Connectionist Semantics and the Collateral Information Challenge

FRANCISCO CALVO GARZÓN

Abstract Jerry Fodor and Ernest Lepore have launched a powerful attack against Paul Churchland’s connectionist theory of semantics—aka State Space Semantics. In one part of their overall attack, they exploit the potentially orthogonal histories of different individuals to introduce what they labeled ‘the collateral information problem’. Aarre Laakso and Gary Cottrell have recently put forward a mathematical technique for measuring conceptual similarity across neural networks. Churchland uses Laakso and Cottrell’s technique to defend State Space Semantics. In this paper I shall highlight a potential problem for Laakso and Cottrell’s technique, and for Churchland’s subsequent defence of connectionist semantics that has been ignored in the connectionist literature. I shall argue that a connectionist sympathiser of Churchland could not make use of Laakso and Cottrell’s neurosimulations to address Fodor and Lepore’s collateral information challenge.

I. INTRODUCTION: CATEGORIZATION AND CONCEPT ORGANIZATION IN CONNECTIONIST MODELS OF COGNITION

Although sensory experiences come in isolated epidodes, humans appear to be very well equipped to form abstract categories out of their similarities. Most of us will probably be able to recognize a Caucasian ovcharka as a dog even if we’ve never seen one before. We can furthermore form abstract prototypes from individual experiences, and agree that German sheppards are typical dogs. Categorization and prototype extraction are well-known examples of cognitive tasks that can be neuromodeled in connectionist theory.

In blunt terms, we may define a connectionist neural network as a device for approximating functions. The basic components are simple processing units and connections between them. Computations are based primarily on the interconnection of the units. By exploiting statistical regularities in the weight space of connections, neural networks are able to pair patterns encoded in an input domain with patterns that belong to an output one. Learning, in the connectionist guise, consists in minimizing the network’s error in the space of potential connection weight assignments. A simple

---

1 The reader unfamiliar with the connectionist literature may care to consult Bechtel and Abrahamsen (1991) for a philosophy-oriented introduction, or McLeod et al. (1998) for a psychology-oriented one.
feedforward network can behave as a universal approximator. That is, for any given function with a finite range and a finite domain, the function can be computed to an arbitrary level of accuracy (Hornick et al., 1989). Any arbitrary input-output mapping can (in principle) be performed—a different issue is whether any particular network hits on the right architecture, training conditions, etc., that allows the network to deliver the goods (see section II, below). Ultimately, the working hypothesis is that connectionist networks will be able to achieve results similar to those achieved by human subjects, while mimicking the coarse-grained structure of cognition.2

Consider a simulation task in which the output layer of a simple feedforward network represents various more or less intuitive categories, say, A, B, and C. To simplify matters, we may think of each unit at the input layer as specifying a particular feature that the exemplars that belong to the different categories possess. On the other hand, each output unit encodes information about a particular category. The idea would be that the network learns to assign various combinations of "features" to categories, by having these features represented (somehow) by patterns of activation at the input layer. A network trained by backpropagation can learn to classify stimuli into an established set of categories. In the training phase, after each input has been presented, activations are calculated upward based on existing connection weights. Activations feed forward through the hidden layer/s to the output one, and as a result one of the three output nodes (for the answers ‘A’, ‘B’, and ‘C,’ respectively) acquires a higher value than the others, eliciting thus an output response. If the network’s response is correct, then the patterns of activation are left intact. Otherwise, the weights on the connections are

---

2 Neural networks bring a great deal of simplification in contrast with ‘biological networks,’ and should be interpreted as abstract models of real nervous systems. However, it must be pointed out that modelers are progressively making use of more and more constraints under the light of neurobiological research, refining thus their models so as to push the neural metaphor of connectionism as far as possible—see Rolls and Treves (1998).
recursively adjusted downward according to the measure of error calculated as a result of the difference between the actual response the network has given and what the correct answer should have been. Once the training phase is completed, the network is tested by presenting it with several previously unencountered exemplars belonging to all three categories. Under optimal training conditions, a network can respond correctly to the novel inputs.

Having three categories, we may for example encode output patterns over three output units by using binary suffixes as shown in Table 1:

<table>
<thead>
<tr>
<th>Category</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1 0 0</td>
</tr>
<tr>
<td>B</td>
<td>0 1 0</td>
</tr>
<tr>
<td>C</td>
<td>0 0 1</td>
</tr>
</tbody>
</table>

Table 1: Target output patterns for the A, B, and C categories

Since these are orthogonal unit vectors, the distance between any two is just the same. So, were we to focus exclusively on the output level there would be no sense in which the network would regard a bulldog as more similar to a labrador than to a corgi, for example. But things are different if we measure the distance between the hidden unit vectors corresponding to the various categories—what we may call the connectionist "concepts" (see below). Once the network has learned the task, the connectionist rendition of the "concept" corresponding to each category would be the average of the hidden unit activation vectors that are produced by input patterns that produce activation at the output unit corresponding to the category. By observing the behaviour of the hidden units during the training phase and how the network performs subsequently

---

3 Someone might want to argue that the (proper) connectionist concept of a “concept” is the shape of the hidden unit activation space that corresponds to a particular categorical output (many thanks to Aarre Laakso for bringing this point to my attention). However, for present purposes, we need not dwell on this issue.
under the presence of new exemplars, we find that the training process has *partitioned* the state space of possible patterns of activation across the hidden units. In particular, the internal space has been split into three sub-spaces that correspond to A-, B-, and C-representations, respectively.

