Representations, Thought and Language:

Some Psychological and Neurological Considerations

In this paper, I will discuss several psychological and neurological issues related to the representation of concepts and propositions, which are basic elements in a general approach to psychological theory broadly called the representational theory of mind.

To begin with basic assumptions, I assume that "representation" is a construct in psychology. Representations are mental constructs; the reflection of a tree in a pond is not a representation of the tree and the thought of a tree by a human involves a representation of a tree. Mental constructs are part of psychology, a special science with its own, largely if not entirely proprietary, database, and a set of primitives and computational operations that provide descriptions and explanations of that database. There is considerable debate about these assumptions, with many scientists and philosophers maintaining, or suspecting, that constructs and operations described in psychology are replaceable by constructs and operations in a theory of neurobiology. I see no reason to believe this is the case now, any more than is the case for reductionism in other areas of science, or that it is likely to become the case in the near future.

Given that psychology is a special science, how is it related to neurobiology? The fact (if it is a fact) that psychology is a special science does not, of course, entail that there is no connection between constructs and operations specified in psychology and those specified in other sciences, just as there is, presumably, some relation between constructs and operations in other sciences (e.g., geology and physics). On the contrary, I think that constructs, operations and events specified in psychology correspond to constructs, operations and events in neurobiology. It does not follow that constructs, operations and events in psychology are caused by the corresponding constructs, operations and events in neurobiology; if there is a causal relation between psychological and neurobiological events, the direction of cause-and-effect is unclear. Does having a particular pattern of neural firing cause a person to activate a French word, or does activating a French word cause a person to have a particular pattern of neural firing? (In a computer analogy, a command in a program cause the electromagnetic events in a CPU to change, or do changes in the electromagnetic events in a CPU cause the command to have the effect that it does in the program?) As long as psychology and neurobiology have the status of independent special sciences, this is an open question. In addition, considering the question more broadly, the fact (if it is a fact) that constructs, operations and events in psychology are causally related to constructs, operations and events in neurobiology, the causal relation may not be necessary; it may be contingent upon these constructs and operations operating in humans. It may be that a silicon-based machine can support the same psychology that the neural carbon-based machine does, as Functionalism maintains.

All that said, even if any relation between psychology and neurology is causal, or its causal relation is contingent on the host of the relation being a human, even if and the direction of causality is unspecified, there is overwhelming evidence that human psychological processes take place in human neural tissue. This implies that these processes must be capable of being instantiated in human neurobiology. Accordingly, in principle, studying neural correlates of psychological theories might provide information about those theories. Whether this is true in practice is a separate question. At the moment, psychological and neurobiological descriptions are not interchangeable. If we want to characterize what a person knows when he knows French, or calculus, or the history of the Holy Roman Empire, we have to use terms such as "verb classes," "variables," and "battle," not terms such as "BOLD signal coherence," or "gene expression." The question I shall discuss here is how much neurobiology contributes to models of the nature of representations, at present.

Views regarding the value of neurobiology for psychological models of cognitive functions have changed dramatically in the past 30 years. The change is closely linked to changes in psychological theories. The type of psychological model that was most utilized in the 1960s and 70s assumed that cognitive primitives could be represented as symbols, and that symbols were related through various types of rules. I will call these "symbolic, procedural" models. These theories made use of computer analogies. Psychological predicates were

like software; the brain was like hardware; little thought was given to what types of software could run on neural hardware. These models largely ignored information about neural elements and systems. Chomsky (1994) famously claimed that "the brain sciences … as currently understood, do not provide any basis for what appear to be fairly well established conclusions about language." In contrast, many more recent models, first widely adopted in psychology with the 1986 PDP books (Rumelhart, McClelland and the PDP Group, 1986; McClelland, Rumelhart and the PDP Group, 1986), have properties that are often inspired – and sometimes justified -- by neural considerations. For reasons that will be clear later, I will call these models "probabilistic, associative" models. Advocates of these models often argue that they are preferred to symbolic, procedural models because they are neurally realistic. It is sometimes claimed – and often felt – that symbolic, procedural models are inconsistent with neural reality, which would, of course, entail that they are wrong. I will outline these two contrasting types of psychological models, and then discuss neural data that are relevant to each. [1]

Psychological Considerations

Symbolic procedural models: The Representational of Mind (RTM)

Symbolic-procedural models have been developed in many areas of cognitive psychology (e.g., almost all models of linguistic structure and many models of language comprehension and production are of this sort). The subject matter of my discussion will be representations of concepts and propositions, constructs that play central roles in psychological models of thought. I shall discuss these representations as they have been invoked in a broad framework that describes thought -- the Representational of Mind (RTM). RTM is developed in the literature in symbolic-procedural terms.

RTM holds that the cognitive psychological states consist of maintaining a propositional attitude – a statement of the form *X believes/wants/expects … P*, where *X* is an individual and *P* is a proposition. Holding propositional attitudes and relating them and their embedded propositions to one another constitutes cognitive activity and is the mechanism whereby people bring cognition to bear on action.

RTM is "a loose confederation of ideas Fodor (1998)," not a single model. Fodor (1998) outlines one version of RTM, in which the following three statements capture the essence of the theory.

1. Psychological explanation is nomic and intentional.
2. Mental representations are the primary bearers of intentional content
3. Thinking is computation

Thesis 1 states that psychological laws exist and are intentional.

