# McCauley’s Demand for a Co-level Competitor

Paul M. Churchland and Patricia S. Churchland

McCauley’s is perhaps the most straightforward of the criticisms. He sees some form of reductive accommodation as the relation most likely to develop between propositional-attitude psychology, on the one hand, and the underlying neurosciences, on the other. In support of this expectation, he cites the typical co-evolutionary process described by Patricia S. Churchland, wherein theories at adjacent levels gradually knit themselves into some appropriate reductive relation or other. McCauley’s crucial move is then to claim that eliminative adjustments of theory are never (almost never?) motivated by considerations of cross-level conflict; rather, they are typically or properly motivated only by conflicts of theory at or within the same level of organization. In the absence of some compelling and comparably high-level alternative to folk psychology, then, we need not see folk psychology (FP) as facing any real threat of elimination. Accordingly, says McCauley, we should stand back and let the gradual interlevel knitting of theory proceed.

McCauley’s portrait of FP’s future may be correct. His guess is as good as ours, and a largely retentive reduction remains a live possibility. But the historical pattern he leans on is not so uniform as he suggests, and any probative classification of reality into distinct “levels” is something that is itself hostage to changeable theory. Consider the highly instructive example of astronomy.

For at least two thousand years (roughly from Aristotle to Galileo), the realm of the heavens was regarded as a distinct and wholly different level within the natural order. It was distinguished from the terrestrial, or sublunary, realm in several mutually-reinforcing ways. Sheer scale was the first difference, then as now. Thanks to Aristotle, Aristarchos, and Eratosthenes, even the geocentric ancients were aware that the moon was 240,000 miles away, that the sun was at least 5,000,000 miles away, and that the planets and the stars were more distant still. Astronomical phenomena evidently took place on a spatial scale at least four or five orders of magnitude beyond the scale of any human practical experience.

Second, the laws that governed our small-scale sublunary realm had neither place among nor grip upon the obviously special superlunary objects. They moved in their (almost) perfectly circular paths, according to their own laws, in a fashion that had no parallel within the terrestrial domain. Third, the realm of the heavens was immutable and incorruptible, in contrast to our own sorry domain. Centuries may flow by, but the heavens remain unaltered. Fourth and finally, the realm of the heavens was evidently the realm of the divine, the home or doorstep of the gods.

Accordingly, even the most casual of observers could appreciate that the discipline of astronomy was attempting to grasp a level of the natural order far beyond what the Lilliputian mechanics of falling stones, taut ropes, and rolling wagons could ever hope to address. Ptolemy was explicit in rejecting the aspirations of “physics” to explain astronomical phenomena, and his voice reflected an almost universal opinion. Astronomy was an autonomous science attempting to grasp the autonomous laws appropriate to the phenomena at a dramatically distinct level of the natural order.

Further, the ancient astronomical theories actually made good on this conviction. Aristotle’s account had a nest of 57 concentric earth-centered spheres, spheres made of the transparent and exclusively superlunary “fifth essence” (Plato’s cosmium), each moving at the behest of its own perfectly circular, perfectly uniform telos. Ptolemy’s different but similarly geocentric account had the familiar nest of perfectly circular deferent circles with eccentrically-placed centers, moving epicycles, and artfully placed "equant" points with which to cheat a bit on the issue of the perfect uniformity of astronomical motions.

We all know what finally happened to these ancient, "high-level" theories. They turned out to be radically false theories, so fundamentally defective that both their principles and their ontologies were eventually displaced, rather than smoothly reduced, by Newton's completed mechanics of motion (cf. the opening sentence of Churchland, 1981). Astronomy as a discipline is still with us, of course, and is more vigorous than ever, but it no longer speaks of crystalline spheres, fifth essences, moving epicycles, and phantom equants. An anisotropic, geocentric, rotating, finite spherical universe was displaced wholesale in favor of an isotropic, earth-indifferent, nonrotating, possibly infinite space. And the laws that govern the heavens turned out to be the very same laws that govern phenomena at the terrestrial level. They are the laws of Newtonian mechanics.

We present this as a presumptive counterexample to McCauley's claim that theories suffer radical displacement only at the hands of co-level competitors, and never at the hands of theories whose primary home is at a different level of scale or organization. Since the Newtonian revolution, modern astronomy has simply become the Physics of the Heavens. What remains, then, among the patterns of history, that would preclude modern psychology from simply becoming the Neuroscience of very Large and Intricate Brains? Perhaps brains differ from sea-slug ganglia only in the scale of neuronal interactions they involve.

