Behavior and Philosophy, Fall/Winter 1991, Vol. 19, Number 2

ON A MODEL FOR PSYCHO-NEURAL COEVOLUTION

Bernard W. Kobes
Department of Philosophy
Arizona State University

ABSTRACT: According to a model of inter-theoretic relations advocated by Patricia S. Churchland, psychology will need to revise its theories so as to fit them for "smooth reduction" to the neurosciences, and this will lead to the elimination of reference to intentional contents from psychology. It is argued that this model is ambiguous; on one reading it is empirically implausible, on the other its methodology is confused. The connectionist program NETtalk, far from exemplifying the model as Churchland claims, suggests a theoretical rationale for employing relations to intentional contents in psychology.

It is surely naive to suppose that all branches of information-processing psychology and all branches of the neurosciences will, or should, develop in complete independence of each other. But the exact nature of the expected coevolution of theories is controversial and conceptually slippery. Some recent advocates of a reductive conception of the unity of science have given reasons, drawn from the history and philosophy of science, to expect that theories of lower-level neural processes will explain the processes and subsume the explanations of the higher level. The thesis is often put descriptively, as a prediction: the neurosciences will eventually subsume psychology. But given an assumption of progress in science, such claims are not merely predictive. They carry a normative load: higher-level processes are ultimately best explained by lower-level processes, and theory construction in psychology ought to acknowledge this.

Such an account of the nature of coevolution has been put forward in the work of Patricia S. Churchland, Paul M. Churchland, and Clifford Hooker. For specificity I will focus on Patricia Churchland's book *Neurophilosophy*[1], and on an analysis by her of the widely advertised connectionist program called "NETtalk". Many critics have fastened on Churchland's (as I shall simply call her) eliminativism with respect to folk psychology. Instead, I want to focus on the more fundamental issue of the nature of coevolution between the neurosciences and information-processing psychology. This issue bears on the expected knit between information-processing psychology and neurally inspired or "parallel distributed processing"

AUTHOR'S NOTES:

Thanks to Kathleen Akins, Stewart Cohen, Steven Hoffman, Steven Reynolds, ASU's "Neurolunch" group, and a journal referee for comments and discussions. Robert McCauley gave me a useful suggestion about NETtalk; I do not know if he would approve of the result. I am also grateful to Arizona State University's College of Liberal Arts and Sciences for a summer research grant. All correspondence concerning this article should be addressed to the author at the Department of Philosophy, Arizona State University, Tempe, Arizona, 85287.

accounts of cognition; indeed, the dispute between Churchland and myself will come to its sharpest point over "NETtalk". In that case, I shall argue, the need to posit relations to intentional contents can be seen to be linked to our need for a certain sort of explanation of success and failure.

REDUCTION-DRIVEN COEVOLUTION

Churchland advocates and predicts coevolution between the neurosciences and information-processing psychology: theories and phenomena from each field will help shape theories and our conception of the phenomena in the other. Thus weakly stated, the thesis is not very controversial, and I will not challenge it. I will treat it rather as setting the stage for a methodological and philosophical inquiry. How does such coevolution occur? What does it tell us about the explanatory categories of psychology?

Neurophilosophy aims to sketch a unified theory of the "mind/brain", which in Churchland's usage does not stand in for the conjunction "mind and brain", but refers to a single subject of scientific investigation. To this end she gives us a history of neuroscience from Galen to Golgi, and gives us surveys of what is known about neuronal firing, neurochemistry, functional neuroanatomy, the laminar structure of the cortex, topographic maps in the brain, clinical lesions, cerebral specialization, lateralization, event-related potentials, new technologies for studying the brain, and some philosophically puzzling cases in clinical neurology. But she ignores what we now conceive of as non-physiological psychology: branches of perceptual psychology and the psychology of memory, the psychology of language, personality theory, social psychology, cognitive-developmental psychology, and the psychology of inferential and practical reasoning, all of which surely have to do with the "mind/brain".

A distinguishing characteristic of the disciplines in the second group as currently practiced is their willingness to posit intentional states and events in the explanations of their characteristic phenomena. An intentional state or event (e.g., a belief or desire) is a relation between a person and a content specifying what the state or event is *of* or *about*. As a result of coevolutionary pressure, Churchland holds, psychological models attributing such intentional states and events as explanatory posits will be dramatically revised. We should expect a unified theory whose key explanatory constructs are neurally rather than linguistically inspired: perhaps neural networks, or vector phase spaces whose dimensions correspond to neural structures.

More precisely, the disciplines of the second group employ in their *explanantia* relational predicates (e.g., 'thinks that ___', 'perceives that ___', 'desires that ___', 'hopes that ___', 'fears that ___', or technical analogues thereof), that relate subjects to elements in turn related semantically or logically to each other. These elements I have called intentional contents. Coevolution, driven by the need for

reduction, will pressure these disciplines into abandoning predicates of this type in favor of predicates expressing "numerical attitudes" (e.g., 'has a mass-in-grams of ___', 'has energy-in-joules of ___', 'fires at vector point ___', and the like), which relate subjects to elements in turn related arithmetically or algebraically to each other (p. 304). Thus will unity of explanation issue in unity of type of relational predicate.

