; W; SW; SE; E; E; W; NE; N; W; W; S; SE; S; E; N; W; N; NW; NE; S; W; SE; E; SE; W; W; N.
5. **DISOD:** N; E; S; S; N; N; W; E; SW; S; E; N; N; NW; SW; E; E; SE; S; W; N; W; NW; NE; S; E; S; SW; SE; N; N; W.
6. **SIDOS:** S; E; N; W; NW; E; N; SE; SW; S; E; E; NW; W; NW; E; E; SW; S; E; S; NW; N; NW; E; S; S; E; W; N; NW; E; W; SE.
7. **SIDOD:** S; E; E; NW; W; N; W; SE; S; E; S; NW; N; N; E; S; SW; E; W; N; N; S; S; E; N; W; N; N; SW; SE.

Figure 3: The route types used in the experiments, and the particular move sequences that realized them

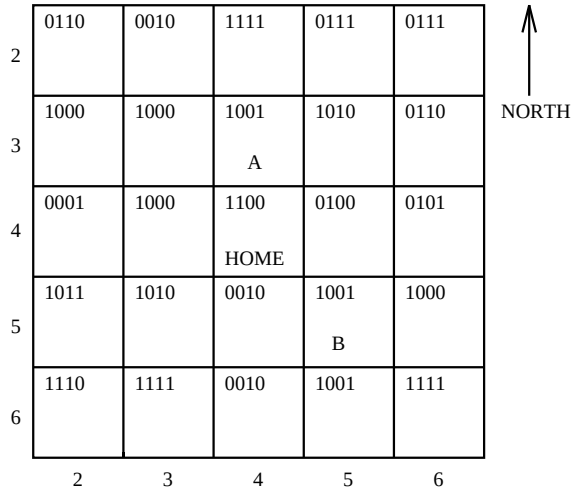

Figure 4: The region of the grid world used in the experiments. The four-bit binary vector at each location indicates the description or sensation vector associated with that location.

7 Experiments & results

To test these predictions, we had the CNM learn 7 routes, each route realizing a different route type (see figure 3). The particular environment that was used is shown in figure 4. The CNM converged on a solution with no errors within, on average, 16330 epochs of training.[6] The learning rate was 0.01, and the momentum was 0.5.

As we surmised (cf section 4.2), the standard Euclidean measure of distance (and attempts at functional analysis based on it, such as cluster analysis) is an unreliable measure of functional equivalence. The non-linear nature of networks means that sometimes codes that are geometrically close will have different functional properties, and sometimes codes that are relatively geometrically distant will be functionally equivalent. An example of this was found in the codes (C_{29}, C_{24} and C_{33}) the CNM learned for the DIDO route (for moves 29, 24, and 33; see figure

[6] In a few of the simulations, there were a few prediction errors (at most 2 on any route) with respect to the learned route, but none of the errors involved the two locations under scrutiny nor their immediate neighbours.

Code 1	Code 2	Distance	F_b	F_p	F_d
C_{29}	C_{24}	0.87	87.50%	62.50%	-1.056
C_{29}	C_{33}	0.74	84.38%	50.00%	-1.517

Figure 5: An example of Euclidean similarity and functional equivalence coming apart.