Thesis 2 requires a propositional attitude to include a mental representation of a proposition, *P*, that is the content of the attitude. Propositions themselves express relations between concepts. The proposition that dogs bite includes the concepts *dog* and *bite* and a relation between them. Mental states thus consist of intentional relations to relations among the mental representations of concepts.

Thesis 3 states that thinking is "computation." In more detailed terms, thinking is the iteration of "causal relations among symbols that respect semantic properties of the relata (Fodor, 1998, p. 10)." Causal relations

[1] I will not discuss "neuropsychology" data – the role of disorders in individuals with known neurological disease to develop models of psychological phenomena and their neural correlates (See Shallice, 1988; Shallice and Cooper, 2011, for extensive presentations). Behaviors of individuals with known neurological disease are themselves psychological data, described and explained (to the extent that they are) in models whose terminology is drawn from psychology (including computational models). In this respect, they can reasonably be considered part of the data that the special science of psychology recognizes as its subject matter (although one might argue that they should be excluded, or treated as boundary cases of limited importance, because they arise in non-normal brains; this is a complex issue). The characterization of the neural state of these individuals (e.g., features of their lesions) is part of the data that contribute to our understanding of the neural correlates of these behaviors.

among symbols that respect their semantic properties occur because, at its basis, the mind/brain operates as a Turing machine, in which the next operation of a machine is based on the state the machine is currently in and the symbol it is reading. The computations of RTM are purely formal, determined by the representations that constitute their input, not the content of those representations. This concept of computation is broad. Though it limits computations to causal relations that respect semantic properties, it does not limit them to causal relations that are truth preserving. Thus association is a computation.

A simple example of the way the process works is Jones walking down the street and seeing a man with a dog. Jones knows that dogs bite (i.e., holds a propositional attitude) and crosses the street to avoid them (bases action on an entailment of the propositional attitude he holds). Jones may also think of dog food (i.e, have an association) and remember he has to buy dinner (i.e, have an further association that might lead to action). I see these two types of operations – entailment and association – as playing two different roles in human psychology. Entailment and inference are the basis for rational thought, justifiable belief (including science) and many successful actions (including applications of science). Association contributes to creativity.

The example above is extremely simple. Much more elaborate chains of mental states underlie more complex decision making and action. In many cases, thought involves considering multiple possibilities (i.e., holding multiple propositional attitudes) and computing their entailments; and social interactions involve attributing similar complex chains of thought to others.

The RTM framework requires characterization of the concepts that are related in the propositions that are the contents of propositional attitudes and of the rules that relate these concepts to form these propositions (among other requirements). I will discuss these topics in turn.

Concepts

The concept of a concept is, to say the least, illusive (see, for instance ((1975, chapters 8 and 12) and Margolis and Laurence (1999, Chapter 1) for summaries and critiques of many commonly held views about what they might be). As a point of departure, I will defer to my thesis advisor, Jerry Fodor. Fodor (2000) advances three premises/assertions about concepts.

1. Concepts apply to things in the world. The concept *dog* is one which, of necessity, all and only dogs fall under.
2. Concepts are word meanings. The concept *dog* is what the word 'dog' and its synonyms and translations express.
3. Concepts are constituents of thoughts. To think that dogs bark is *inter alia* to entertain the concept *dog* and the concept *bark*.

Premise/assertion 1 deals with the content of a concept. Though, as noted above, RTM holds that the content of an expression does not determine its role in mental operations, the content of an expression does determine the truth of a proposition that contains the concept, which is of course critical to the utility of propositions.

Premise/assertion 1 claims that a concept necessarily applies to all and only the items that fall within its extension. What does this imply about the representation of concepts?

The classic, Aristotelian, answer is that the representation of a concept is the set of properties that are individually necessary and jointly sufficient to characterize items in its extension. In this model, concepts are related to features by the relations of dominance and sisterhood. Thus, the concept MODE OF TRANSPORT dominates VEHICLE, BOAT, PLANE, etc.; VEHICLE dominates CAR, TRUCK, SUV, etc. At the bottom of the hierarchy are primitive features. Concepts inherit features from the nodes that dominate them and transmit them to the nodes below.[2] Despite many positive features,[3] and its venerable history, the Aristotelean view of the

[2] This view is closely related to the view that words have definitions.

content of concepts has been firmly rejected by cognitive psychology and cognitive neuroscience.[4] A number of models of content have arisen in place of the Aristotelean view of the content of concepts. They fall into two broad categories - inferential role semantics models and information models. Theory theories exemplify the first type; prototypes exemplify the second. Prototype models are important to the discussion here because they are universally accepted by "associative-probabilistic" models; I shall discuss them below.

Premise/assertion 2 states that the concept *dog* is what the word 'dog' and its synonyms and translations express. This relates concepts to language (words). It points to the important fact that there is more to a concept than its content, understood as the items in the world that it designates. This is a familiar conclusion, associated with Frege's (1892) distinction between reference and sense. As Frege pointed out, "the evening star" and "the morning star" refer to the same object but are not the same concept. The reason is related to the existence of intentional contexts (a critical feature of RTM). Different terms cannot substitute freely in intentional contexts and preserve truth. It is possible for John to believe that the evening star is the evening star and not believe that the evening star is the morning star. Therefore different co-extensive terms must be different concepts.