One way to grasp this partitioning is via statistical techniques such as *cluster analysis*. A cluster analysis is a systematic way to measure and display distances between representations. It consists in pairing each pattern of activation with its closest neighbour. An average activation value is then calculated, and the process of pairing neighbours is repeated for the new patterns of activation. This technique is hierarchically applied, arriving in the end at a final clustering in space where points are located in several specific regions as a function of the similitudes shared with other points, culminating each sub-space in a *hot spot* (see fig. 1). The hot spot in the A-like space is taken to represent the *prototypical* A exemplar; and the same goes for the other sub-spaces. All the vectors whose activation values correspond to A/B/C-inputs are seen as points in a A/B/C-region. We may then judge how representative of a category an exemplar is as a function of the geometrical distance, in a pre-specified Euclidean space, that separates the vector in question from one particular hot spot. In this way, the task of the network can be better understood as a process of discrimination, not only between A-like, B-like, and C-like representations, but also between more or less prototypical A, B, and C exemplars. Hence, under a novel input pattern, the network is discriminating the proximity of the new pattern of activation produced—represented as a point—to one or other of the hot spots. In short, the key point to bear in mind is that when a hot spot is activated, it represents the network’s concurrent understanding of a given environmental feature.
Paul Churchland makes use of this framework in order to articulate a connectionist-inspired theory of mental representation—aka State Space Semantics. Churchland (1986) proposes that we understand concepts as points in a partial state space of a connectionist network. These points correspond to the tips of the vectors determined by the levels of activation of the different units in hidden layers. The semantic characteristics of a concept can then be seen as a function of the place that that concept—i.e., point—occupies in a geometrically characterized hyperspace. In this way, Churchland proposes, we may talk of semantic similarity between concepts in terms of the proximity of their respective absolute positions in state space, as identified in relation to a number of semantically relevant dimensions. In short, State Space Semantics tells us that the semantic connection of a concept A with properties x and y can be analyzed in terms of the position of concept A in a semantic space which is delimited in part by the x- and y-dimensions.

We now have an elementary picture of the basic components and dynamics of connectionist networks. The bulk of the paper will be devoted to analyzing a powerful attack against connectionist models of cognition put forward by Jerry Fodor and Ernest Lepore, and its implications for the theory of mental representation.

II. STATE SPACE SEMANTICS AND THE COLLATERAL INFORMATION CHALLENGE

As we saw in section I, the interconnectivity among the units in a network, and the flow of activation from one layer to another are sufficient to master complex tasks such as categorization. Complexity, through the connectionist lens, emerges from the non-linear

---

4 This is a simplification of Churchland’s (1986) original proposal. Since then Churchland has developed State Space Semantics further, in order to meet a number of challenges (see Tiffany, 1999; and section II, below).
properties of simple dynamical systems. Further qualification, however, is required to pave the way for an appraisal of Fodor and Lepore’s attack on connectionist semantics.

The representation of concepts that a connectionist model of categorization generates cannot be dissociated from the strategies the network makes use of in order to put exemplars together into categories. Connectionist activity remains rooted in representations that are context-dependent, driven by the inherently dynamical character of the interconnection of many simple units that operate in parallel. The way in which exemplars get represented in a connectionist network has an important role to play in the mapping to be performed. Exemplars are encoded as patterns of activation across the input space. At the other end, output units receive a weighted sum of activation from lower level units, in such a way that the exemplars are assigned to specific categories as a function of the weight matrix obtained. The network thus works exclusively at the level of numerous context-dependent activation patterns. In this way, different patterns of connectivity, different numbers of hidden units (which result in hidden unit vectors of different lengths in different networks), and different ways of encoding the same set of input features will have an effect on the specification of the network’s representational space, and therefore, on the performance of the network. Ultimately, in the connectionist dynamical approach, we may say that the knowledge the network has of a target domain resides in the connection weights that have been generated during learning in accordance to a learning algorithm. The learning rule operates locally and concerns exclusively how the weights will change as a result of an incoming flow of activation. Performance, thus, is highly sensitive to the idiosyncracies of processing, and variations in hidden representational space will have causal consequences for the network’s behaviour.
It seems, under the light of these considerations, that some of the vexing philosophical problems for the theory of mental representation remain. Fodor and Lepore (1999) have launched a powerful attack against Churchland’s connectionist theory of semantics. In one part of their attack, Fodor and Lepore argue that the aforementioned architectural and functional idiosyncrasies of connectionist networks preclude us from articulating a notion of conceptual similarity applicable to State Space Semantics. Their objection can be summarized as follows: State Space Semantics understands conceptual similarity across networks as similarity in the activation patterns across those dimensions that specify the networks’ representational spaces. However, under this connectionist framework, it seems that two individuals—i.e., networks—cannot possibly entertain the same concept. Processing in connectionist networks is highly idiosyncratic. Differences in the encoding of the input data, in the architecture of the model, and in the dimensionality in hidden space, strongly constrain how a network proceeds in order to achieve successful performance. Idiosyncrasies in encoding, architecture, or hidden dimensionality, Fodor and Lepore argue, make it impossible to talk of similarity of patterns of activation across networks. It then seems to follow straightforwardly that we cannot talk either of similarity of positions in state space across networks.