We anticipate the reply, from McCauley, that this historical elimination of an ancient astronomical theory was not a cross-level displacement at all, but rather a displacement by a theory (Newtonian mechanics) that encompassed phenomena at the same dynamical level as the old theory. It is just that astronomical phenomena turned out not to be unique or special after all: They are distinguished only by their vast scale.

The reply has a point, and McCauley may succeed in pressing this interpretation upon us. But this reply entails what should have been clear anyway: that science can be profoundly wrong about what counts as a nomically-distinct level of phenomena, and profoundly wrong in its estimation of which theories do and do not count as genuinely "co-level" theoretical competitors. And if McCauley accepts this point, as we think it clear he must, then he is in no position to insist that the psychology/neuroscience case must turn out differently from the astronomy/physics case. Psychological phenomena, perhaps, are distinguished only by the unusual scale of the networks that display them.

Our conclusion, then, is as follows. The claim that psychology comprehends a distinct level of phenomena comprehended by a distinct set of laws uniquely appropriate to that level is not an assumption that our opposition can have for free. It is part of what is at issue - empirically at issue - in this broad debate, and the historical fate of ancient astronomy should caution against any premature convictions in its favor.

Astronomy aside, there are other historical examples that contradict McCauley's generalization about the agents of ontological displacement.1 Eliminative cross-level impacts on conceptual structure, both upward and downward, seem to us to be historically familiar, not rare or nonexistent. But we need not explore further examples here. Instead, let us explore directly the popular conviction that psychological phenomena really do belong to a more abstract level of analysis. If they do, would that really serve to insulate PP or other propositional-attitude theories from the threat of wholesale displacement?

Not in the least. Even if an abstract or higher-level explanatory framework were somehow essential to grasping psychological phenomena, it would remain an open question whether our current FP is the correct framework with which to meet this challenge. Legitimating the office need not legitimate the current office holder. This point is important because a priori there are infinitely many comparably high-level alternatives to FP; and because it is arguable that the conceptual framework of neo-connectionism is one of them.

As we sketched the fate of ancient astronomy a few paragraphs ago, it turned out that astronomical phenomena were not distinctly higher-in-level after all. But we might just as well have expressed the outcome by saying that the assembled laws of Newtonian mechanics turned out to be, when suitably articulated to fit the astronomical context, exactly the high-level theory that was needed to do the relevant high-level job. The analog of this latter stance, within psychology, will now be explored.

The claim on the table is that a psychological-level competitor for FP is already here and is already staring us in the face. It is the framework in which the occurrent representations are patterns of activation (or sequences of such patterns) across millions of neurons. It is the framework in which the computations are synapse-driven transformations of such patterns (or sequences thereof) into further such patterns across further neuronal populations. It is the framework in which such transformations are dictated by the learned patterns of synaptic connection strengths that connect one population of neurons with another. It is the framework, in short, of contemporary connectionist theory.

A frequent judgment about connectionist models of cognition is that they constitute at most an account of how classically conceived cognitive processes might be implemented in an underlying neural hardware. A quarter-century from now, we predict, this dismissal will be celebrated as one of the great head-in-the-sand episodes of twentieth-century science. Our confidence here is born not primarily of confidence in the ultimate correctness of connectionist models of cognition. (They must chance their hand to fate along with every other approach.) Rather, it is born of the recognition that the kinematics and dynamics of current connectionism already constitute an account of cognition at a decidedly abstract level. Allow us to explain.

When one sees a standard introduction to the connectionist modeling of cognitive processes, one is typically presented with a diagram of several layers of neuron-like units connected to one another by way of axon-like projections ending in synapse-like contacts (see figure 6.3.1, for example). One is then told about the variable nature of the weights of such contacts, about the multiplication of each axonal activation level by the synaptic weight it encounters, about the summation of all such products within the contacted neuron, and finally about the great variety of real-world discriminations such networks can be trained to make. We have given such accounts ourselves, and any audience can be forgiven for thinking that they are witness to an account of the underlying wheels and gears that might or might not realize the many abstract cognitive faculties that psychology presumes to study.