Frege proposed that words had both sense and reference. He suggested that the sense of an expression is the way by which one conceives of the denotation of the term. He proposed that when a term (a name or a description) follows a propositional attitude verb, it denotes its sense, not its referent. In Frege's system, while "John believed that the evening star is the evening star" and "John believed that the evening star is the morning star" have the same truth value (if the morning star and the evening star are coreferential), they are different propositions and ascribe different thoughts to John, because the senses of "the evening star" and "the morning star" differ.

As Fodor points out, however, sense is not quite fine grained enough to deal with all intentional contexts. In Frege's view, terms with the same sense are synonymous. However, synonyms cannot substitute *salve veritate* in so-called Mates contexts. "Pupil" and "student" are (arguably) synonyms, but it is possible for Bill to wonder whether John understands that pupils are students even though Bill does not wonder whether John understands that pupils are pupils. Fodor (1998) argues that what is added to a concept aside from its reference is its "mode of presentation (MOP)." There seem to be as many MOPs as there are words and phrases. Fodor (1998, p. 17) says "If it is stipulated that MOPs are whatever substitution *salve veritate* turns on, then MOPs have to be sliced a good deal thinner than [Fregean] senses. Individuating MOPs is more like individuating forms of words than it is like individuating meanings."

MOPs have roles in the operations of RTM discussed above. The same MOP can play the same mental role in two people even if the concepts it refers to are non-coextensive (the most notorious example is imaginary: "water" on Hillary Putnam's Twin Earth and water on Earth). The role of a MOP rather than a referent in mental processes can be illustrated by considering a person who thinks "Jill's new boyfriend" is Bill and one who thinks

---

[3] One strength of this view is that simple concepts and complex concepts have the same structure – the former are constituted by sets of features, the latter by sets of primitive concepts. Because a concept entails and is entailed by its constituents, the necessity of the application of a concept to items that fall under its extension (premise/assertion 1) is guaranteed. A corollary is that this model accounts for analyticity; it is because "bachelor" entails and is entailed by "unmarried man" that "bachelors are unmarried men" is analytic.

[4] A widely cited problem, attributed to Wittgenstein (1953), is that no examples of severally necessary and jointly sufficient features have ever been provided. A second is that there is no principled distinction between features and concepts (correspondingly for definitions, between *defieniens* and *definiendum*); the features that are postulated often appear to be at least as complex as the concepts whose constituents they are said to be (e.g., Jackendoff (1972) defines "keep" as "cause a state to endure"). A third, related to the second, is that there does not seem to be a way to identify the set of primitive concepts or features. A fourth is that the hypothesized definitional features of a concept never seem to be activated in psychological processes. A fifth is that the semantic phenomena that the Aristotelian account explains probably don't exist. The analytic/synthetic distinction (see note 3), for instance, is widely regarded as having been seriously discredited, if not destroyed, by Quine (1951).

it is Henry. They can discuss many topics about Jill and her boyfriend without realizing they are not referring to the same person – the phrase "Jill's new boyfriend" is mentally active, not its referent.

To summarize, Premises 1 and 2 lead to the view that "a concept is a MOP together with a content (Fodor, 1998, p. 17, fn 16)." The content of a concept is a nomic relation between the concept and some set of items or events in the world (the notion of item or event is broad; *honesty* is a concept). MOPs are mental objects that have semantic content and that play causal roles in the rules that combine representations in the propositions in propositional attitudes. MOPs capture both what Frege's theory of sense and Turing's notion of computation require, a fact that Fodor (1998) says is one of the most important convergences in psychology. Basic concepts, and unanalyzable MOPs, are not constellations of individually necessary and jointly sufficient features. I will argue below that they are also not constellations of probabilistically associated features. Fodor (1998, and elsewhere) says that the content of a concept is whatever the mind "locks" or "resonates" to when presented with prototypical instances of the concept -- the content of the concept DOG is "dogginess." Underspecified as this view is, and unhelpful as it seems to most people I know, I shall leave this section with this idea, a key feature of which is that basic concepts are unitary.

Computation

Fodor's (2000) assertion/premise 3 is that "concepts are constituents of thoughts. To think that dogs bark is *inter alia* to entertain the concept *dog* and the concept *bark*." To think that dogs bark also requires that one be able to combine the concepts *dog* and *bark* to form the proposition (*dogs bark*). This is one aspect of computation.

Computations have three properties that Fodor describes as compositionality, systematicity and productivity. These features can be illustrated with an example. If you understand the words "the," "boy," "girl" and "push" and you are capable of having the thought "The boy pushed the girl," you must be capable of having the thought "The girl pushed the boy." Note that this capacity does not depend on preservation of truth; asymmetry of action is commonplace. Nor does it depend upon epistemic factors; even if it is true that "The girl pushed the boy," you may never have evidence that this is the case. It is a property of the human mind that, if it is possible for a person to represent "The boy pushed the girl," it is possible for him/her to represent "The girl pushed the boy."