Churchland (1998) hopes to bypass Fodor and Lepore’s attack by equipping State Space Semantics (see section I, above) with a non-absolute measure of conceptual similarity.
similarity. By putting the emphasis on the similarity of the relative positions of different activation patterns, Churchland defines conceptual similarity across networks in terms of the position of a given pattern of activation in relation to other patterns in the same representational space. In this way, we may say that two networks share the same conceptual repertoire if the set of relations among the activation patterns in the first network is isomorphic to the set of relations obtained in the second network.

Churchland’s new account shows some promise in the fact that a non-absolute definition of similarity relaxes the demands on State Space Semantics. Note that now we can ignore the different dimensionality, as well as the particular microcontent of each dimension of each state space. All we need then—or so it appears to Churchland—is to establish a set of necessary and sufficient conditions for a relative definition of conceptual similarity, and an explanation of how to implement it in computational models of cognition. To achieve these goals, Churchland turns to some empirical research carried out by Aarre Laakso and Gary Cottrell (2000).

According to Laakso and Cottrell, we do have a criterion for judging conceptual similarities across different connectionist networks. Namely, by measuring distances among points within the hidden space of a given network, and correlating those measures with the measures obtained within the hidden space of a distinct network. Laakso and Cotrell considered the following simple case: Suppose we take two simple feedforward networks—call them NET1 and NET2—with one and two hidden units, respectively. NET1 and NET2 are trained on the representation of three unspecified things, say, A, B, and C. Whereas NET1 comes to represent A, B, and C with the coding vectors:

---

concept (given that human brains have different numbers of neurons, which are differently connected to each other, and which exhibit different patterns of causal connectivity)—see Calvo Garzón (2000).
A = <0>, B = <50>, and C = <100>,

NET2 learns to represent the same three things with the coding vectors:

A = <0, 0>, B = <30, 30>, and C = <80, 0>.

We can then form the following matrices (see Table 2) by considering the distances between all the representations within NET1, and also comparing the distances between all the representations in NET2.

<table>
<thead>
<tr>
<th></th>
<th>1. Unit network</th>
<th>2. Unit network</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>100</td>
<td>80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>50</td>
<td>42</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>80</td>
<td>58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>42</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

TABLE 2. Comparison of distances between points in two different vector encodings (Laakso and Cottrell, 2000, p. 57)

Since both matrices are symmetric we can extract the respective vectors. In our toy example, the two vectors are:

<50, 100, 50>, and <42, 80, 58>.

Given the fact that they have the same dimensions, they can be easily compared. In this way, by computing these distances, we can employ a mathematical measure of similarity—e.g., Pearson’s $r$ correlation—with which to compare the representations of NET1 and NET2. In Laakso and Cottrell’s toy example, Pearson’s $r$ correlation between <50, 100, 50> and <42, 80, 58> was .91. Values close to 1 would suggest that the hidden representational spaces of NET1 and NET2 have been structured in similar ways. A perfect match (correlation = 1) would indicate that representations across networks...
are identical. The idea, in short, is that points in different hidden spaces stand for the same, or similar, things in case there is a high correlation between the distances among the sets of points—i.e., concepts—in the respective networks. Different dimensionality, architecture or encoding, Laakso and Cottrell argue, bring no trouble, insofar as correlated distances between points in the respective spaces are preserved.

Laakso and Cottrell tested this strategy on a colour-categorization task. They employed networks with different internal dimensionality (between 1 and 10 hidden units), as well as different input codings. Once the networks mastered the categorization task, Laakso and Cottrell compared relative positions across state spaces. The mathematical measurements were computed, and the correlations obtained were very high, independent of the number of hidden units employed by the networks. Laakso and Cottrell’s mathematical measure of similarity may well constitute a robust criterion of content similarity—robustness achieved by the hand of correlations where different input encoding, different architectural connectivity, and specificity in the weight assignments is involved (although see below). With Laakso and Cottrell’s neurocomputational results at hand, Churchland reaffirms the credentials of State Space Semantics:

The account we are currently piecing together ... is not just a syntactic account; for it promises to do what we have always expected a semantic theory to do. It ... provides a criterion for assigning the same contents to the representational vehicles of distinct individuals. It gives us, that is, a criterion for accurate translation across the representational/cognitive systems of distinct individuals. (Churchland, 1998, p. 31)