And so witness they are. But the real story only begins there, and strictly speaking that beginning is inessential. Neuronal details are no more essential to connectionist conceptions of cognition than vacuum-tube or transistor details are essential to the classical conception of cognition embodied in orthodox AI, Fodorean psychology, and FP itself. What is essential is the idea of fleeting high-dimensional patterns being transformed into other such patterns by virtue of their distributed interaction with an even higher-dimensional matrix of relatively stable transforming elements. The fleeting patterns constitute a creature's specific representations of important aspects of its changing environment. And the relatively stable matrix of transforming elements constitutes the creature's background knowledge of the general or chronic features of the world.

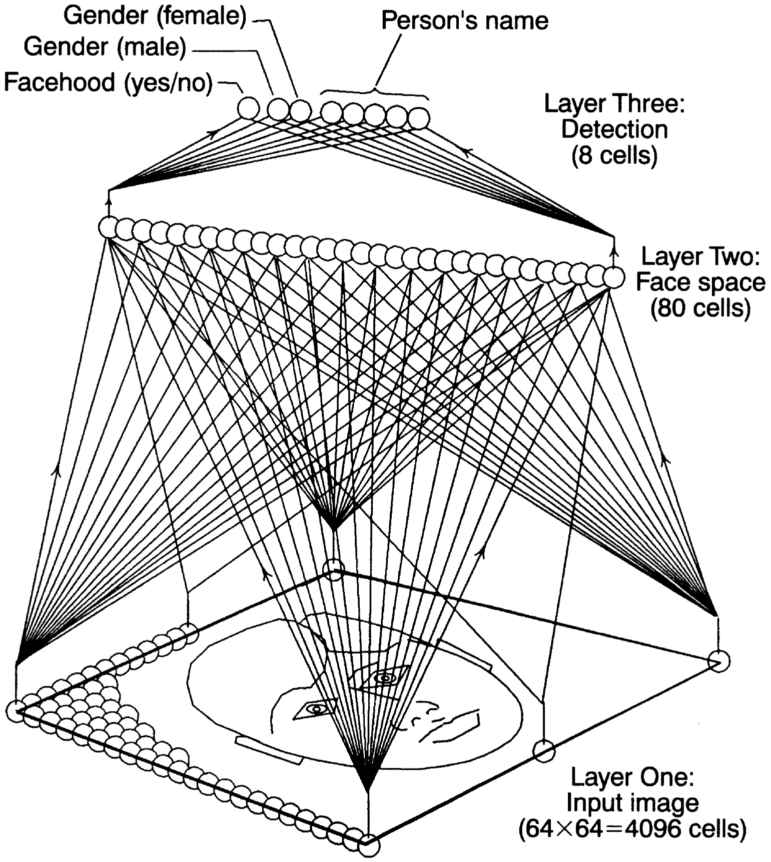
Gender (female)
Gender (male)
Facehood (yes/no)
Person's name
Layer Three: Detection (8 cells)
Layer Two: Face space (80 cells)
Layer One: Input image (64×64=4096 cells)

Figure 6.3.1 A feedforward network for discriminating facehood, gender, and personal identity as displayed in photographic images (adapted from Cottrell, 1991)

The abstract nature of this new conception of cognitive activity is revealed immediately by the fact that such activity can be physically realized in a wide variety of ways: in sundry biological wetwares, in silicon chips etched with parallel architectures, and even in a suitably programmed serial/digital machine, although this third incarnation exacts an absurdly high price in lost speed.

Its abstract or high-level nature is further revealed when we explore its kinematical and dynamical properties. Each population of elements (such as the neurons at the retina, or at the LGN, or at the primary visual cortex, and so on) defines a high-dimensional space of possible activation patterns across that population, patterns that are roughly equiprobable to begin with. But their relative probabilities gradually change over time as the system learns from its ongoing experience. Learning consists in the gradual modification of the many transforming matrices through which each activation pattern must pass as it filters its way through the system's many layers. Each matrix is so modified as to make certain activation patterns at the next layer more likely and other patterns less likely. The space of possible activation patterns at each layer thus acquires an intricate internal structure in the course of training.

Visual models are helpful here, and two standard display types are shown in figure 6.3.2. Their purpose is to illustrate the background cognitive state of the network of figure 6.3.1 after it has been trained to discriminate faces from nonfaces and female faces from male faces, and, within each gender, to recognize the specific faces of 11 named individuals displayed in the original set of training images (Cottrell, 1991).