Compositionality, systematicity and productivity immediately follow from the view that a person who has the thought "The boy pushed the girl" has a rule that represents a relation between representations of "the," "boy," "girl" and "push," which allows "the," "boy," "girl" and "push" to combine in the way that expresses the propositions "The boy pushed the girl" and "The girl pushed the boy." Because systematicity and productivity are general properties – the relation between the ability to have the thoughts "The boy pushed the girl" and "The girl pushed the boy" applies to an infinite number of concepts that can substitute for "the," "boy," "girl" and "push" -- the rule establishes a relation not between representations of specific concepts but between higher level characterizations of concepts. The higher level characterizations of concepts that figure in the rules are not simply superordinate levels of a concept hierarchy; a single rule applies to items in many semantic domains. The rules could take many forms. They might be formulae in predicate calculus; i.e., a rule might state that objects are variables (x,y) in expressions of the form $f(x,y)$, where f is an action.

Linguistics provides candidates for the necessary higher level characterizations of concepts in the form of syntactic categories, which are independent of semantic class to the needed extent, and for the rules that relate concepts, in the form of the syntactic structure of a language. The linguistic approach to compositionality, systematicity and productivity is to postulate lexical items that are assigned syntactic categories, informally known as "parts of speech" – noun, verb, determiner, etc. (for issues with this claim, see Croft, 1990) -- and syntactic structure in which the terminal symbols dominate lexical entries and in which dominance and sisterhood relations defined over nodes in these trees determine semantic relations among the lexical items. The system is highly specific -- similar relations among nodes differ in whether they allow semantic relations to be established between the lexical items they dominate (Figs 1, 2).

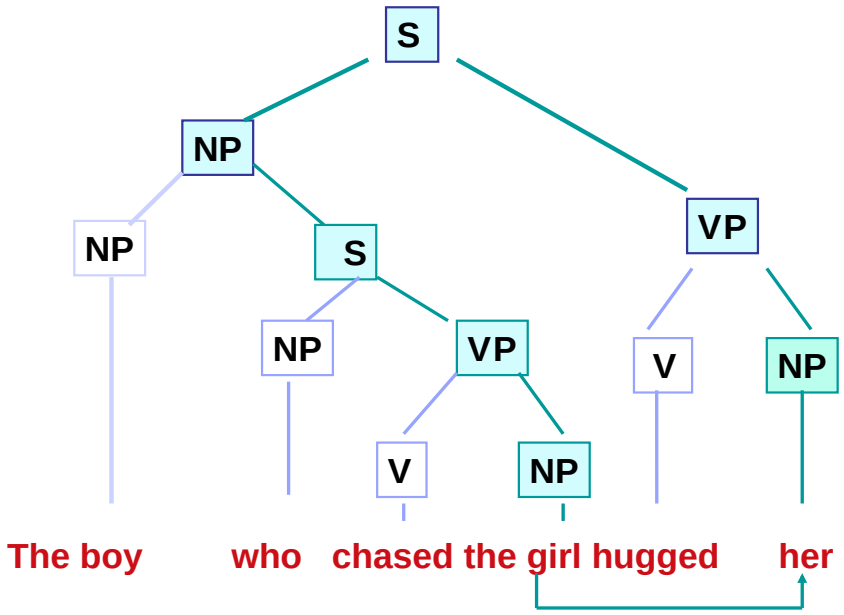

Figure 1: Interpreted path over phrase marker.

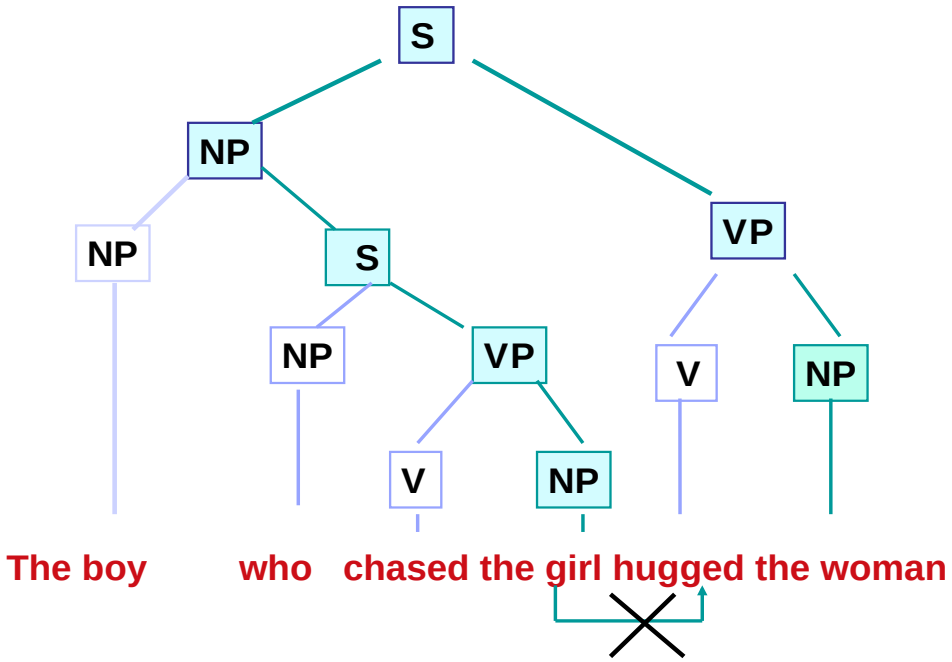

Figure 2: Uninterpreted path over phrase marker.