Laakso and Cottrell’s results get connectionist semantics off the ground, and seem to shed new light on the Fodor-Lepore/Churchland debate over the fate of State Space

---

8For the details of the experiment, see Laakso and Cottrell (1998).
Semantics.9 Bearing in mind this setting, in the remainder of this paper I’d like to draw the reader’s attention to a problem for Churchland’s defence of State Space Semantics—‘the collateral information problem’; a problem that to the best of my knowledge has been completely ignored in the connectionist literature (although see below)10

A typical categorization task involves the production of one and the same output under the presence of inputs that are not identical. Traditionally, a great deal of attention has been given to the mastery of semantically interpretable—cognitive—categories. However, we may distinguish between perceptual and cognitive tasks in categorization. The categories obtained in Laakso and Cottrell’s colour-categorization task involve sensation and perception. Nevertheless, it is worth pointing out that both Laakso and Cottrell’s colour-categorization task and cognitive tasks in general involve the same statistical process of input/output approximation that neural networks must achieve in order to subsume a specific input pattern of activation under a general output one. In this way, having demonstrated by the hand of Laakso and Cottrell neurocomputational results how to judge similarities across different connectionist networks with an emphasis on perceptual tasks, we may now zoom back to enlarge the picture, and turn to questions concerning the broader role of categorization in human cognition as it relates to Fodor and Lepore’s critique of connectionist semantics.

9 The reader familiar with the literature on connectionism will have adverted that by the time Laakso and Cottrell’s results came out, a number of (post hoc) statistical techniques, such as cluster analysis (section I, above) had already been successfully deployed in order to compare representational spaces across networks. However, the task became increasingly difficult, the more the clustering profiles tended to diverge. Lack of measurement techniques wasn’t the problem. The problem was rather lack of objective measurement techniques. As Cottrell pointed out when presenting Laakso and Cottrell’s neurocomputational results: “When I eyeball the dendograms [i.e., the tree-structured hierarchical clusterings] for two distinct networks, I may say, ‘yes, they are fairly close’, but that’s just my reaction. We need an objective measure of such things” (quoted from Churchland, 1998, p. 18).

10 Elsewhere (Calvo Garzón, 2000), I’ve argued against Churchland’s general approach to semantics. I considered that it would be of interest to the general reader to discuss this particular problem because of its absence in the literature, and the fact that Churchland (personal communication) has acknowledged that Fodor and Lepore’s collateral information problem has not been addressed in his reply.
In one part of their overall attack, Fodor and Lepore exploit the potentially orthogonal histories of different individuals to introduce what they labeled ‘the collateral information problem’:

The point is that if a semantics recognizes dimensions of state space corresponding to all the properties of dogs about which our beliefs differ, then even assuming that your state space has exactly the same dimensions as mine, the location of the dog concepts in our respective spaces is likely to turn out to be quite significantly different. This should be all sounding like old news; it’s just the worry, familiar from attempts to construct a notion of content identity, that a lot of what anybody knows about dogs counts as idiosyncratic; it’s “collateral information”, the sort of thing that Frege says belongs to psychology rather than semantics. If we are to have a notion of meanings as shared and public property, a robust notion of meaning, we must somehow abstract from this idiosyncratic variation. (Reprinted in McCauley, 1996, pp. 156-7)

In this passage, Fodor and Lepore target Churchland’s earlier reading of State Space Semantics (that is, Churchland’s definition of semantic similarity in terms of absolute positions in state space with respect to a given set of dimensions—see section I, above). We may update their charge to address Churchland’s latest approach to connectionist semantics (that is, Churchland’s definition of semantic similarity in terms of relative positions in state space, such that different dimensionality across networks becomes harmless). Their charge becomes: Even assuming that the set of relations among the patterns of activation in your state space is similar to the set of relations obtained in mine, and even assuming that the distances among the sets of vectorial representations in our respective state spaces are highly correlated, the location of the dog concepts in our respective spaces may still differ significantly.

Fodor and Lepore exploit the fact that different individuals are likely to have had very different encounters with diverse environmental features. The reader, nevertheless, may wonder why this should be a problem. As Fodor and Lepore point out (ibid., pp.
157-ff.), it is at first sight difficult to appraise how we can entertain the same concepts if, according to State Space Semantics, all dimensions in hidden space determine the semantic content of our conceptual repertoires. Fodor and Lepore consider two ways out for Churchland: On the one hand, the defender of State Space Semantics may help herself to an analytic/synthetic distinction, in order to discriminate between those hidden dimensions that are highly relevant in determining content—and which hopefully we all share!—and those dimensions which are less relevant, or not relevant at all—and which hopefully correspond to those axes in state space that reflect historical idiosyncracies. On the other hand, Fodor and Lepore argue, we may appeal to the empiricist assumption that all concepts are (statistical) functions of ‘sensory’ concepts. This would also furnish us with a robust account of conceptual similarity since all dimensions would then correspond to sensory properties. Hence, we may say that two individuals share their conceptual repertoires if they have relevantly similar sensory connections. Unfortunately, Fodor and Lepore would recommend neither of these two options to their connectionist enemy. Regarding the first option, honouring an analytic/synthetic distinction may bring well-known problems that the reader familiar with the philosophy of language literature will be aware of. On the other hand, Fodor and Lepore wouldn’t recommend the second option either. Although dressed in connectionist clothing—viz., statistical, rather than boolean functions—the assumption is an embarrassing one that would inherit all the problems that afflicted classical empiricism. In short, State Space Semantics is caught on either of two horns—honouring an analytic/synthetic distinction,

11Note that in the passage just quoted Fodor and Lepore assume, as a best-case scenario for Churchland, that different state spaces enjoy the same dimensionality.

