Aporia vol. 17 no. 2—2007

Bad Attitudes: Rethinking the Role of Propositional Attitudes in Cognitive Science

Jonathan Martin

Introduction

Mental representation is a central concept in cognitive science. It is the causal story involving "symbolic structures" that form the explanatory schema of classical cognitive science. Objects that populate the particular ontology of a scientific theory must have an instrumental role in explaining the processes in question. As such, it is requisite that the explanatory vehicles within a theory be in some way informative with regards to a particular inquiry. Due to this methodological presupposition, the entities invoked in a theory can be evaluated in terms of their explanatory value or their status as causal-explanatory entities altogether. In pursuing a more explanatory cognitive science, I will be advancing an eliminative program which addresses the need for representational language but questions the status of propositional attitudes as causal objects in computational explanations.

Propositional attitudes are sentences consisting of an intensional verb followed by a content clause; for example, "x believes that y." It is not my aim simply to reiterate Paul Churchland's claim that folk psychology is an empirical theory (and thus falsifiable). Rather, I will show that (1) propositional attitudes as causal entities are inadequate and largely uninformative for explaining representation computationally; and (2) understanding the dynamics of networks of neurons (though I will rely on artificial models) with relation to representation will require description in non-propositional terms using a dynamic neurobiological model. I will argue that the functional story, which will inform us about how the brain represents, will be

Jonathan Martin is a senior majoring in philosophy at the College of Wooster and is currently applying for graduate programs in philosophy. His academic interests are cognitive science, philosophy of mind, and philosophy of science.

concerned with causal neurobiological functions and will most likely account for representation in a highly non-propositional way.

Classical Minds: Folk Psychology and Propositional Attitudes

The accounts of the mind in classical cognitive science (of which Fodor's language of thought hypothesis is paradigmatic) have been committed to two fundamental propositions: (A) mental content is determined by the causal effects and input/output relations of a representational system; and (B) the causal story of A is best described in the language of propositional attitudes (PAs). I will focus on B. Formulations of mental events under this framework take the form of citing certain intensional states (beliefs, desires, hopes, etc.) as explanations for behavior. For example, it is Jack's desire to drive, together with his belief that his car is in the garage, that explains why he grabs his keys and heads to the garage. Though this action is surely dependent on other intensional states (he believes that his keys are necessary to start the car), the rough picture of this view of the mind should be clear. One can even present predictive counterfactuals such as "If Jack wanted to take a ride and believed his car was in the garage, *but also* believed that cars are very likely to explode at any moment, he would not grab his keys and head to the garage." The explanatory ability of giving such propositional reasons as causes, at first glance, seems obvious. Indeed, it is hard to conceive of what a denial of these reasons might even mean.

Still, there are further commitments of classicism to reveal before we proceed. One important commitment is noted by Andy Clark:

> The folk framework provides both a model of our computational organization and a set of contents [propositional attitudes] which have reasonably close analogues in the inner economy . . . [such that] strings of inner symbols can stand in sufficiently close relation to the contents cited in ordinary mentalistic discourse for us to speak of such contents' being *tokened* in the string and having causal powers in virtue of the causal powers of the string. (*Associative Engines* 6)

The classical view in cognitive science is committed to a system of "syntactically structured representations . . . [in which] the computational processes . . . can be described by transition or derivation rules" (46). While classicists such as Fodor might insist there should be only the requirement of a *correspondence* between a PA's proposition and the representational and structural processes of a cognitive system, the causal efficacy of these sentences in the brain is indispensable to the classical picture.