Churchland's account of inter-theoretic relations derives in its main features from logical empiricism.[2] Central to the account is the notion of reduction, which is a relation between a *reduced* theory T_R (e.g., non-physiological information-processing psychology) and a more *basic* theory T_B (e.g., some branch of the neurosciences). Reduction involves finding a theory T_A, in the vocabulary of T_B, such that T_A is an "analogue" (or "image") of T_R, and T_A can be logically deduced from T_B together with certain statements specifying limit conditions and idealizations (pp. 282-283, 310). Thus each putative law of T_R corresponds more or less well to some image statement in T_A, which though not a law of T_B, is derivable in T_B (p. 283).

Churchland modifies the standard logical empiricist account in allowing that the old high-level theory may have to be modified more or less drastically in order to accommodate the demands of the reduction (p. 282). In the best case, when the properties and processes in the high-level theory find exact analogues in the lower-level theory, she calls the reduction "smooth". Sometimes there will need to be a large amount of revision before smooth reduction can take place.[3] Thus we have an account of what drives coevolution between theories at different levels. If the processes and properties of T_R don't "line up" properly with those of some analogue T_A expressible at the more basic level, the high-level theory is under pressure to revise its postulated entities and processes in a direction that facilitates reduction. The reduction *per se*, in which inter-level identities are specified, is an empirically trivial consummation; all the scientific interest lies in the pre-reductive "fitting and revising" (p. 285).

Churchland allows that coevolution can operate from the top down, in that high-level theory may guide construction of hypotheses at the low level and our conception of what are the phenomena to be explained. (On this ground she vigorously rejects the charge of being "anti-psychology".[4]) But a notable asymmetry remains. The role of what we now consider non-physiological psychology seems to be that of identifying phenomena to be explained. The role of the neurosciences is to explain them. "Higher-level capacities and structures will be explained in terms of lower-level capacities and structures..."(p. 297;see also p. 283,p. 296). So reductive explanation is bottom-up explanation. "Upward" coevolutionary pressure is driven by the need to provide smoothly reductive explanations. "Downward" coevolutionary pressure is driven by a stream of *explananda*.[5]

We may distinguish two ways in which reductive pressures might be held to warrant the elimination of reference to intentional elements in psychological

explanation. For we may ask: is it supposed to be the case that, as a result of successful eliminativist reduction, we explain at a lower level (the level of T_B) what we formerly and misguidedly tried to explain at a higher level (the level of T_R)? Or is it rather supposed to be the case that we obtain a multi-level theory that includes some radically new explanation at the higher and characteristically psychological level, but also systematically relates high level entities and processes to complexes of entities and processes at the lower level?[6]

Some aspects of Churchland's discussion suggest the first, stronger interpretation. The interpretation certainly has the advantage that it makes the nature of the coevolutionary and eliminativist pressure plain. Reference to intentional elements may be eliminated because intentional explanation is pitched at too high a level, and empirical inquiry yields no way to construe the vocabulary of information processing psychology as "shorthand" for complexes of neural vocabulary. This seems to be Churchland's picture when she writes: "Should the old theory [T_R] be largely displaced, then bridge laws are typically dispensed with" (p. 282)--bridge laws being strict identity statements connecting entities as described in the vocabulary of a low-level theory with entities as described in the vocabulary of a high-level theory. And she writes: "The analogue [i.e., T_A] can then be logically deduced from the reducing theory T_B plus sentences specifying special conditions (e.g., frictionless surfaces, perfect elasticity)...Under these conditions, the old theory reduces to the new." (pp. 282-283) If the "old" theory does not reduce smoothly, it is replaced by a "new" theory couched in the vocabulary of the more basic science (e.g., neuroscientific vocabulary). The possibility of a "new" theory at the higher level is neglected.

Elsewhere Churchland's discussion suggests the second, weaker interpretation, for she emphasizes the multiplicity of strata in the brain and correspondingly in the neurosciences. Some of the higher levels of neuroscience must involve representation, computation, and "information processing" (pp. 298, 380-1, 457). Consequently she rejects David Marr's famous distinction between three independent levels of explanation: the computational level, the algorithmic level, and the level of neural implementation. Instead of only three independent levels, we get many mutually constraining levels. (p. 359) And on the interpretation of Churchland we are now pursuing, the distinction between levels survives the demise of old theory due to reductive pressure. We may agree that on occasion an old theory is eliminated because it attempted to explain at a high level what should instead have been explained at a low level, but that is not Churchland's claim about intentional explanation. Her claim is rather that the high-level theory evolves in response to the demands of smooth reduction, and that eventually such pressures will cause intentional explanations to be supplanted by radically new high-level theory.