12Treating these problems would take us far afield. However, since Churchland himself rejects the analytic/synthetic distinction, we may for present purposes agree with Fodor and Lepore, and Churchland, and ignore that option.
or resurrecting empiricism; and neither of these alternatives is very attractive, both in Fodor and Lepore’s, and in Churchland’s view.

In his latest defence of State Space Semantics, Churchland does not address the collateral information problem. Laakso and Cottrell’s neurosimulations showed us that conceptual similarity could be objectively measured “regardless of how inputs are encoded, and regardless of number of hidden units.” (1998, pp. 595-6) This however, I claim, does not address the collateral information problem since in Laakso and Cottrell’s simulations information across networks was never compared regardless of the training histories of the networks. In the experiment reported earlier, several networks were trained on a colour-categorization task. The networks were trained on four different encodings of the input data: a real encoding, a binary encoding, a gaussian encoding, and a sequential encoding (ibid.) All encodings, however, were variations of the same set of data. The idea was to illustrate that individuals with contrasting sensory modalities may categorize the world in similar ways. Commenting on their methodology, Laakso and Cottrell claim:

Formally speaking, our method can be used to compare measurements from any two state spaces. In fact, however, [...] we imposed additional constraints on the state spaces we compared. The spaces were generated by presenting identical stimuli to subjects who “spoke similar languages” (all of the network “subjects” were trained with the same labels on input stimuli). Using feedforward connectionist networks, it was both possible to conduct such experiments and reasonable to assume that activations caused by identical stimuli were comparable. Nevertheless, the fact that we imposed those constraints might give rise to a number of objections [...]” (ibid., p. 71).

13It may be the case that Churchland drops the issue after his exchanges with Fodor and Lepore, confident that the battle has been won. I doubt that this is the case, but I won’t press on this issue here (see Fodor and Lepore, 1999, for a very lucid exposition of some of the vexing problems that remain for any theory of semantics—classical or connectionist). The purpose of this paper is more modest. I shall simply argue that Churchland wouldn’t be able to appeal to Laakso and Cottrell’s results to address the “collateral information problem.”
Although they discuss a number of objections to their strategy (ibid., pp. 71-5), Laakso and Cottrell do not compare at any point networks trained under different histories. It must be stressed, however, that some of the potential problems that they envisage are intimately related to our current concerns, and do bring some light to the issue. Thus, they acknowledge for example that their technique cannot be applied to recurrent neural networks of the sort required to represent time implicitly in connectionist processing. The employment of feedforward networks forces Laakso and Cottrell to consider exclusively “subjects” who process static inputs, in such a way that between time $t$ and time $t+1$ of processing the state the network is in does not get preserved. However, as Laakso and Cottrell humoristically point out, “It is likely that even humans sitting on a couch watching TV are not such perfectly passive receptacles. Rather they bring internal state to the processing of input” (ibid., p. 72). In short, Laakso and Cottrell’s experiments somehow pressupose that the neural networks to be compared do not have a history of their own. On the other hand, they see a further objection with their proposal. Namely, that their experiments don’t take into account the fact that different networks may process different numbers of “concepts.” Given their strategy of measuring correlations between distances among sets of points across networks, it is obvious why this would be a problem: “If you have 40 [concepts] and I have only three, how is it possible to compare our representational states?” (ibid., p. 73).

14 Indeed, Cottrell claims (personal communication) that they never intended to do so. The fact that people may not share the same conceptual repertoires should not be seen as a problem. Obviously, if someone is exposed to a class of colours more often than someone else, the representational structure of the first one will be somehow more elaborated. However, as long as certain fundamental concepts are shared, Cottrell maintains, communication and exchange of information can proceed. Scrutinizing these comments would take us far afield. However, since Cottrell agrees entirely with Laakso’s line of response (see below), we may ignore it for present purposes, and concentrate on Laakso’s reaction.

15 The reader may care to consult Churchland (1998) for a theoretical proposal that attempts to extend Laakso and Cottrell’s results with feedforward networks to recurrent ones of the sort required to process, for example, grammatical structures that unfold in time—see also Calvo Garzón, 2000.
Despite the acknowledgement of these shortcomings, Laakso would not be sympathetic with my contention that their simulations don’t address the collateral information challenge. He argues (personal communication) that in a sense the issue of people with different histories was addressed. As he points out, in the simulations they reported the exemplars were selected randomly, without replacement, within each epoch of training. In this way, while all the networks were exposed to the same set of data, it is extremely unlikely that any two of them, let alone all of them, saw it in the same sequence. In Laakso’s view, taking into account the task in question and the fact that the training set consisted of a finite number of colours (which are even perceptually discriminable by humans), each of the networks should be regarded as having had a different history—each network saw the colours in a different order.