Syntactic structures interface with expression and perception. They therefore map meanings onto sequences of words. The mapping is many-to-many. The same meaning can be mapped onto different word sequences (e.g., thematic roles in actives and passives); a given sequence of words can be ambiguous; and word sequences

(and other surface forms) encode several aspects of meaning (e.g., a sentence-initial NP can be the subject of a sentence, the agent of a verb phrase, and the focus of a discourse). One way to characterize the surface form of syntax and morphophonology is as a compression of multiple aspects of meaning into a linear signal. A major issue in theoretical linguistics is how syntactic representations capture this compression. All options involve an extremely complex set of rules that relate concepts in propositions to surface forms of phrases and sentences.

Syntactic structures (and morphology) provide rules that can account for some of the combinatorial properties of concepts. The question of whether these representations support thought has largely been answered negatively (e.g., Fodor, 2008; Pinker, 1994). The most important objection seems to be that mental content is ultimately unambiguous (the meanings of ambiguous sentences must themselves be unambiguous). Therefore, it is argued, we must think in a "language of thought (LOT)," not a natural language. However, the unambiguous nature of mental representations is only incompatible with an un-annotated surface structure of a natural language being the entirety of the representation of a sentence. A conceptual (semantic) level of language – Logical Form (LF) – would be a possible LOT. A sufficiently annotated surface structure can also be unambiguous and could be the form of thought. A viable position, I believe, is that syntactic (including morphological) operations underlie the combination of word meanings at the propositional level.

Although it is possible that some of the properties of LOT are due to direct utilization of features of natural languages, others clearly cannot be. For instance, associations based on words and inferences based on propositional meaning are not derived by combination of concepts based on syntactic rules. Because of this, many aspects of meaning are not created by the application of rules of language to lexical items. As I mentioned above, an important set of these preserve truth. These include rules of valid inference. I would like to emphasize that such rules have the properties of compositionality, systematicity and productivity that characterize the rules that combine concepts to form propositions. Syllogisms relate their elements systematically and productively, and any human who can draw a valid conclusion from the syllogism

All plumbers are electricians

All electricians are carpenters

can draw one from the syllogism

All electricians are plumbers

All plumbers are carpenters[5]

To summarize, RTM holds that psychological laws are intentional and that mental states consist of propositional attitudes in which the propositions relate concepts. Concepts connect nomically to items in the world, though how they do so remains mysterious. The content of a concept is not the aspect of a concept that plays a role in mental computational processes. What plays such a role is how a person "grasps" the concept. There seem to be at least as many ways of grasping concepts as there are lexical items and their combinations – the MOPs of concepts. Words are the best available candidates for the elementary MOPs that play such causal roles. The syntactic operations of language allow the combination of concepts into phrases and into the propositions in propositional attitudes. These operations, like those of valid logic and mathematics, consist of rules that operate on symbols.

[5] Because of this, syllogisms also can express counterfactuals, and cannot be learned by inductive generalization in which the truth of a conclusion is associated with the truth of the premises. Consider:

Some plumbers are electricians

Some electricians are carpenters

The fact that the inference that some plumbers are carpenters is not valid can be endorsed by person who has only encountered plumbers who are carpenters.

Probabilistic, associative models

There are many contentious aspects of the picture I have presented. One area of contention deals with "internal" issues. Debates of this type accept the idea that mental operations include symbolic representations of concepts and combinatorial rules but dispute aspects of particular models. For instance, many models of syntactic structure are based on philosophical and linguistic considerations and make little contact with experimental studies of perceptual identification, comprehension and production of words and sentences. Alternative models of syntactic structure have arisen within these disciplines -- e.g., Lewis and Vasishth's (2005) implemented parser use of modified X' categories -- that do not appear in theoretical linguistics, as far as I know.

The second set of objections is much more fundamental. It rejects the notions of unitary symbols and combinatorial rules; it sees mental activities as probabilistic at their core. Spivey's (2007) description of "representation" makes some of the differences in the approaches clear:

> [the] term representation . . . need not refer to an internal mental entity that symbolizes some external object or event. . . The word can. . . refer to a kind of mediating stand-in . . . between sensory stimulation and physical action, which is implemented largely by neuronal assemblies . . . [that] never settle into truly stable states . . . the bottleneck that converts fuzzy, grey, probabilistic mental activity into discrete easily labeled units is *not* the transition from perception to cognition – *contra* cognitive psychology. Rather that conversion does not take place until the transition from motor planning to motor execution (p 3 – 6)"

There are many models fall into the "probabilistic, associative" category. The best known and most widely used in cognitive psychology are connectionist (PDP) models and dynamic state and phase models. Developers of probabilistic models have claimed that these models are superior to symbol-based procedural models in describing and explaining a wide range of psychological phenomena.

I find it instructive to consider the areas in which success has been claimed. A complete review of work using these approaches was beyond me; I compiled a list of representative areas of study from three relatively recent sources: Spivey's (2007) book on dynamic models; Joanisse and McClelland's (2015) review of studies of language; Rogers and McCelland's (2014) review introduction to a special issue of *Cognitive Science*. By my count, Spivey reviews 9 simulations, Joanisse and McClelland 7, and Rogers and McClelland 25. Almost all of the work reviewed deals with single items. A small portion (5 studies) deals with structural relations over time; a portion of this latter work is presented as showing the capacity of probabilistic/associative models to extract "syntactic-like" relations among words in sentences.