Concepts, Recombination, and Systematicity

As established, the classical cognitive scientist describes the inner workings of the human mind as a series of operations over syntactically defined representational states. Andy Clark cites a good, working definition from Newell and Simon:

> A physical-symbol system . . . is a physical device that contains a set of interpretable and combinable items (symbols) and a set of processes that can operate on the items (copying, conjoining, creating, and destroying them according to instructions). (*Mindware* 28)

The model of classical cognitive science takes its power from the general usefulness and breadth of abilities realized on serial processing computers (which are syntactic, physical-symbol systems) and the knack people seem to have for recombining words and concepts. For example, anyone who has the concept of a bird should be able to imagine a bird in different places (in a bush, in a person's hand, etc.). Thus, concepts are said to be atomistic in that they can be used in multiple domains or recombined scenarios. If one can understand "Sarah found the bear" then one should be able to understand "the bear found Sarah."

Another strength of the classical paradigm is that it provides a straightforward answer to the issue of the systematicity of cognition. Because of the symbolic nature of the representations in classical theories, systematic abilities (such as logic) are easily put into propositional form. For example, if David knows that all triangles have three sides *and* he knows that x is a triangle, then *ceteris paribus* he knows that x has three sides. The property of systematicity plays a large role in accounting for cognition in classical computational cognitive science because it provides a picture of what types of mental contents there are as well as the different kinds of computational transition rules that might be part of the computational hardware. This is meant to show why a person who comprehends "Kelly is a sister" and "sisters are female" will be able to deduce that Kelly is a girl. While the explanatory weight of this propositional view of mental content is clear, I will now address some objections to this view.

A Major Position on PAs and Objections

Though it is my intention to reject PAs as part of explanatory cognitive science, there are a few major positions which offer reasons for retaining these entities.

Fodor's Realism

Jerry Fodor's Language of Thought Hypothesis relies on the claim that PAs are the only possible causal constituents of explanatory cognitive theories. As such, he is committed to the idea that the ontology of PAs (beliefs, desires, fears, etc.) traces real symbolic entities in an "inner" computational economy. He states that functional cognitive states must be described as law-like relations between PAs, for explanations below the propositional level lose essential functional significance. He writes that

> having a propositional attitude is being in some relation to an internal representation. In particular, having a propositional attitude is being in some *computational* relation to an internal representation. . . . It won't be possible to construct a psychology of the kind that I have been envisioning unless organisms have pertinent descriptions as instantiations of some or other formal system . . . that for each propositional attitude . . . there is some causal state of the organism such that . . . the state is interpretable as a relation to a formula of the formal system, and . . . being in the state is nomologically necessary and sufficient for . . . having the propositional attitude. (198–99)

This statement leaves us with a fairly clear picture of how Fodor expects to vindicate our causal-propositional theories. According to this view, there are nomological relations between the structures of certain propositions and the inner symbols of a syntactic computational system. These relations and causal effects are generally consistent with the expectations of folk psychology. For example, suppose Karen believes it will snow tonight and also fears it will snow tonight. These two states are, on Fodor's account, two different causal effects of the same symbolic representational items. The representation must, according to Fodor, bear causal relationships appropriate to belief in one case, and fear in the other.

Objection 1 to Fodor's Realism

Jerry Fodor provides a clear idea of realism regarding the status of propositional attitudes. If his position is correct, then folk psychological explanations and predictions work because they give an accurate account of the computational states of cognitive functioning. But if the ontology of folk psychology is basically accurate and the explanations and prediction using PAs are dependent on this essential accuracy, then it seems we must be able to provide nomological relationships between the posits of folk psychological cognitive explanations. In other words, if folk psychology is an accurate account of the causal/computational states of cognition, then there must be law-like relations between certain mental contents. When the consequences of folk psychological predictions are not due to law-like relations, then they must be logical or analytic implications. Assuming that a certain bit of knowledge allows one to explain or predict certain outcomes, there must be nomological regularities that make the predictions or explanations informative. But if such implications are analytic, then the results of such predictions would be scientifically uninformative; one would be able to explain a state of affairs or predict an outcome purely through conceptual analysis with no need for empirical data.