The space in figure 6.3.2a represents the possible activation levels of three of the 80 units that make up layer two. As you can see, the space has been partitioned into a hierarchy of subspaces. Nonface images (strictly, nonface activation patterns) at the input layer are transformed into activation triplets at layer two (strictly, they are transformed into 80-tuples, but we are here ignoring 77 of those dimensions so that we can have a coherent picture to examine), triplets that always fall into the smallish subvolume near the origin of this 3-space. Evidently, most of the dynamic range of the units at layer two has been given over to the representation of faces. For all face images at the input layer get transformed into triplets that fall into the much larger subvolume to the right of the small triangular partition.

Within that larger subvolume is a second partition, this time dividing the range of activation triplets that represent female faces from the range of activation triplets for male faces. Activation triplets that fall anywhere on that speckled vertical partition are the network's mature responses to input-layer face images that are highly ambiguous as to gender. Activation triplets within each of the 11 small volumes scattered on either side of that partition represent slightly different photographs of the 11 different individuals represented in the training set. The network has thus developed six further subcategories within the male subvolume and five subcategories within the female subvolume. The relevant partitions have been left out of figure 6.3.2a so as to avoid visual clutter, but the 11 prototypical "hot spots" within each final partition are saliently represented.

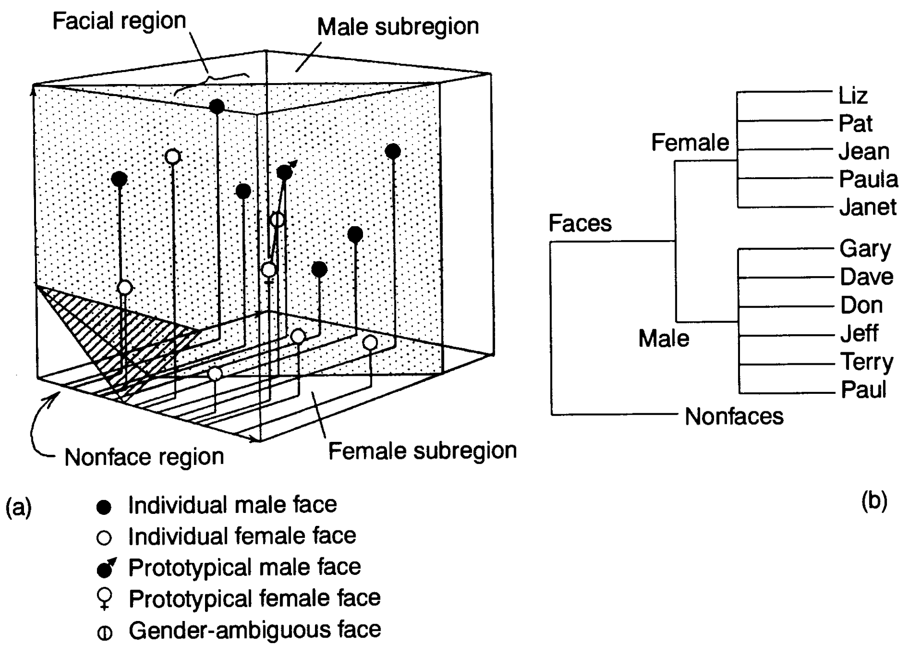

Figure 6.3.2 (a) An activation state space whose three axes represent three of the 80 units at the middle layer of face-discrimination network. The partitions into subcategories are visible, as are the small volumes that code input images for each of the eleven individuals variously portrayed in the training set This 3-space is frankly a cartoon, in that the distinctions displayed cannot effectively be

drawn within only three of the 80 dimensions available at Layer Two. If they could, the network would need only three units at that layer. In fact, only the bulk of those 80 units working together will draw all of those distinctions reliably. However, the 3-space does represent fairly the kind of partitioning that training produces, except that the true partitions are high-dimensional hypersurfaces rather than 2-D planes.
(b) A dendogram representing the same set of hierarchically organized categories. Since they make no attempt to portray partition surfaces and hypersurfaces, dendograms are indifferent to the dimensionality of the activation space at issue.