Churchland's view on this second construal is more plausible than it was on the first construal in this respect: it treats eliminative coevolution as a special case

of the more general fact that we often construct multi-level theories relating distinct levels such that explanations within the higher level are a proper part of whole. The notion of "multi-level theory" I have in mind may be clarified by way of example. It is plausible that the theory of visual afterimages is pitched at a single, low level, while the theory of the visual ambiguity of the Necker cube is multi-level. No information-processing account explains why staring at a uniform red field results in a greenish after-image. The explanation refers only to the structure of cones in the retina, the bleaching of photopigments, and the like. By contrast, the visual ambiguity of the Necker cube is explained as a consequence of the fact that the visual system is confronted with two equally plausible but jointly inconsistent three-dimensional interpretations of a two-dimensional play of light on the retina. This theory makes essential reference to visual representation of the world. The full story adds to this an account of how visual representation is neurally instantiated, and an account of perceptual reversal in terms of fatigue in a bi-stable neural network. But the neural part of the theory supplements and illuminates the representational explanation; it does not replace it with a theory pitched at a lower level. In this sense the theory is multi-level.[7]

Plausible as it may be, the second interpretation leaves us with no convincing account of how pressure for smooth reduction *per se* could result in elimination of a type of high level explanation. For it is not constitutive of successful reduction that each high level property or process be identified with a complex of low level properties and processes that is a "natural kind" at the low level. Of course we could consider the complex a natural kind derivatively, in virtue of its role in reducing high level explanations, but it need not be a natural kind in virtue of its role in explanations at its own lower level. If this were a general requirement of reduction, then since each level of nature is supposed to be reducible to a lower level, any given high-level process must be identifiable with a natural kind at the level of fundamental physics--a highly implausible consequence.

But if the lower-level complex may seem unnatural or gerrymandered from the perspective of low level theory alone, then the demand that reductions be smooth is without epistemic force. For if low level naturalness is not constitutive of reductive smoothness, then "smooth" is too vague to offer epistemic guidance, and if it is constitutive, then smoothness is not a general methodological desideratum. Whether smoothness in this sense is desirable in a reduction cannot be answered with the generality to which philosophy of science aspires; it depends on evidential constraints that are peculiar to the domain of inquiry, notably including the evidential support T_R enjoys at its own level.

If the evidence might show that inter-level relations are extremely tangled in the case at hand, then how can the desirability of reduction have any coevolutionary teeth? Suppose an old theory T_R finds a low level analogue T_A employing a complex of low level properties tangled and undistinguished except by virtue of its role in reducing T_R; should we aspire to a high level theory T_S that reduces more

smoothly? Surely the answer depends on whether T_R can muster strong evidential support at its own level. If it can, then we have some reason to believe that reductive smoothness is not here in the offing, and is not desirable in any sense stronger than that of personal taste.

Let us take stock. We have identified two ways to interpret Churchland's account of how reduction-driven coevolution will radically transform psychology. The first does not comport well with her emphasis on levels of explanation within the mind/brain. But the second leaves us without a notion of reductive "smoothness" that is both methodologically desirable and has coevolutionary teeth. It is time to flesh out this abstract skeleton by reference to a particular style of empirical theory.

VECTOR THEORIES OF NEURAL REPRESENTATION

In support of her account Churchland discusses the notion of representation in Pellionisz and Llinas's tensor theory of sensorimotor coordination in the cerebellum. The theory concerns the neural mechanisms that underlie the subject's ability to catch an object that she sees tossed toward her. An input vector space is defined in terms of an array of neurons projecting into the cerebellum. Intuitively, these input fibers carry visual information about the spatial location of the object that the subject wants to catch. An output vector space is defined in terms of an array of neurons projecting from the cerebellum to certain motor centers. The function of the cerebellum in this case is to transform the visually specified "intention" vector into a finely tuned "motor" vector which causes accurate reaching.

On Churchland's view, representations in this case are points in the relevant vector spaces, and computation consists in transformation from coordinates of the input visual vector space to the coordinates of the output motor vector space according to a generalized mathematical function known as a tensor matrix. Her thesis is that the tensor matrix hypothesis is a computational hypothesis on a par with other computational hypotheses in psychology, and a suggestive model for coevolution:

> ... as a result of considering a theory of the basic principles of sensorimotor control, we can begin to see the shape of a new and powerful paradigm for understanding computational processes executed by the mind-brain and for understanding how the mind-brain represents at a variety of organizational levels. The multi-dimensional phase space appears to be a very powerful means for representation-in-general. (p. 457)

Suppose that we, like Churchland, are impressed with the power of vector phase spaces as a formal tool for describing high-level neural processing. What follows from this about the potential for coevolutionary pressure on intentional psychology? Note that the Pellionisz and Llinas theory concerns a process in which there is, in the language of the "old theory", at most one intentional content: both

the input vector and the output vector might be said to have the content "Reach here". Or better, since the input vector coordinate system is visually determined while the output vector coordinate system is identified by the motor response that it determines, the input vector has the content "Reach here$_v$" while the output vector has the content "Reach here$_m$". Successive coordinate systems constitute a succession of anatomically related and mathematically tractable ways of demonstrating the relevant location. The tensor computation is defined not over distinct contents but over distinct coordinate systems in which one content is represented. The representations are not about vector spaces or coordinate systems, just as the sentences of this paper are not about English or about sentences. But intentional explanations in information processing psychology typically involve a plurality of semantically related contents; it does not address the type of sensorimotor coordination of Pellionisz and Llinas's theory. So there is a dramatic gap between the explanatory domain of the tensor theory of the cerebellum and that of "old" theories of higher cognition.