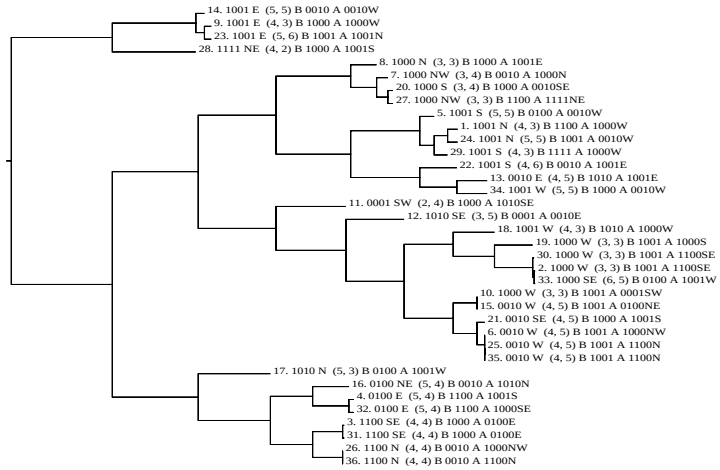

Figure 6: Cluster analysis of all location codes used in the DISOS route. Labels indicate the move number that produced the code, the description vector for the location, the move made, the coordinates of the location, the description vector of the previous place, the description vector of the following place, and the move taken to get there.

5). Although the distance between C_{29} and C_{33} was less than than the distance between C_{29} and C_{24}, the functional equivalence of the former pair was less than that of the latter pair, on all three of our measures of functional equivalence.

7.1 Qualitative analysis

One can use cluster analysis to get a rough idea of the different degrees of systematicity developed in learning the different types of routes. Figure 6 shows the cluster analysis of the location codes developed in learning the DISOS route. Note how the codes corresponding to (5, 5) are found in several parts of the tree, suggesting low functional equivalence between them. The same applies to the codes for (4,3). Note also that codes for (5,5) and (4,3) are often clustered together, suggesting a high functional equivalence between them. Both of these factors indicate a very low degree of systematicity.

In contrast, the cluster analysis of the codes developed for the SIDOD route (figure 7) suggests a high degree of systematicity. The codes for (5,5) are all clustered together, as are the codes for (4,3), and the (4,3) and (5,5) codes are in different (albeit neighbouring) sub-clusters, suggesting that they might be functionally divergent, despite the sensory equivalence of the two locations.

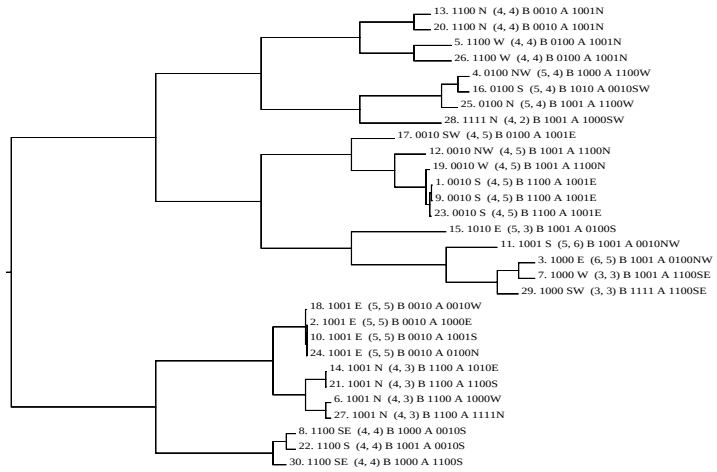

Figure 7: Cluster analysis of all location codes used in the SIDOD route. See figure 6 for an explanation of the labels.

7.2 Quantitative analysis

However, in order to provide a more detailed comparison of the systematicity of the location codes developed for each route type, we need a quantifiable measure of systematicity. In keeping with the two components of the definition of systematicity, systematicity should be maximized when the functional equivalence between codes for the same location is maximized, and when the functional equivalence between codes that correspond to different locations is minimized. Thus, systematicity can be seen as the average functional equivalence of codes for the same location minus the average functional equivalence of codes that represent different locations. For the particular cases considered in the experiments, this can be formalized as:

$$S(A,B) = 2\frac{\sum_{i=1}^{N}\sum_{j=i}^{N} F(a_i,a_j)+F(b_i,b_j)}{N(N-1)} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} F(a_i,b_j)}{N^2}$$

where:

- A and B are the distinct yet sensorily equivalent locations the codes for which are under consideration (in our case, the A and B were the locations (5,5) and (4,3) in all routes);
- N is the number of times that the route enters the places A and B (in our case 4);
- F is the functional equivalence measure being used, be it F_p, F_b, F_d (see section 4.2); and
- a_i and b_i are the location codes that are active on the ith visits to A and B, respectively.