The work dealing with single items has been directed towards two related phenomena: extracting patterns in what are called "quasi-regular" domains, such as spelling-sound correspondences and past tense formation in English, and modeling interactions between factors that determine categorization, interpretation, aspects of memory, naming and other behaviors. These models have greatly expanded our understanding of possible ways that structure may be extracted from stimuli in a domain and that factors might interact. They effectively model "graceful degradation" of performances in these domains following brain damage. They have done much to dispel the encapsulation view of processing associated with Fodor's theory of modularity. That said, many of the best developed models of processing single items are highly controversial (see Shallice and Cooper, 2011, for a review of issues that have arisen in models of word reading, an area in which PDP models have been quite successful).

There is much less work dealing with extracting structure from sequences of items, and very little pertaining to extracting structure in domains that, from the viewpoint of symbolic models, involve operations that combine categories in the way that the computations reviewed above do. I will review work dealing with the aspects of combinational computation discussed above.

As discussed above, RTM maintains that one set of mental states consists of maintaining a propositional attitude – an intentional relation to a relation of concepts that forms a proposition. This requires a theory of concepts and their combination into propositions. Beginning with the former, to my knowledge, though details vary from model to model, all probabilistic/associative models of concepts adopt the view that concepts are prototypes.

Prototype theories share features with the Aristotelian view. In both, concepts are related by the relations of dominance and sisterhood and concepts inherit features from the nodes that dominate them and transmit them to the nodes below. The difference between Aristotelian and prototype models is that, in the Aristotelian model, the relation between nodes is one of necessity and, in prototype theory, it is probabilistic. Properties are assigned to items probabilistically in two ways. First, whereas an Aristotelian view of the concept DOG might maintain that dogs necessarily have tails, bark and are man's best friend, a prototype model might maintain that they are likely to have these properties but need not have any of them. Second, properties may be graded -- it is possible to rate man's affection for an animal. Items that have higher values of properties associated with a concept and those that have more of these properties are more likely to be instances of the concept than items that do not. The prototypical item at each level of the hierarchy is the item that shares the most features with other items at that level.

Unlike definitions, the effects of prototypicality are manifest in virtually every psychological domain. People produce words corresponding to prototypical concepts faster than to non-prototypes. Children learn names for prototypical members of a category before non-prototypical items. Aphasic patients generalize semantic feature training for naming from non-prototypical to prototypical members of a category more than *vice versa* (a superficially paradoxical result that some computational models of recovery simulate).

Despite these successes, prototypes cannot be the content of concepts for many reasons. To begin with, they confront the same problems in specifying features that Aristotelian models face. In addition, they encounter problems specific to themselves.[6]

One is that notion of similarity upon which prototypes build presupposes the notion of identity. In order to grade how similar two items are with respect a feature, we must know what the feature consists of. A second is that prototype theory cannot account for "Boolean concepts," such as "not a cat". As Fodor (1998) points out, a bagel is not a cat, but items that are more like bagels are not better instances of the concept "not a cat." A third problem for prototypes is that they are found in domains where definitions do exist. 3 is the prototypical odd number and the prototypical prime number. Maintaining that prototypes are the content of concepts in these domains is clearly wrong and leads to the inverse of the Boolean concepts problem: 5 is not a better odd number or prime number than 29 for being more like 3 than 29. Most important, prototypes do not combine to form complex concepts. The prototypical fish for most Americans is a medium sized fish -- perhaps a trout, perch, salmon, etc.; the prototypical pet is a dog or a cat; but the prototypical pet fish is a goldfish or another small fish that survives in a glass bowl, not the combination of a trout and a dog.

To my knowledge, no probabilistic models of complex concept formation solve this last problem. Zadeh's (1997) fuzzy logic does not (see discussion of earlier work by Zadeh (1965) in Osherson and Smith (1981)). Smolensky (1990) developed a tensor product model of complex concept formation in which vectors representing elementary concepts combined to form new vectors representing complex concepts. The model was productive and systematic in the sense that it allowed any number of vectors to be combined. However, it was not compositional. The contributing vectors could not be extracted from the resulting vector; the vector representing the complex concept was, in essence, a unitary item.

[6] Jean-Michel Fortis pointed out to me that Rosch herself (e.g., Rosch, 1978) treated prototypes as fictions, claiming only that there are judgments of typicality.

With respect to extracting structure from sequences of items, specifically the combinatorial operations that determine propositional meaning, the reviews I cited above showcase work by Elman and Socher. Elman (1991) used recurrent nets to extract syntactic structure. However, Elman's model dealt with a very limited range of structures. Christiansen and McDonald (2002), which is rarely cited, reported training of a simple recurrent network to predict the next word in sentences. The model had a number of successful predictions (e.g., more errors predicting verbs of object relative sentences than other words) but it made many false and ungrammatical predictions. The most recent work on the topic, cited by Joanisse and McClelland, is by Socher and colleagues, who developed a parsing model using multiple hidden layers that mediate the input from the output (so-called "deep learning"). Socher et al (2013) used this recursive neural network to model the Stanford Sentiment Treebank. The model is important because it not only assigns structure but also interprets phrases, as positive or negative in emotional valence. The model is successful at assigning words the proper valence, but creates incorrect syntactic structure, illustrated in the following figures.