Suppose that the mathematical formalisms associated with vector space theories, including the wide class of neural network theories, are sufficiently powerful to describe high-level neural processing. There remains the question whether the structures and processes of the mature system are neurally and anatomically "natural". If they are, this outcome might sustain eliminativist claims on the first interpretation: we will have explained at a neural level what we had misguidedly tried to explain at the higher intentional level. But the Pellionisz and Llinas example does little to shorten the odds against this empirical bet. We have no reason to suppose that the computation defined on the vector space would be neurally natural in domains where the old theory postulates more than one content.

A more likely result is that, if and when powerful vector space formalisms are successfully applied to high-level neural processing, they will yield structures and processes in the mature system that look tangled and gerrymandered from a neural perspective. In such cases the theory will seem incomplete. As I will argue below with respect to the program NETtalk, it will leave us feeling the lack of an explanation of how this system succeeds at its computational task. Thus we will be pushed to acknowledge the need for a higher level of explanation. The vector space description may be opaque with respect to real computational processes going on in the same spatio-temporal room as the domain of the vector space formalism. But now our criticism of the second interpretation of the coevolutionary thesis applies. For the explanation we seek will have to come from a theory justified by independent evidence available at its own level, rather than by the smoothness of its reductive relations to events described at the neural level. There is no reason to believe the reduction will be smooth in this instance, and so reduction *per se* exerts no special coevolutionary pressure on the high level theory.

NETTALK

This line of argument is best illustrated by reference to a particular connectionist model. Churchland and Terrence Sejnowski have recently focused on a system called NETtalk, designed by Sejnowski and Charles Rosenberg, which pronounces English text, and which is "perhaps the most complex network model yet constructed".[8] After completing a training period, NETtalk is able to take as input a written transcript of English sentences that it has not previously encountered and produce as output a phonemic representation of the text. The phonemic representation is hooked up to a voice synthesizer to produce comprehensible speech. Text pronunciation represents a non-trivial engineering achievement, since the correct phoneme typically depends not only on the immediately corresponding letter, but also and in complex ways on surrounding letters. For example, English maps the letter 'a' to different phonemes in 'hat' and 'hate', and the letter 's' to different phonemes in 'is' and 'this'. NETtalk masters these complex mappings.[9]

NETtalk has three layers of forward-feeding neuron-like processing elements, which I shall call "units" or "neurodes".[10] The input layer consists of 7 groups of units; each group is stipulated to represent a letter, and the 7 groups represent a text window consisting of the target letter and three adjacent letters on each side. The window "steps" through the text one letter at a time, and each time NETtalk computes a phoneme for the window's middle letter. Each of the output layer neurodes is stipulated to represent a phonemic distinctive feature, such as Voiced, Labial, or Stop, and a pattern of activity in the output layer thus represents a phoneme, such as /b/. Each input neurode is directly connected to each of 80 intermediate neurodes, and each intermediate neurode is directly connected to each output neurode. Each of these 20,000 or so connections has a modifiable weight, which may be positive (excitatory) or negative (inhibitory). The strength of firing of each intermediate and each output neurode is a non-linear function of the sum of the weighted inputs it receives from neurodes directly connected to it.

The weights are initially set at random, but NETtalk undergoes a period of "training up", during which its weights are gradually modified by a part of the system called the "Teacher". The Teacher continually computes a measure of the difference between NETtalk's output and the correct phonemic output, and, going backward through the network, adjusts each weight slightly so as to reduce its contribution to that error. Such a feedback algorithm is called "back-propagation". The Teacher does not "know" any of the rules of pronunciation; it only knows the correct phoneme for each window, and a perfectly general formula for reducing the contribution of each weight to the output error. In the course of massive quantities of training, NETtalk's output increasingly resembles human speech. In the end NETtalk has a welter of weights that collectively perform moderately well on English text, even text not in the training corpus. But in this welter one cannot

readily discern any morphological structure that could be said to represent a rule of pronunciation.

Wherein lies NETtalk's philosophical significance? Note that NETtalk is in the first instance an engineering achievement. No one claims that the system is psychologically or biologically realistic in its details. Compared to humans, the system has a markedly "toy" quality, since it elides several large components of reading. Typically a child does not learn to read until it has learned to understand and produce ordinary conversational speech. But NETtalk in no sense parses or understands what it pronounces. (Thus NETtalk cannot distinguish pronunciations that hinge on syntactic analysis or semantic interpretation. 'Lead', for example, is differently pronounced as a verb than it is as a chemical term.) Human learning-to-read builds on pre-existing articulatory skills and knowledge of pronunciation. It also requires development of visual recognition and visual motor skills, which NETtalk ignores. The feedback that NETtalk receives from the Teacher is different in quality, and greater in quantity, than the feedback a child receives from other speakers in pronunciation learning. Biologically, NETtalk's neurodes differ from neurons in many important respects. For example, in the brain, each neuron is such that its connections to the next layer are either uniformly excitatory or uniformly inhibitory. Moreover, no plausible neural mechanism for back-propagation is known.