The first term sums up the functional equivalences of the four codes that represent A and the functional equivalences of the four codes that represent B, and then divides by the number of such comparisons (in our case 6) to get an average; the second term sums up the functional equivalences of the codes that represent different places, and then divides this sum by the number of such comparisons (in our case 16) to yield another average. The difference then expresses the degree of systematicity.

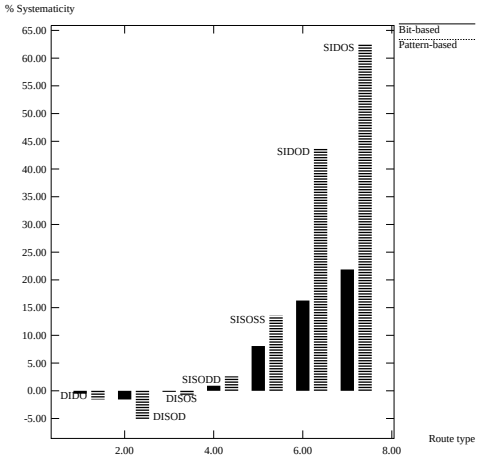

Figure 8: The observed bit- and pattern-based systematicity of the 7 tested route types.

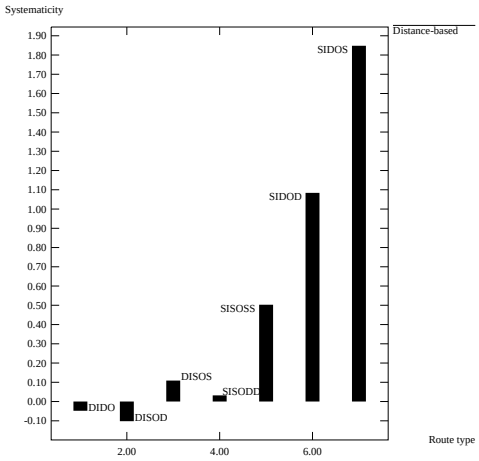

Figure 9: The observed distance-based systematicity of the 7 tested route types.

The systematicity results using pattern- and bit-based functional equivalence measures are shown in figure 8. We also calculated the systematicity of the 7 scenarios using the distance-based measure, shown in figure 9 (note that this is not a measure of the Euclidean distance between the codes, but a measure of the distance between the sensations that two codes predict).

8 Discussion

The data are fairly univocal. The SIDO scenarios produce the most systematic codes, the SISO scenarios to a lesser extent, and the DISO scenarios even less. This agrees with our predictions, and thus supports the postulates, principles and hypothesis of sections 5 and 6.

However, two aspects of our predictions were not borne out. First, although our predicted general ranking of the SIDO, SISO, and DISO scenarios was correct, our predictions of the rankings of the sub-types within those groups were not. In particular, we expected SIDOD to be more systematic than SIDOS, and SISODD to be more systematic than SISOSS, yet the converse was found in both cases. It seems that having a common move in helps the codes for the two locations to diverge from each other, which suggests that in some cases principle 1 and postulate 2 (cf section 6) do not hold. Further work, then, should include an investigation into that principle and postulate.

Second, we expected the systematicity of the DIDO route to be between that of the SIDO and SISO routes, but it is in fact near the bottom of the scale, better only than the DISO routes. This suggests that the commonality of the codes for the same location has a greater weight in the determination of systematicity than divergence between codes of different locations. This would also explain the negative systematicity results for the DISO routes.

Note that, predictably, the high systematicity of the SIDO routes yielded perfect generalization. That is, since the CNM was trained until it got all the predictions on its routes correct, the four codes that it used for each of the four times in (5,5) each make a correct prediction for what would result from moving north, east, south and west, respectively. But since the CNM in this case developed a systematic representation of (5,5), these four codes are highly functionally equivalent, and thus a correct prediction is made if the CNM considers moving south from (5,5) at a point in the route where it normally would have gone east. This shows that the CNM is doing more than just memorizing a route (cf section 4).