In the pursuit of demonstrating that folk psychology is an empirical theory, Paul Churchland provides several comparisons between folk psychological predictions and familiar theories in physics:

> (1) if (x) (f) (m) [((x has a mass of m) & (x suffers a net force of f)) then (x accelerates at f/m)].
>
> (2) if (x) (p) [(x fears that p) then (x desires that $\sim p)$].
>
> (3) if (x) (p) (q) [((x desires that p) & (x believes that (if q then p)) & (x is able to bring it about that q) then (barring conflicting desires or preferred strategies, x brings it about that q)]. (*Philosophy of Mind* 570)

There is clearly a structural similarity in the relation of the implications Churchland uses to formulate these theories. But let us look at (2), as it is the most simple of his folk psychological theories. What is the relationship between x's fear that p and x's desire that $\sim p$? It seems that this *might* be an analytic implication, as "fear that p" might conceptually include the "desire that $\sim p$." If this is the case, then we are not making an important prediction or comprehending the relationship between two states, because both states are redescriptions of the same state (to fear p *is* to desire that $\sim p$). I will create a new law involving PAs that should further explicate this problem:

> (4) if (x)(p) (x hopes that p) & (x finds that p) then (*ceteris paribus*, x is pleased).

This relationship should hold in all cases of wishing and hoping, given *ceteris paribus* (all things being equal) conditions hold. But does this relationship hold because of nomological relations between hoping and being pleased or because the implication is analytic? Another consideration relevant to the empirical acceptability of folk psychological theories is that the theories rely heavily on *ceteris paribus* clauses (barring confusion, higher motives, etc.) and require that the attitudes (for example, wishing that x and being pleased that x) are conceptually and logically distinct. Here, we can ask some questions about the similarities and differences between these theories. Could force turn out to be something other than mass × acceleration, or would this have been a conceptual impossibility? Observation confirms the reliability of these physical terms and the law-like relationships between these entities as they occur in physical theories. Could we have arrived at this conclusion merely by analyzing the concept of force? To put it simply, the truth of folk psychological theories, if this account holds, relies on the fact that the implications are the result of analytic and not nomological relations. The prediction that (2) makes, namely that if x fears that p, then x desires that $\sim p$, can be arrived at through conceptual analysis, even in the absence of empirical observations.

Objection 2 to Fodor's Realism

My second objection to Fodor's propositional-computational account deals with the notion that satisfying computational accounts of cognition must be propositional so they do not miss the functionally significant features of the computational system. The argument I am objecting to is committed to the proposition that cognition is describable only at the level of abstraction found in folk psychological/propositional language. I shall begin with a basic example. Imagine a person playing a game of chess (a notably advanced cognitive ability). Each turn, the player engages in a process of deliberation in which a number of important cognitive abilities must be engaged in highly integrated ways in order to produce appropriate reactions to given scenarios. The process (I rely on some intuitions here) most likely involves an ability to discriminate pieces by shape. Now this discriminative capacity will necessarily be successful in a seasoned chess player given high levels of variation in the shape of, for example, knights. In examples of successful discrimination, the relevant features for the identification of chess pieces' shapes must be computationally related to the process of identification. Since the various relevant physical features must be identified primarily through retinal stimulation and admit to a high degree of physical variation (shape, orientations etc.), the computational system involved must be able to rely on a largely general discriminative capacity for using shifting patterns in retinal input. This is only the first step, for associating appropriate movements with given pieces is significant in the computational processes involved in strategizing over the next move.[1] In addition, note the generality of an account of this cognitive process citing PAs. On this account, the player's cognitive processes are most adequately described as a causal web of beliefs, desires, and strategies which

[1] As I will explore later, given the representational account in the non-propositional neurocomputational view (analysis of activation patterns, prototypicality judgments, etc.), the types of discriminative capacities and associative abilities here are well accounted for.

are nomologically related and produce causal effects roughly identical to our ordinary expectations. So let us cash out the intensional story which cites PAs. The player desires to put the opponent in check, believes certain strategies might work, and has sufficient beliefs about the shapes of pieces and their appropriate movements. But this account, at the propositional level of abstraction, would be true of any system capable of playing chess, no matter what goes on computationally in order to achieve the relevant cognitive task. Deep Blue, the famous chess-playing computer, and the human brain play chess successfully by virtue of entirely different types of computation. The intensional psychology of PAs, however, could be used with equal predictive success.