The claimed philosophical significance is rather that NETtalk exemplifies broad principles of information storage in the brain. NETtalk does not employ what Churchland and Sejnowski call "the sentence logic model" of the mind. It does not pronounce text by manipulating symbol-tokens that express pronunciation facts or rules. Certainly Sejnowski and Rosenberg did not program rules of pronunciation directly into the system, as did the programmers of a text-pronunciation system named DECtalk, an expert system in the mainstream tradition of artificial intelligence. Instead, NETtalk's progress during the training period can be modeled as a tracing of a gradual meandering path through a hyper-space of almost 20,000 dimensions (one for each connection weight), until it settles into a location in weight space where its error rate is minimized. "[NETtalk] processes information by nonlinear dynamics, not by manipulating symbols and accessing rules. It learns by gradient descent in a complex interactive system, not by generating new rules."[11]

I want to focus specifically on the claim that NETtalk exemplifies revision of high-level theory driven by the need to fit and revise it in preparation for smooth reduction to neuroscientific theory.[12] We model NETtalk's behavior in terms of "the trajectory of a complex nonlinear dynamical system in a very high-dimensional space."[13] Since the dimensions are the weights between neurodes, this high-level model might be thought to bear the "smooth reduction" relation to low-level theory. Churchland disavows the claim that we understand NETtalk's behavior simply as the aggregate of enormously many particular causal interactions at the neurode level. She labels this view "Cajal's dream". Each time NETtalk is trained, a

different pattern of connection weights results. So it is quite clear that the aggregate of neurode interactions for a particular "training" of NETtalk provides no theoretical illumination. Yet Churchland holds that the high-level model must be bridged to a complex of posits and properties at the lower level. The notion of a dynamical system in high-dimensional space is explained by reference to processes of local causation at the neurode level, hence causation of a sort broadly familiar to neuroscience. We therefore modify our conception of the information processing involved in cognition, and we abandon the more specific postulation of relations to contents about a subject domain such as that of text pronunciation. The system's postulated employment of intentional rules, rules about the text-to-speech function, is eliminated in favor of more smoothly reducible posits. In this respect NETtalk points the way for all of what we now consider non-physiological psychology: the desideratum of smooth reduction dooms relations to intentional contents.

I shall argue that this interpretation of NETtalk's significance is mistaken. Of course it is a dangerous practice to extrapolate from general features of a system as psychologically crude and unrealistic as NETtalk to other branches of psychology. There are also genuine disputes about the psychological value of the wider class of neural network models. I shall not address these questions directly. Instead I want to make a more fundamental conceptual point about the notion of level in such systems.

We must distinguish the theory of how NETtalk is trained from the theory of its mature performance. The model of gradient descent in a high-dimensional space is a description of NETtalk's training. But the model leaves us epistemically undernourished: we are not yet told, except in terms of the utmost vagueness, how the fully trained system pronounces text. The network has programmed itself, but not perspicuously. The question concerns NETtalk's mature performance: how, we want to ask, does NETtalk encode the text-to-speech information that accounts for its success?

It is a truism of the philosophy of science that radically new theories do not so much answer old questions, or address old explananda, as show that old questions and explananda are misconceived. But evidently this question about NETtalk is not misconceived, for it is a question that motivates much of Sejnowski and Rosenberg's research. NETtalk correctly pronounces many words that are not part of its training set. Therefore it does not simply encode the pronunciation of each member of the training set; it must encode pronunciation regularities. But which regularities does it encode?

There is a sense in which any transformation can be computed by a connectionist network with a single layer of sufficiently many hidden units. For example, with N binary input neurodes, there are 2^N possible input patterns; given 2^N hidden units, with each hidden unit hooked up to detect one of the possible inputs, any desired function from input patterns to output patterns can be wired. But such a direct wiring is extremely costly in the number of hidden units it

requires. NETtalk is much more efficient than such a direct wiring algorithm would be, because it exploits redundancies in the text-to-speech function. Economy is all-important, and so therefore are pronunciation regularities.[14]

Sejnowski and Rosenberg estimate that the mature network can be defined by 80,000 bits of information, whereas the pronunciation dictionary of 20,012 words that NETtalk masters is defined by 2,000,000 bits of information. NETtalk must achieve this economy by discovering and exploiting regularities in English text-pronunciation.[15]

Our need for a theory of pronunciation regularities encoded is also highlighted by the fact that connectionist networks generalize best with an optimum number of hidden units. Networks with too many hidden units, as well as networks with too few, generalize poorly in their responses to stimuli not in the training set. (The precise optimum depends, of course, on the problem the network solves.) With too few hidden units the network lacks sufficient capacity to form an adequate partition of the training set itself. But a network with too many hidden units finds a lazy solution: it forms for each particular member of the training set an independent representation and response, rather than discovering a computationally efficient general strategy or heuristic for partitioning the training set. The lazy solution generalizes poorly to novel stimuli.[16]