Unfortunately, I don’t think that this “sequential reading” of the collateral information challenge can really address the worries prompted by Fodor and Lepore. The idiosyncratic variations in the training histories of different individuals that Fodor and Lepore’s argument highlights have nothing to do with the order in which the (same) exemplars are presented to different individuals. Rather, it has to do with the exposition of the subjects to different training histories altogether. Laakso (personal communication) anticipates this point, and accepts the fact that the number of possible experiences that two individuals may be confronted with may be indefinitely large, and the fact that plausibly no two subjects share very many of these experiences. Since the debate as framed by Laakso and Cottrell (and, by extension, by Churchland) is meant to go beyond simple colour categorization, the burden of proof is on the sympathiser of State Space Semantics to demonstrate that subjects exposed to different sets of

---

16 Although further qualification is required (see below), it must be said in fairness to Laakso and Cottrell that each of their networks does have a “history” in the sense that it has been trained on a particular set of
experiences (and not merely to different sequences of experiences) still represent things similarly.

In short, if we are to make use of Laakso and Cottrell’s neurosimulations to address the collateral information problem, I contend, we need to train different networks on different sets of data (that would amount to equipping individuals with different histories of categorization and concept acquisition). For purposes of illustration, let us consider the following way in which this may be accomplished.

To adopt McClelland and Rumelhart’s classical example, consider two subjects who see different dogs, cats, and bagels in the course of their day-to-day experiences. Even though exemplars differ from each other, the subjects do learn the prototype that belongs to each category. The analogous neural simulation would be a task in which the output layers of two different networks must learn to represent the categories DOG, CAT, and BAGEL. We may feed the networks with input patterns that are distortions of the prototypical dog, cat, and bagel, respectively. Under adequate training conditions, connectionist processing will ensure that the networks learn the prototypes of the categories (even without ever seeing any particular exemplar that matches its prototype directly, but only distorted patterns—see below).

What we need now is a clear account of how the notion of collateral information would be captured in this hypothetical situation. Since representational dimensions of hidden space may not correspond to anything very intuitive, it seems natural to suppose that human subjects who make use of different collateral information will be modelled by networks that differ in some way at the input unit level. The difference would not just

---

17 McClelland and Rumelhart (1986) imagined a situation in which a child had to learn about dogs, cats, and bagels. They run a number of simulations and showed how a network that operates in parallel can form categories from individual experiences, can come to represent the central tendency of a set of
be one of input encoding, or of architecture (e.g., number of input units); rather it should be a difference in what features are represented. In short, differences may be modelled, not as differences in the networks themselves, but as a difference in the training sets.  

Simplifying matters, I propose the following five-fold strategy to create different histories of training. We may make use of different:

(i) numbers of encounters with the whole set of exemplars.
(ii) numbers of encounters with particular exemplars.
(iii) (non-overlapping) data sets.
(iv) distortions of the prototypical input patterns.
(v) ordering of the sequence of patterns to be presented.

On the one hand, (i) we may implement a mechanism for stopping training automatically when the error falls below a pre-established range. Effectively, this would amount to training different networks with different numbers of sweeps—the outcome would be the existence of different number of encounters with the whole set of dogs, cats, and bagels that we feed the networks with. On the other hand, (ii) we may deploy data files of different length. Effectively, the outcome would be that the networks would be exposed to different number of encounters with particular dogs, cats, and bagels. In addition, (iii) we may train different networks on different (non-overlapping) sets of related exemplars without experiencing the prototype itself, can produce information about a prototypical exemplar of a category, and can generalize to new experiences.

18 An anonymous referee of *Mind and Language* wonders whether comparing training regimes exclusively is the right way to model differences in collateral information. Certainly it is not. Brain circuitry will have an effect on processing, for instance, or on the initial state a subject is in. In fact, Laakso and Cottrell make some comments on a number of neuroimaging techniques (e.g., fMRI, or PET) that are employed to average information over subjects, removing thus idiosyncratic data. However, we need not worry about these divergences. Although Laakso and Cottrell’s method can be employed to measure degrees of similarity—i.e., correlation—between any two vector-coded representations, we may for simplicity’s sake obviate architectural differences or differences in the encoding, and concentrate instead, for the purposes of this paper, on differences in the training sets themselves.