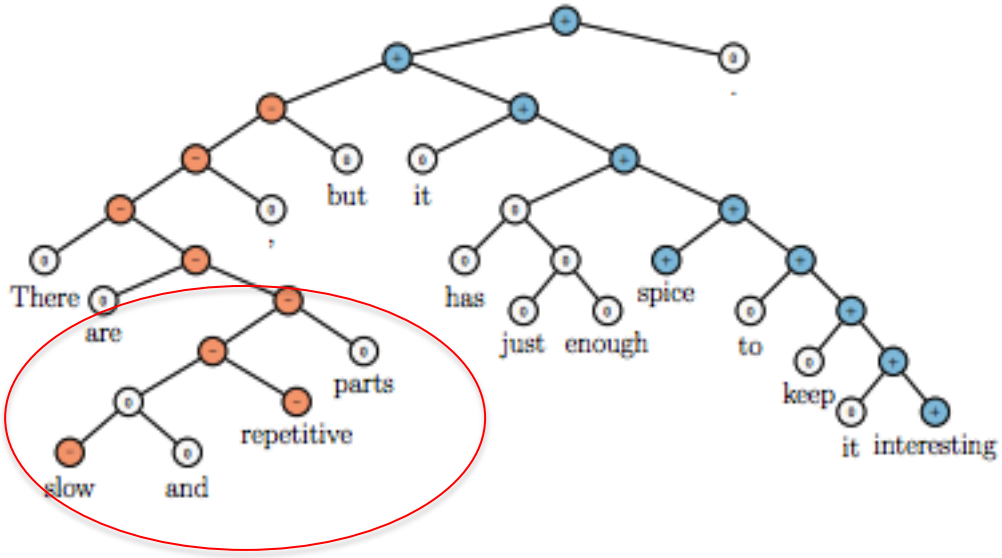

Figure 3: Incorrect parse (1) in Socher et al. (2013).

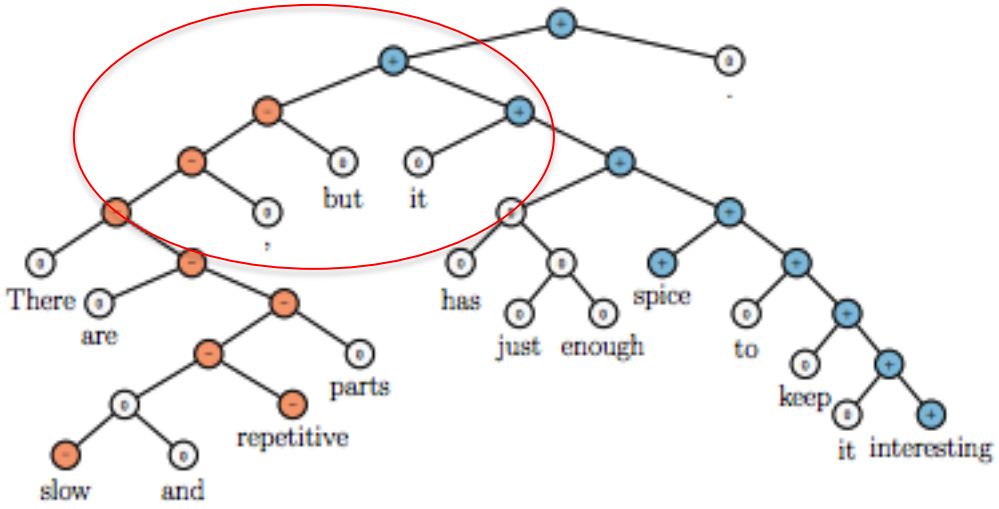

Figure 4: Incorrect parse (2) in Socher et al. (2013).

It is, of course, premature and dangerous to predict the limits of a vigorous line of research, but, at present, it seems to me that probabilistic/associative models have failed to relate concepts based on their forms in psychologically realistic ways. One possible direction of research using these models that may fill this gap are dynamic models (Spivey, 2007). These models describe sequences of activation of categories (actually, approximations to categories – see quote from Spivey above) as transitions between attractors in phase or state space. Shallice and Cooper (2011) suggest that this is the basis for combinatorial operations over categories. Whether these models will succeed where existing ones have not remains to be seen.

Dynamic models also provide a means for a potentially interesting combination of aspects of symbolic/procedural models and probabilistic/associative models. They allow for apriori values of transitions between attractors in phase and state space (Smolensky and Goldrick, 2014). These apriori values might be a basis for rules such as those that create syntactic structure, underlie valid inference, and others. Note, however, that, if this were to be developed, the success of these models in describing and explaining combinatorial operations over categories would not be entirely due to the probabilistic/associative nature of the computations in individual nets but also to pre-specified, non- probabilistic/associative constraints on relations among attractors. These constraints would correspond to the rules of symbol-based procedural models.

A final consideration regarding the rules that govern relations among categories and categories themselves is that many are not features of the world but products of the human mind. Imaginary numbers do not exist outside the human mind; the same seems to me to be true for syllogisms, syntactic relations among grammatical categories and categories other than natural kinds. If our science is right, gold and tigers exist in the real world, and the concepts “gold” and “tiger” could be learned by observation of items in the non-human world (in fact, the concepts “gold” and “tiger” depend upon scientific observations), but shrubs are not a biological entity – shrub is a category whose origin is totally human. The fact that the origin of many combinatorial relations among categories and of many categories themselves is the human mind entails that what a human has to learn is how other humans categorize the world and relate the categories they assign. The available database that probabilistic/associative models consult to extract these aspects of cognition is the overt behavior of humans. This is an extremely limited source of information compared to the information available to humans (people

learn about imaginary numbers, or the multiplication table, or syllogistic logic by exposition from teachers, not observation of behavior) but, even in domains where it could conceivably be sufficient (e.g., acquiring word meaning), the question of how the mind created these relations among categories and the categories themselves needs to be explained. As best I can see, this leads back to apriori, innately human, constraints on these aspects of cognition, not obviously phraseable in probabilistic/associative terms.