It does not take much analysis of the folk psychological explanation of what the player is doing and representing to see that the predictions and descriptions given are superficial, uninteresting, and not explanatory. In fact, it is the level of abstraction (cited by Fodor for the indispensability of PAs) that makes the theoretical success of PA attribution irrelevant to understanding anything but the most general computational features of a cognitive process. To put it simply, if we wish to understand how a certain cognitive process is achieved computationally, PAs will work regardless of the computational system involved. A person inclined toward the neurocomputational account will be able to specify computational functioning at a much more interesting and causally less dubious level of description. How do we identify chess pieces (given damage, partial obstruction, and variation across sets of pieces)? The answer, which includes a number of beliefs (knights look like horses, are shorter than the queen, etc.), will give a computational account of PAs, but will most likely prove inadequate at providing truly explanatory or nomologically predictive cognitive stories. The story might still be true in some computationally uninteresting way. The agent is described and predicted accurately as wanting to win or strategizing that x, but the truth of these higher-level descriptions and their predictive success must be *explained* at a lower level through the description of representation and cognitive functions best defined non-propositionally. In fact, when we look at similar examples of cognitive capacities in parallel distributed processing, we will find systems which store representations in a manner very incompatible with folk psychological descriptions and expectations. Still, this incompatibility does not seem to have any real effect on whether the attribution of PAs works on these non-propositional systems. The impotence of this incompatibility demonstrates my central point. A facial recognition network can be described in folk psychological/propositional language ("This network recognizes the face as Sarah"), but this seems to work no matter what the causal properties of the system. This is because the PAs describe cognition at a level of generality highly divorced from computational or causal processes. There is, however, a more telling and satisfactory account of this cognitive ability which occurs when we treat populations of neurons as a parallel computational system whose functioning is non-syntactic or propositional. The fact that PA attribution works as a tool for predicting cognition does not mean that it is a satisfactory way of providing computational accounts.

Getting to Know the Artificial Neuron Network

Since my examples employ artificial neuron network research, it is necessary for me to outline and describe the relevant features about these types of processing systems. The Artificial Neuron Network (ANN) is a cluster of interconnected individual processors which are organized into layers. These layers are often categorized as: (1) Input Units (2) Hidden Units, and (3) Output Units. Each input unit is respectively connected to every hidden unit, and each hidden unit is connected to either additional hidden unit layers or to the output units. The unit is a simple processor able to give and receive graduated outputs of some decimal value between zero and one. This description is admittedly simple, but it conveys the basic idea.

Now we come to what these simple processors do when their behavior is coordinated. Researchers in neural-nets make use of learning algorithms of which, for my purposes, only a rough introduction is needed. Through the use of learning algorithms, researchers are able to train ANNs to represent certain significant patterns in their inputs and subsequently to give outputs consistent with specific domains of representation. In order to do so, the network is given a training set—a fixed group of inputs—and the output is compared with some desired input-output relation. Over time, the network learns by reconfiguring the weight (from zero to one) of the connection to the other units through various methods of comparisons to correct input-output relations. I will provide an example that further elucidates this concept.