19 The list is by no means meant to be exhaustive. My proposal is meant to be orientative and may need some recasting before it can be applied directly to Laakso and Cottrell’s type of simulation (see specially fn. 20, below). On the other hand, it must be noted that Laakso and Cottrell do take into account points (i) and (v) below in their experiments.
data. We may consider this as a strong derivative of (ii). Effectively, the outcome would be that the networks would be exposed to different dogs, cats and bagels (rather than being exposed to different numbers of particular dogs, cats and bagels). Furthermore, (iv) we may represent the (increasing or decreasing) physical similarity of different exemplars by (increasingly or decreasingly) distorting the prototypical patterns by probabilistically flipping, for example, in a random fashion the sign of some of their elements. Effectively, if we present each experience of a dog, a cat, or a bagel as a string of binary digits that represents the signal generated by its physical appearance, the outcome would be the representation of visual similarities among different exemplars by distorting the prototypical patterns of the three categories. The distortions would correspond to the *varying appearance of individual exemplars*. An encounter with a new exemplar would be modelled as a presentation of a novel distortion of its corresponding prototype. And finally, (v) we may train different networks sequentially and/or randomly. Effectively, that would amount to presenting the patterns to the networks in different *orders*.

We now have the basic toolkit to gauge the chances of State Space Semantics to survive Fodor and Lepore’s attack. Networks trained to perform a given function are not required to see all possible input/output mappings in order to master their task. In fact, a network exposed to a very small set of exemplars can learn successfully. The trick is simply to expose the network to a *representative portion* of its overall training domain. On the other hand, the less representative the samples are (i.e., the more restricted they are to specific regions of the training domain), the bigger the amount of data the network will have to see in order to induce the correct generalization. Generally

20 It must be noted that in Laakso and Cottrell’s experiments the were no prototypical patterns. Nevertheless, artificial “noise” may be introduced into the data set in a number of ways. The point,
speaking, given the particular nature of the training regime that we choose in terms of combinations of factors (i)-(v), networks may be trained either under what I shall call an *inductively robust* training regime, or under an *inductively weak* one. An inductively robust training regime comprehends a set of data that allows a network to induce a given regularity with the employment of the (ideally) minimum number of samples (see below). On the other hand, by *inductively weak* training regime I have in mind a set of data such that a network being trained on this set must ‘see’ a large number of samples before being able to induce the same regularity from the environment[^22].

Take two hypothetical networks trained on different inductively robust regimes—call them training regimes $a$, and $b$, respectively. A network exposed to training regime $a$ may be trained on a pool of data that consists of a subset of the base data set generated by distorting the prototypical patterns by flipping the sign of elements chosen at evenly-spaced intervals with a probability of .01. Or under training regime $b$ we may expose another network to a pool of data that consists of a different subset of the same base data set where, this time, distorted patterns are generated by flipping the sign of different elements, again chosen at evenly-spaced intervals, with the same probability. On the other hand, consider a hypothetical network trained on an inductively weak regime—call it training regimes $c$. Under training regime $c$ we may expose the network to a pool

[^21]: Many thanks to Aarre Laakso for bringing this point to my attention.
[^22]: It goes without saying that the input patterns presented to a pair of networks under inductively weak, and inductively robust training conditions, respectively, must belong to the same base data set. Notice that no correct generalization can be learned outside a given training space—cf. Elman (1998). The dubbing ("inductively robust" versus "inductively weak" training regimes) is due to Bill Casebeer. The boundaries, on the other hand, between inductively weak and inductively robust training regimes are not rigid. As Laakso (email communication) points out, “there are both procedural and statistical techniques for improving a network’s ability to generalize from a given set of data. The procedural techniques include more frequent presentation of input patterns from poorly represented regions of state space. The statistical techniques would include adding a “penalty term” to the backpropagation error that is proportional to the variance in the output unit activation. This has the effect of “regularizing” the learning function to improve network generalization in regions of state space that are over-represented in the
of data that consists of a different subset of the same base data set, but this time distorting prototypical patterns by flipping the sign of a number of elements *chosen at random* with probability .5. Distortions, in this hypothetical situation, will be further away from the prototypes. Notice that, unlike in training regimes a and b, where the subset chosen spans the full range of possible outputs at evenly-spaced intervals, training regime c draws its examples from areas chosen at random from its input space, such that idiosyncratic variations are more likely to occur. A network trained under these conditions is exposed to less representative samples, and will thus require to see many more patterns before being able to induce the same regularity discovered by those networks trained under ‘inductively robust conditions.’

When the distorted patterns that we feed the networks with are close to the prototypes, the central tendency of the set of exemplars (that is, the hot spot—see section I, above) will produce the strongest response, and the hidden representational space will be partitioned in such a way that clusters get neatly defined around their prototypes. However, when the distortions are farther away from the prototypes, the training exemplars themselves will produce the strongest activations, and a larger number of different distortions will be required for the network to induce the correct mapping.

I shall argue next that Laakso and Cottrell’s technique can deliver high correlations *only* when measuring similarities across networks that have been trained on inductively robust regimes. Unfortunately for the friend of State Space Semantics, to address the problem of collateral information we need to compare networks that have training regime while simultaneously reducing network bias in areas of the state space that are under-represented in the training regime. (See Bishop (1995), chapter 9, for details).”

23 Obviously, if the distribution of data is not representative at all of the full range of possibilities, the network will learn a somewhat different function. For argument’s sake I assume that the network ends up
been trained under inductively robust conditions with networks trained under inductively weak conditions, I contend. Let me elaborate.