Another look at the cluster analysis of the SIDOD route (figure 7) suggests that another kind of generalization might be at work. Even though the SIDOD structure of the route was only expected to foster a systematic representation of (4,3) and (5,5), it appears that *every* developed location code meets the systematicity requirements (contrast the ordered grouping of co-referential location codes in figure 7 with the relatively jumbled groupings in figure 6). Perhaps developing systematic representations for a few locations can serve as a catalyst that bootstraps systematic representation in general, for locations that have not been the focus of a SIDO strategy. This will be the subject of future work.

9 Comparisons with other work

Independently of the work done in (Chrisley, 1993), there has been work on applying simple recurrent nets (SRN's) to the task of learning finite state automata (FSA's) which suggests their capability to develop systematic representations of those automata. In particular, the cluster analysis in (Cleeremans, 1992, chapter 2) of an SRN trained to predict the next letter in a sequence constructed from a simple grammar suggested that the net had indeed developed hidden

unit patterns that were active if and only if the portion of the string processed so far corresponded to a particular node in the FSA representation of that grammar. Furthermore, Cleeremans went on to examine one parameter that seems crucial in determining whether or not such systematic representations will develop: the number of hidden units. Given too many hidden units, the network will use different regions of the hidden unit space to represent the same state, thus preventing generalization.

The work here can be thought of as complementing Cleeremans', in that it highlights external, rather than internal, constraints on the development of systematic representation. However, there are several other differences between that work and this,

Cleeremans used a very different training strategy. In order to train his network, he used 60,000 sequences with an average of 7 patterns per sequence, yielding 420,000 training patterns (as opposed to our use of 30 patterns or so). He thus had little idea of which of those patterns were crucial for systematic representation. Our lighter approach allowed us to investigate the relatively minimal requirements for the development of systematicity.

Also, the nature of the task is different: the CNM learns by predicting the sensory properties of states, while Cleeremans's model predicts possible sequences of letters, which are analogous to the *actions* in the CNM (i.e., they are what effect state transitions). Perhaps the CNM could be modified to use both kinds of learning to further constrain the development of systematicity. This would also improve the CNM's ability to aid in navigation, since it would be able to rule out some possible action sequences as being "ungrammatical". Correlations between these restrictions on movement and sensory properties could then be learned, yielding a deeper understanding of the causal properties of its environment.

In his analysis of (what we would call) the systematicity of his network's representations, Cleeremans relied only on Euclidean distance and cluster diagrams, which as we have shown do not always indicate the true functional equivalence of representations. However, since his behavioral tests were so exhaustive and high-scoring (e.g., his network correctly categorized 130,000 randomly generated strings as grammatical or ungrammatical), perhaps the high number of training patterns ensured a tight connection between distance and functionality.

Given that the development of spatial representations in the CNM can be thought of as the development of "place permanence" (Chrisley, 1993, p 342), recent research into connectionist models of the development of object permanence (Mareschal & Plunkett, 1994) are highly relevant to the work done here, especially since the task in that work is (visual) prediction using an SRN. However, despite these similarities, there are some serious differences. For one, Mareschal & Plunkett's work has the advantage of successfully addressing and explaining actual human developmental data. Also, their evaluation of their network is entirely behavioural; i.e., they do not analyse the representations their network develops via cluster analysis or calculate systematicity measures.

Finally, there has been some investigation into the conditions under which SRN's learning FSA's can transfer what they have learned to other domains (Dienes, 1994), which can be seen as another kind of generalization. For example, it was found that in order to achieve transfer, a network should be presented with sequences that have repeated elements. Perhaps the principles that were useful here in predicting and explaining the conditions for the development of one form of generalization could be of use in explaining why repeated elements are so crucial to developing this other kind of generalization.