Putting a Name to a Face

Paul Churchland cites a powerful example of a three-layered network that learned to effectively represent various relationships in its inputs to a specific cognitive domain: facial recognition. Not only could this network quite accurately discriminate between faces and non-faces, but it could also determine the gender of the face and the person's name. Churchland writes:

> [The network's] input layer or "retina" is a 64x64-pixel grid whose elements each admit of 256 different levels of activation or "brightness." . . . [The] *training set* contained 64 different photos of 11 different faces, plus 13 photos of nonface scenes. . . . Each input cell projects a

> radiating set of axonal end branches to each and every one of the 80 cells in the second layer, which . . . represents an abstract space of 80 dimensions. . . . This second layer projects finally to an output layer of only eight cells. (*The Engine of Reason* 40–41)

The final eight units of the network give outputs (albeit numerical ones) which signify various features of the picture input. These correspond to the picture's being either: 1. (face/non-face) 2. (male/female) 3. (name: Sarah [for example]). Churchland writes:

> It achieved 100 percent accuracy, on the training set of images, with respect to faceness, gender, and the identity of the face presented. . . . A more severe and more relevant test occurs when we present the network with photos it has never seen before, that is, with various photos of the same people drawn from outside the training set. Here again the network comes through though. It identified correctly 98 percent of novel photographs of the people encountered in its training set, missing the name and gender of only one female subject. . . . A third and highly intriguing experiment tested the network's ability to recognize and identify a "familiar" person when one-fifth of the person's face was obscured by a horizontal bar across the input image. Surprisingly, the network's performance was hardly impaired at all. (*The Engine of Reason* 45)

In this case one can see that the synaptic weightings of the network have settled into a configuration which bears appropriate outputs to certain "visual" patterns on the input units. By this I mean that certain general relationships in facial input patterns are reflected in the synaptic weightings due to the learning process of repeated synaptic adjustment and not computed in terms of identifying common-sense facial features. Let us look at possible methods of understanding the processes involved in making these impressive judgments.

The Inside Story

Before we move on to issues of mental causation, we should look at a few of the key concepts for understanding the representational story of the Parallel Distributed Processor.

Representational Spaces

After hearing such a result, one might wish to understand just *how* this network is making these "smart" discriminations. There are a few key ways in which the network's "concepts" can be understood, and becoming familiar with these methods will reveal more about the representational story involved. By virtue of the network's development of appropriate connection weights, we can talk of the network as having an 80-dimensional facial "concept space" (Churchland, *The Engine of Reason* 46–47). This means that all possible faces that can be represented by the network can be located somewhere in a geometric "space" where distance between two possible faces will be indicative of the similarity or difference along some specific domains which are represented by the system. Each point in such a space stands for a specific pattern of neural activation in the network. If we were to understand the behavior of this computational system using the resources provided by the model of PAs, we would quickly find that this project fails as a causal story. If one were to cite the network's "belief that females have eye distance/forehead proportion *x*" and predict the system's outputs consistently, it would live up to the expectations of the intensional stance. But there is reason to believe that this is decidedly *not* a good theory of the system's computational processes. First, there is nothing remotely like such a belief being represented or playing a role in the network (there are no structures which are causal by virtue of semantic content). Second, such a theory of its computational economy would both likely fail to make interesting or informative predictions of future behavior and will make false predictions when tested more in depth. But perhaps most importantly, by analyzing patterns in the network, "one can see immediately that each cell comprehends the *entire surface* of the input layer . . . they seem to embody a variety of decidedly *holistic* features or dimensions of facehood, dimensions for which ordinary language has no adequate vocabulary" (47). But alas, though the intuitive propositional partitioning of facial features may fail us, there *are* some ways of finding what is going on within the neural network.