Following the above distinction between inductively robust versus inductively weak training regimes, we may interpret the collateral information problem in at least two ways. In a best-case scenario for Churchland, we may identify the histories of two individuals with two networks both trained under different, although inductively robust, conditions. In that case, even though the networks are exposed to different input patterns, they will partition their state spaces similarly by sampling, for example the base data set at different evenly-spaced intervals (as training regimes a and b, above, exemplify). A high degree of correlation would then be expected. Given that the different training sets span the full range of possible outputs, we may expect the set of internal relations among points in one hidden space to be isomorphic with the set of internal relations in the other network’s hidden space. In a worst-case scenario, however (although see fn. 24), one network would be trained in an inductively robust regime, and the other network in an inductively weak regime. These networks will be trained on input patterns that are, respectively, more and less representative of their common task. I take it that this is a more plausible interpretation of the collateral information problem. After all, my concept dog and your concept dog may plausibly be associated with highly divergent inferential regimes—think of me as a dog breeder (an inductively robust training environment), and you having had spare encounters with dogs in your life (an inductively weak training environment) who has eventually come to be able to tell dogs from non-dogs as well as I can. Unfortunately, under this second scenario (inductively below a pre-established area in the error landscape, so that it performs optimally. There are further technical subleties that we may obviate for present purposes.
robust versus inductively weak training regimes), we will expect (although see below) to find a relatively low correlation across our networks. Networks trained under inductively different conditions will be exposed to different, and maybe orthogonal, experiences on input pools of data—while acknowledging the fact that a public concept is being shared. Assuming that Fodor and Lepore’s challenge is to be read in this way, different networks are highly unlikely to come out with similar solutions, or to partition their state spaces in similar ways (cosmic flukes apart!). The set of internal relations among points in one hidden space won’t be isomorphic with the set of internal relations in another network’s hidden space.

In fairness to the friend of connectionist semantics it must be stressed that the issue remains an open empirical question. Nevertheless, let me end up with the following comment from Laakso and Cottrell on their current results; a comment that (temporarily) seems to back the theoretical point being raised here against the defender of State Space Semantics:

Our networks might have had more or fewer categories (output units) in their repertoire; some might have been trained to distinguish only two colors; others to distinguish six or more. If that had been the case, would it have made sense to compare representational similarity between networks with different numbers of categories, and what would the results have been? Because we have not yet done the experiments, we cannot report the results. One can, in any event, use our technique to compare the representational structure of two such networks, because we can still present them with the same stimuli. We would hypothesize that the representational similarity between networks trained on different numbers of color terms would be rather low, since networks that had to learn more categories would also have to create more partitions in their hidden unit activation spaces. (Ibid., p. 74)

I conclude that a connectionist sympathiser of Churchland could not make use of Laakso and Cottrell’s neurosimulations to address Fodor and Lepore’s collateral

---

24To be precise, comparing networks both being trained under diverse inductively weak regimes would furnish a more realistic setting to address Fodor and Lepore’s challenge. Although to make the point
information challenge. Obviously, the above considerations are not unsurmountable, and, in fairness to Churchland, the case against State Space Semantics is less than conclusive. For one thing, someone may be ready to bite one or other of Fodor and Lepore’s bullets, and grant either an analytic/synthetic distinction, or empiricism. However, since Churchland explicitly rejects both options, I didn’t consider them in the present discussion. On the other hand, it would be good, ideally, to run the different kinds of training regimes described above with subsets of Laakso and Cottrell’s data. Unfortunately, Laakso (personal communication) worries that we might run into the problem of not having enough data in any given subset to train on. It would certainly be interesting and worthwhile to run those experiments with an expanded version of Laakso and Cottrell’s data, and that could be a topic for future research.25, 26

FRANCISCO CALVO GARZÓN

University of Murcia

REFERENCES


more vivid, I shall compare inductively robust with inductively weak training regimes.

25 An alternative would be to generate a larger (artificial) set of data in the way I described above—points (i)-(v)—and run control experiments similar to those run by Laakso and Cottrell. Laakso, nonetheless, would be willing to bet that the expanded pool of data would not make any difference with respect to the results they report.
26 I am grateful to Bill Casebeer, Paul Churchland, Gary Cottrell, Jim Edwards, Jeff Elman, Evan Tiffany, and an anonymous referee for *Mind and Language* for helpful comments and suggestions. I am especially indebted to Aarre Laakso for many helpful discussions, and comments on earlier drafts of this paper. An earlier version of this paper was presented at the 8th Italo-Spanish Meeting of Analytic Philosophy in Santiago de Compostela (Spain) in November of 2001. I wish to thank the audience at this meeting as well. This research was supported by a grant from *Fundación Séneca*.
Churchland," *Philosophical Psychology* 13, pp. 77-95.


Hornick et al. (1989) "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks* vol. 2, pp. 359-66.


Fig. 1: Hidden representational space regions of categories A, B, and C with regions of prototypical A-like, B-like, and C-like vectors highlighted.