10 Future work

In addition to the future work already mentioned (cf sections 3, 8 and 9), some other possibilities should be mentioned.

The generalization exhibited by the CNM so far only involves different combinations of transitions that it has made before. Another important kind of generalization is to transitions not made before, but for which one has been given enough information to make a successful prediction. For example, suppose the an agent using the CNM has never moved south from (4,4) to (4,5), but it has been to (4,5) via moving east from (4,4) to (5,4), then south to (5,5), and then west. It would be very significant if the CNM could develop representations so systematic that the code for "south from (4,4)" was functionally equivalent to the code for "east, south and west from (4,4)", even though it had never moved south from (4,4) before. Specifically, it might be useful to consider under what conditions does the CNM develop codes such that the action vectors cause systematic movements in the principal component space of the location codes.

Perhaps the CNM is on its way toward this, as evidenced in the unexpected systematicity in figure 7. But another idea for how the CNM might achieve this is for it to have a small core of systematic location codes which it repeatedly redeploys in order to represent different areas. This might also address the scaling-up problems of SRN's that have been observed (Cleeremans, 1992, p 66). It is unclear, however, how such a structure might be learned, so it is unclear to what extent such an architecture would be furthering a connectionist naturalization of epistemology, as opposed to assuming a complex innate symbolic mechanism.

Finally, if the CNM were to be refined so that it might be used to model and explain actual data of some sort, a natural area of application is the spatial learning of rats. The "place cells" (O'Keefe & Nadel, 1978) observed to be in the rat hippocampus, cells which are maximally active if and only if the rat is at a particular location, sound very much like the systematic location codes developed in the CNM.

Acknowledgements:

The authors wish to thank Terence Cain, Dave Cliff, Zoltán Dienes, Bob Ives, Philip Jones, Derek Parkinson and Matthew Taylor for their assistance.

References

Chrisley, R. (1990). Cognitive map construction and use: A parallel distributed processing approach. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *The Proceedings of the 1990 Connectionist Models Summer School*. San Mateo, CA: Morgan Kaufmann.

Chrisley, R. (1991). A hybrid architecture for cognitive map construction & use. *Artificial Intelligence & the Simulation of Behaviour: Special Issue on Hybrid Models of Cognition*, (78), Autumn/Winter 1991.

Chrisley, R. (1993). Connectionism, cognitive maps & the development of objectivity. *Artificial Intelligence Review*, (7):329–354.

Cleeremans, A. (1992). *Mechanisms of Implicit Learning*. Cambridge, MA: MIT Press.

Cussins, A. (1990). The connectionist construction of concepts. In Boden, M. (editor) *The Philosophy of Artificial Intelligence*, pages 368–440. Oxford: Oxford University Press.

Cussins, A. (1992). The limitations of pluralism. In Lennon, K. and Charles, D. (editors) *Reduction, Explanation and Realism*, pages 179–224. Oxford: Oxford University Press.

Dienes, Z. (1994). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. Talk given to PDP Discussion Group, School of Cognitive & Computing Sciences, University of Sussex, 18 November 1994.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, (14):179–212.

Glass, A. and Holyoak, K. (1986). *Cognition* (2nd edition). New York: Random House.

Holland, A. (1994). *Simple Recurrent Networks, Non-conceptual Content & the Development of Objectivity.* Unpublished MSc thesis. University of Sussex, Brighton.

Langton, C. (1989). Preface. In Langton, C. (editor) *Artificial Life: Proceedings of the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems, Los Alamos, NM, Sept, 1987.* Addison-Wesley: Volume VI in the series of the Santa Fe Institute Studies in the Sciences of Complexity.

Luria, A. (1968). *The Mind of a Mnemonist.* New York: Basic Books.

Mareschal, D. and Plunkett, K. (1994) Object permanence and visual tracking: A connectionist perspective. *The Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.*

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map.* Oxford: Clarendon Press.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by back-propagating errors. *Nature*, (323):533–536.