The point is that, by contrast, a network with an optimum number of hidden units must discover some strategy that allows it to categorize novel stimuli more or less correctly. For a given working network we would like to know exactly what strategy it uses. We want to ask, in other words, the sorts of questions that cognitive psychologists ask about humans. This epistemic thirst cannot be slaked by neurode data alone, even when modeled on a large scale, as for example by the model of learning as gradient descent in weight space. We need, in addition and at least, an analysis of the text-pronunciation regularities that NETtalk encodes.

It might be objected that NETtalk merely behaves in a way that can be described by the pronunciation regularities, or that the strategy is merely part of the theorist's strategy for understanding the system, but that NETtalk does not stand in any real cognitive relation to either. This opens up a large and venerable debate; but the response strikes me as unmotivated. The regularities and strategy are attributable to NETtalk in virtue of complex and changeable internal modifications that the system undergoes, even though the relevant internal states are not morphologically salient at the neurode level. In some cases the attributed regularity might be strictly false, or the strategy unreliable, so it is plainly not merely the theorist's belief or explanatory strategy. And the regularity or strategy would seem to function as a general resource usable across a certain range of textual inputs, just as beliefs or problem-solving algorithms do in people. So it is hard to see how the theorist could use the pronunciation regularity or strategy without assuming that NETtalk stands in some real cognitive relation to it.

"NETtalk is a fortunate 'preparation'."[17] That is to say, during training the system acquires some sort of strategy or program (in the broadest sense), and it is a promising research project to treat NETtalk as if it were a psychological subject whose neurodes can be poked and probed, without ethical qualms, for clues as to what strategy it has come up with. In fact, Sejnowski and Rosenberg performed the following manipulations. Given a particular letter-to-sound correspondence, e.g., the correspondence between the letter 'c' and the soft-c sound /s/, they found the set of all 7-letter windows whose middle letter was a 'c' pronounced as /s/. Each 7-letter window in the 'c' → /s/ set gives rise to a characteristic pattern of firing in the array of hidden units. Averaging these over the set yields an average pattern of hidden-unit firing for the 'c' → /s/ correspondence. This whole process was repeated for each of 79 letter-to-sound correspondences, yielding 79 average patterns of hidden-unit firing. Cluster analysis was used to obtain a hierarchy of similarity relations among the 79 average patterns of hidden-unit firing. It turned out that the hidden-unit firing averages for all correspondences involving vowels are very similar. And within the vowels there are further similarities; e.g., the hidden-unit firing averages for correspondences involving the letter 'o' cluster together. There is also a rough clustering by phonetic feature within the hidden-unit firing averages for correspondences involving consonants.[18]

Churchland reports this research, but fails to appreciate its philosophical significance. That significance, as I see it, is twofold. First, we are not epistemically content with saying that NETtalk learns to pronounce text by gradient descent to a low-error position in a many-dimensional weight space. This already should make us suspect a confusion in Churchland's attempt to make gradient descent displace older theoretical posits such as cognitive strategies. For the extent to which the clustering result is scientifically illuminating, or explanatory, is the extent to which it illuminates a strategy that NETtalk exploits. NETtalk's exploitation of a strategy displays a kind of quasi-intentionality, since it is "about" text-pronunciation and is "directed at" getting the correct phoneme for each window.[19]

Second, the strategy that NETtalk exploits has the status of a strategy only against the background of a notion of correct text-pronunciation, which is a social construct and which the theorist must simply presuppose. NETtalk's acquisition of a strategy derives its integrity as an explanatory posit partly from its role in explaining NETtalk's successes and failures as measured against the presupposed notion of correctness. This fact gives the posit a degree of immunity from a strict demand of smooth reduction to a complex of neurode-level posits. Note the intrinsic complexity of the cluster-analysis manipulations. The results of the manipulations would be pointlessly gerrymandered and recondite, from the standpoint of large-scale neurode-level description, but for their potential to illuminate a strategy.

The value of NETtalk *qua* collection of linked and weighted physical neurodes occupying a low-error position in weight space is as a convenient "preparation" for further study, and as an existence-proof of the bare compatibility of high level quasi-

intentional explanations with neurode-level explanations. Its value is not that of a low-level foundation for legitimating explanatory kinds. If we must draw methodological morals from NETtalk to what we now consider non-physiological psychology, two reasonable morals to draw are these: wherever psychology deals with states or responses that are characterized as correct or incorrect as measured against a public standard, there is reason to exempt its explanatory kinds from the straightjacket of smooth reducibility to neural kinds, and prima facie reason to expect that relations to intentional contents will continue to figure in psychological explanations in the future.