Figure 6.3.2a indicates how the regularities and variances implicit in the set of training images have come to be represented by an acquired set of structures within the activation space of layer two. The job of the network's Layer Three is now the relatively easy one of the discriminating just where, within this hierarchy of Layer Two subspaces, any fleeting activation pattern happens to fall. This it does well. Overall, Cottrell's network achieved 100 percent reliability on the (roughly 100) images in the training set, in facehood, gender, and individual identity. More importantly, its acquired perceptual skills generalized robustly to images it had never seen before. It remains 100 percent accurate on faces vs. nonfaces; it remains almost 90 percent accurate on arbitrary male and female faces; and to any novel face, it tends to apply the name of the individual among the original eleven to whom that novel face bears the closest resemblance, as judged by relevant proximity in the space of figure 6.3.2a.

What we are looking at in this figure is the conceptual space of the trained network. (Or, rather, one of its conceptual spaces. The fact is, a network with many layers has many distinct conceptual spaces, one for each layer or distinct population of units. These spaces interact with each other in complex ways.) We are looking at the categorial frame-work with which the network apprehends its perceptual world.

Here it is important to appreciate, once more, that it is the overall activation pattern across all or most of layer two that is important for the network's cognitive activities. Because each element of the network contributes such a tiny amount to the overall process, no single unit is crucial and no single synapse is crucial. If any randomly chosen small subset of the units and synapses in the network is made inactive then the quality of the network's responses will be degraded slightly, but its behavioral profile will be little changed. It is the molar-level properties of the network - its global activation patterns and its global matrix configurations - that are decisive for reckoning the major features of its ongoing input-output behavior. A single unit is no more crucial than is a single pixel on your TV screen: its failure is unlikely even to be noticed.

Evidently, this "vector/matrix" or "pattern/transformer" conception of cognition comprehends a level of abstraction beyond any of its possible implementation-level counterparts. It is not itself an implementation-level theory. The fact is, we have long been in possession of the relevant implementation-level science: It is neuroscience. Connectionism is something else again. What connectionism brings is a new and revealing way of comprehending the molar-level behavior of cognitive creatures, a way that coheres smoothly with at least two implementational stories: the theory of biological neural networks, and the theory of massively parallel silicon architectures. If McCauley insists upon a suitably high-level competitor for FP, fate has already delivered what he deems necessary. FP is already being tested against a new and quite different conception of cognition.

## Note

[1] First, the rather feeble conceptual framework of early biology - sporting notions such as telos, animal spirits, archeus, and essential form - was eventually displaced by an entirely new framework of biological notions (such as enzyme, vitamin, metabolic pathway, and genetic code), notions regularly inspired by the emerging categories of structural and dynamical chemistry, a science that addressed a lower level of natural organization.

Second, the molar-level theory of classical thermodynamics, which identified heat with a macroscopic fluid substance called "caloric," was displaced by the molecular/kinetic account of statistical thermodynamics, a theory that addressed the dynamical behavior of corpuscles at a submicroscopic level.

Third, the well-established conceptual framework of geometrical optics, while a useful tool for understanding many macro4evel effects, was shown to be a false model of reality when it turned out that all optical phenomena could be reduced to (i.e., reconstructed in terms of) the propagation of oscillating electromagnetic fields. In particular, it turned out that there is no such thing as a literal light ray. Geometrical optics had long been inadequate to diffraction, interference, and polarization effects anyway, but it took Maxwell's much more general electromagnetic theory to retire it permanently as anything more than an occasionally convenient tool.

Fourth, the old Aristotelian/alchemical conception of physical substance (as consisting of a continuous but otherwise fairly featureless base matter that gets variously informed by sundry insubstantial spirits) was gradually displaced in the nineteenth century by Dalton's atomic/structural conception of matter. Once again, we may count this an intralevel displacement if you wish, but it is clear that most of the details of Dalton's atomism - in particular, the relative atomic weight and the valence of each elemental atom - were inspired by higher-level chemical data concerning the intricate web of constant weight ratios experimentally revealed in chemical combinations and dissociations. Bluntly, a maturing chemistry had an enormous and continuing impact on the shape of a still-infantile atomism. In this case, note well, it was a higher-level science that was dictating our theoretical convictions at a lower level of natural organization.

## References

Churchland, P. M. 1981: Eliminative materialism and the prepositional attitudes. Journal of Philosophy, 78, 67-90

Cottrell, G. 1991: Extracting features from faces using compression networks: Face, identity, emotions and gender recognition using holons. In D. Touretzky, J. Elman, T. Sejnowski, and G. Hinton, Connectionist Models: Proceedings of the 1990 Summer School. San Mateo, CA: Morgan Kaufmann.