Learning Methods

An essential feature of ANNs is that they learn over time through experience. How they do this depends on which type of learning algorithm is used. There are a number of strategies which are used in training ANNs, and the various strengths and weaknesses of each (as well as their biological feasibility) are not central to my project. Still, a short example of how such learning procedures work will shed some light on what a neuro-computational approach to mental representations might highlight. It seems at least that understanding both how networks in the biological brain represent and how they change in response to experience would be of prime importance to forging a new dynamic view of the mind. Many ANNs learn by gradient descent learning, and Paul Churchland gives a good example of how this procedure works:

> Think of an abstract space of many dimensions, one for each weight in the network . . . plus one dimension for representing the overall error of the output vector on any given trial. Any point in that space represents a unique configuration of weights, plus the performance error that the configuration produces. What the learning rule does is steadily nudge that configuration away from erroneous positions and toward positions that are less erroneous. The system inches its way down an "error gradient" toward a global error minimum. Once there, it responds reliably to the relevant [inputs]. (*A Neurocomputational Perspective* 166–67)

There are reasons to believe that this sort of gradient descent learning is not what is going on in the brain's actual networks. For one thing, the comparison to a desired output would require that the biological brain already contains correct synaptic configuration which it used to train itself. In fact, there is a large degree of uncertainty as to "what features of the brain's microstructure are and are not functionally relevant. . . . Even so, it is plain that the models are *in*accurate in a variety of respects" (181). But concerning their relevance to the brain as a parallel computational system, "it is true that real nervous systems display, as their principal organizing features, layers or populations of neurons that project their axons *en masse* to some distinct layer or population of neurons, where each arriving axon divides into multiple branches [which] make synaptic connections of various weights onto many cells at the target location" (183). Yet a good model of the causal dynamics of cognition requires only that it models the functionally relevant features of these systems. Whatever is found to be the true coding strategy (or strategies) of real networks in the brain, there is reason to believe that a comprehension of learning will be a matter of understanding a process of changing various patterns in the system toward a functionally more effective representational "space."

Conclusion

Cognitive science should seek to understand cognition through the functioning of a computational system. Still, there are various ways of modeling computation. The symbolic paradigm inspired by traditional functionalism, where cognition is a series of operations over relationships between attitudes toward various symbolically structured propositions, is only one model of a computational system. However, there is reason to doubt that the intensional psychology of PAs will provide a satisfactory model for understanding the operations of human cognition. At most, this model will survive as a useful shorthand for the complex representational systems involved in neuro-computational cognition. As a way of highlighting relevant perceptual or informational features of temporally extended and superficial neural activation patterns, the language of PAs will serve an important social purpose. However, as cognitive science should give us an account of the causal features of neural computation, the language of folk psychology is too abstract to provide a satisfactory causal story. This is not to say that no symbolic computational system could instantiate cognitive processes (this is manifestly not the case), but merely that both PAs and folk psychology are unsatisfactory as computational theories; how such an architecture *achieved* cognitive tasks would still require a causal story at a lower level of description. I will allow that propositional attitudes serve the important function of helping us to see broad similarities between different computational systems. We can say that a robot is looking at a chessboard without commitment to any account of how this is achieved, and folk psychology allows us to notice that his human opponent is *also* engaged in a similar activity. As far as computational accounts of cognition are concerned, though, the ontology of PAs and folk psychology are terribly insufficient. Given this consideration, the explanation of cognitive tasks ought to be supported by a computational study of neurobiology. The result would be a computational account of cognition which would be continuous with research in neurology and would benefit from the lessons of functionalism (content is still causally determined) while taking the causal talk of folk psychology much less literally.

Works Cited

Churchland, Paul M. *A Neurocomputational Perspective: the Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press, 1989.

——. *The Engine of Reason, the Seat of the Soul: a Philosophical Journey Into the Brain*. Cambridge, MA: MIT Press, 1995.

——. "Eliminative Materialism and the Propositional Attitudes." *Philosophy of Mind: Classical and Contemporary Readings*. Ed. David J. Chalmers. New York: Oxford UP, 2002. 568–80.

Clark, Andy. *Associative Engines: Connectionism, Concepts, and Representational Change*. Cambridge, MA: MIT Press, 1993.

——. *Mindware: an Introduction to the Philosophy of Cognitive Science*. New York: Oxford UP, 2001.

Fodor, Jerry A. *The Language of Thought*. Cambridge, MA: Harvard UP, 1975.