As I have noted, Churchland rejects what she calls Cajal's dream, and she emphasizes the many strata in the brain and correspondingly in the neurosciences. There are levels corresponding to the study of brain biochemistry, membranes, single cells, cell circuits, brain subsystems, brain maps, and the entire central nervous system. Each level studies structures that serve functions identified at the next higher level.[20] Thus her hierarchy of levels of explanation is organized along the scale of spatio-temporal room. We might call this the "powers of 10" conception of levels.[21] Consequently on her view a representation, for example a many-dimensional weight space, is an information-processing construct at large spatio-temporal scale, which must be properly bridged to neurode or neural processes at a smaller spatio-temporal scale, and whose intentional content is either non-existent or explanatorily idle. I have argued that this conception leaves us epistemically malnourished. We want to know about the strategy that NETtalk uses, but this has a quasi-intentional character, and is identified in part by its role in explaining successful performance, where success is publicly characterized. This epistemological source of the hypothesis of a strategy gives it a degree of immunity from the straightjacket of smooth reduction beyond any that could be conferred just in virtue its being located at a large spatio-temporal scale.

SENTENTIALISM AND "THE COMPUTER METAPHOR"

Churchland argues that digital computers of the sort widely used in standard artificial intelligence (AI) research will no longer be seen as a fruitful metaphor for understanding the mind. In particular, she argues at length that there is little evidence for neurally characterized sentence tokens that are manipulated by a cognizer in a way analogous to the way that an AI computer manipulates syntactically characterized data structures.

However, the appeal to propositional attitudes *per se* in cognitivist explanations does not need a backing of neural sentence tokens. Propositional attitude explanations do not attribute causal interactions between people and *symbols*, or between people and *propositions*, any more than explanations in physics attribute causal interactions between objects and *numerals*, or between objects and *numbers*. Instead, they attribute causal/explanatory relations among propositional

attitudes themselves, just as physics describes causal interactions among numerically described physical states and events themselves. So propositional attitude explanations survive the demise of neural sentence tokens.

The manipulation of sentence tokens according to their formal or syntactic features is a leading idea behind the implementation of cognitive processes in AI research. I am sympathetic with Churchland's rejection of this aspect of "the computer metaphor". However, other more fruitful computer metaphors exist: the notion of machine simulation in automata theory, and the concept of a "virtual" machine in operating systems technology. These concepts show that it can make sense to speak of a plurality of machines occupying the same spatio-temporal region. If one merely points to a running AI computer and says 'that machine...', in the absence of an appropriate context, the reference will be ambiguous. Computing theorists speak of LISP as constituting a machine, and a particular AI program written in LISP as constituting yet another machine. It is easy to conceive of an alien investigator studying a high-level machine and being misled through pitching his investigation at the wrong level; the structure of the higher machine may not be greatly illuminated by the structure of the lower machine. For example, if the alien--someone not involved in creating the machine--wanted to explain a LISP program's success in drawing reasonable inductive inferences, or in identifying the referent of a pronoun (assuming that a LISP program could do this), an inquiry directed at understanding processes at the level of circuitry would be misdirected. Even an inquiry directed solely at uninterpreted LISP formulas would be misdirected. The explanatory value of the AI program for psychology or psycholinguistics depends on our understanding the machine as employing a strategy--depends, that is, on the investigator's attending to the intentionally characterized virtual machine.[22]

The point of present interest is that this aspect of "the computer analogy" carries over to non-standard artificially intelligent systems such as NETtalk, and illuminates the failure of the "smooth reduction" model of coevolution. In the case of PDP systems, unlike that of standard AI, we are relevantly like the alien investigator. Sejnowski and Rosenberg did not program a strategy into NETtalk--not directly, not under that description. Yet a presupposition of their further work is that NETtalk has acquired one, even if it is not smoothly reducible to neurode-level posits.

CONCLUSION

Articulating the nature of theoretical coevolution remains an interesting problem for further philosophical research. I doubt that the general problem is susceptible of a neat solution, for no one model, such as Churchland and the logical empiricists before her tried to construct, is likely to be adequate for domains as disparate as those of thermodynamics/statistical mechanics and psychology/neuro-

science. Insight will come, I suggest, from an analysis of the "deep sources" of the need to describe and explain at a given level.

In the domain presently of interest low-level facts have been shaped historically by high-level facts. Just as facts about anatomy and physiology have been shaped by facts about adaptation and reproductive fitness, so facts about the brain have been shaped by environmental and social facts. Similarly, *mutatis mutandis*, for NETtalk, in the course of training. The historical process gives purchase to the normative language of correctness and rationality. We need to explain cognitive successes and failures as characterized in this language. Even if the neural correlates of a putative type of relation to intentional content are so rough as to frustrate smooth reduction, there may be sources of evidence and explanatory coherence within the psychological theory such that the explanatory integrity of the intentional state is not impugned.

NOTES

[1] Patricia Smith Churchland, *Neurophilosophy* (MIT Press, 1986). Page references to this work will be given simply in parentheses.

The currency of the term 'coevolution' derives from *Neurophilosophy*. It is perhaps worth emphasizing that the term refers to a certain predicted and endorsed dynamic relation between neuroscientific and psychological theories, and not to some aspect of our biological evolution.

[2] After expounding the neo-empiricist account of reduction, Churchland writes, "I neither wish nor need to be doctrinaire about this. On the contrary, I adopt this theory only as a first approximation..."(p. 294), and she expresses unease at modeling explanations as logical relations between sentences.

Paul Churchland has since argued that "explanatory understanding" ought to be taken as more basic than "explanation", and that explanatory understanding ought to be identified with activation of a neural prototype-vector. See chapters 9 and 10 of Paul M. Churchland, *A neurocomputational perspective*, MIT Press, 1989. He suggests that reductive explanatory understanding is the apprehension of a subordinate neural prototype-vector as being an instance of a superordinate neural prototype-vector (*A neurocomputational perspective*, pp. 214-216). I believe that the features of Patricia Churchland's account that I am concerned with in this paper will be retained or find exact analogues in a detailed working out of the prototype-vector activation account of inter-theoretic reduction.

[3] For a similar and earlier view see section 11 of Paul M. Churchland, *Scientific realism and the plasticity of mind* (Cambridge University Press, 1979).

[4] See especially Patricia S. Churchland, "Reply to Corballis", *Biology and Philosophy*, *3*, (1988), pp. 393-397.

[5] See also pp. 373-374, where psychology and neuroscience are compared to "two rock climbers making their way up a wide chimney by bracing their feet against the wall, each braced against the back of the other." Despite the symmetric image, in the same paragraph neuroscience is clearly if implicity given epistemic privilege. Psychology's contribution to the joint task is apparently just to give "high-level specifications of the input-output properties of the system"--and even these specifications may turn out to be "virtual governors" [i.e. misconceived] in light of neuroscience.

[6] For a review of the literature on reduction and levels, see William Bechtol, *Philosophy of science: An overview for cognitive science* (Lawrence Erlbaum Associates, 1988), chapters 5 and 6. Especially pertinent in this context is Robert N. McCauley, "Intertheoretic relations and the future of psychology," *Philosophy of Science, 53*, (1986), 179-199.

[7] This example is drawn from David Marr, *Vision* (W.H. Freeman, 1982), p. 25. Marr uses the example for a slightly different purpose.

[8] Patricia Smith Churchland and Terrence J. Sejnowski, "Neural representation and neural computation," in *Neural connections, mental computations*, edited by L. Nadel, L. Cooper, P. Culicover, and R. M. Harnish (MIT Press, 1989). The article is likely to be influential in the philosophy of cognitive science, as it has already been reprinted twice, in James E. Tomberlin, ed., *Philosophical perspectives, 4: Action theory and philosophy of mind* (Ridgeview Publishing Company, 1990), and in William G. Lycan, ed., *Mind and cognition: A reader* (Basil Blackwell, 1990).

The quotation is from p. 236 of the Lycan reprinting, the pagination of which I also use below.

[9] See Terrence J. Sejnowski and Charles R. Rosenberg, "NETtalk: A parallel network that learns to read aloud", in James A. Anderson and Edward Rosenfeld, eds., *Neurocomputing: Foundations of research* (MIT Press, 1988), with editor's introduction, pp. 661-672.

[10] This use of 'neurode' for connectionist units (cf. the biological term 'neuron') follows Maureen Caudill and Charles Butler, *Naturally intelligent systems* (MIT Press, 1990).

[11] "Neural representation and neural computation", pp. 237-238 in Lycan's anthology.

[12] Churchland and Sejnowski (ibid. p. 229) repeat and endorse the account of reduction-driven coevolution I have described above.

[13] Ibid., p.234.

[14] See Terrence Sejnowski and Charles R. Rosenberg, "Learning and representation in connectionist models," in Michael S. Gazzaniga, ed., *Perspectives in memory research* (MIT Press, 1988), page 144.

[15] Ibid., page 161.

[16] This feature of network learning is reported by P. M. Churchland in *A neurocomputational perspective* (MIT Press, 1989), pp. 179-181, where it is attributed to Sejnowski, but used for a quite different philosophical purpose.

[17] "Neural representation and neural computation", p. 239.

[18] Sejnowski and Rosenberg, "Learning and representation in connectionist models", op. cit.

[19] In saying this I am, however, not committed to the relation between NETtalk and the strategy being one of explicit representation. See the following section. Nor am I committed to the relation being properly called by any mentalistic predicate of ordinary English such as "believes" or "knows".

[20] *Neurophilosophy*, pp. 359-360. See William G. Lycan, *Consciousness* (MIT Press, 1986), chapter 4, for a similar view.

[21] Christopher Cherniak calls this the "Scientific American" perspective; both terms play on P. Morrison and P. Morrison, *Powers of 10* (W.H. Freeman, 1981).

[22] For supporting arguments, see my "Individualism and artificial intelligence", in James E. Tomberlin, ed., *Philosophical perspectives, 4: Action theory and philosophy of mind* (Ridgeview Publishing Company, 1990), pp. 429-459.