To summarize, looking at probabilistic/associative models from the perspective of the view of thinking and cognition broadly set out by RTM, these models have had considerable success in modeling categorization, interpretation, naming and other behaviors affecting the processing of individual items, though, even there, their success is qualified by problems associated with basic features of their models, such as the adoption of prototype theories of concepts. These models have fared poorly in describing, and therefore have necessarily failed to explain, phenomena related to combinatorial operations over categories. As best I can see, this failure is due to intrinsic features of these models, which are not capable of extracting "rules" that determine relations among categories that appear very infrequently in the dataset.

Neurological Considerations

My reading of the literature and discussions with researchers leads me to believe that neural considerations – the belief that probabilistic/associative models are neurologically realistic and symbol-based procedural models are not -- are a large part of the motivation to model combinatorial operations probabilistically. As I said at the beginning of this talk, it is often implied, and sometimes stated, that symbol-based procedural models of the sort I have outlined above in connection with RTM are not neurally realistic and therefore not to be taken seriously. In not a few papers, researchers appeal to the neurological realism of probabilistic models in adjudicating between theories of different types. Thus, for instance, Seidenberg and Plaut (2006) acknowledged that models that postulate word nodes often provide better accounts of empirical findings in word naming than models that have distributed representations but endorsed a model of the latter type because it was considered more neurologically plausible. In this part of this talk, I will consider neurological issues.

Rogers and McClelland (2014) list seven "central tenets" of connectionist models, four of which are widely thought to be related to – even derived from -- neurological elements and events:

1. Cognitive processes arise from the real-time propagation of activation via weighted connections
2. Active representations are patterns of activation distributed over ensembles of units
3. Knowledge is encoded in connection weights
4. Learning and long-term memory depend on changes to connection weights

In connectionist models, these four properties are instantiated by units at one level being connected to multiple units at other levels and activation of a unit changing the strength of the connection between it and those it connects to (1). Information (knowledge) is distributed in the sense that it resides in multiple connection weights (3). The pattern of activity over multiple units at one (or sometimes more) level(s) corresponds to a piece of information (2); the activity of individual units is not sufficient to individuate a piece of information. Learning is achieved by changes in connection weights through feedback from the difference between activation patterns in an output layer and the expected pattern (4), where expectations can be provided externally or internally by predictions about upcoming events.

There have been many discussions about the neurological reality of models with these properties. I will focus on the neurological elements that correspond to the units themselves. For these models to be neurally realistic, assuming the properties of the units are those listed in Rogers and McClelland (2014), there must be elements in the nervous system that have the set of properties specified in 1 - 4. Elements that have only one or two of these properties are not suited to fulfill the computational functions found in the models Rogers and McClelland (2014) refer to.

The obvious hypothesis is that the relevant neural elements are individual neurons, and depictions of units in these models often reinforce this view (Figure 5). Many individual neurons have efferent and afferent connections (synapses) onto and from many other neurons (consistent with 1, 2). Axonal activity releases neurotransmitters that alter post-synaptic membrane voltage (consistent with 2) and induce changes in post-synaptic neurons that alter its responses to further stimulation (consistent with 3, 4). Firing rates of neurons are related to discernable properties of input at a variety of levels of abstraction (from oriented lines through individual faces and people) and to responses (2). Nonetheless, the units in probabilistic/associative models are not likely to correspond to individual neurons.

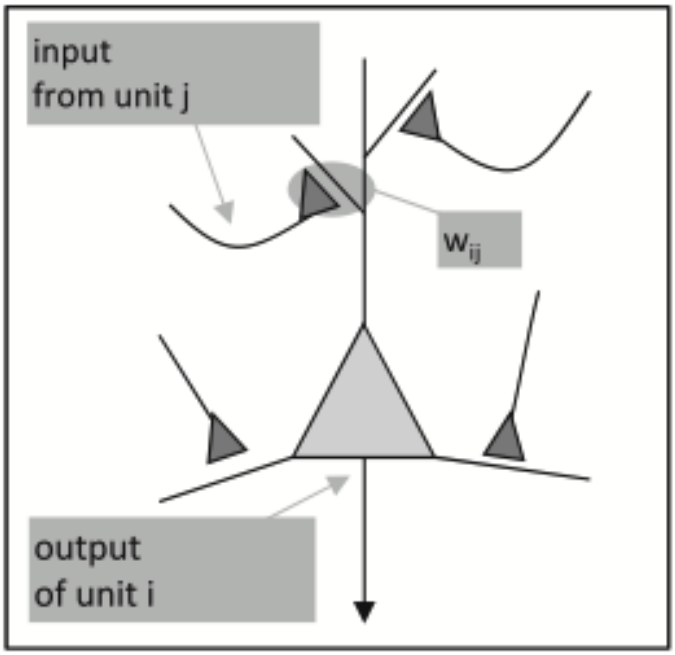

Figure 5: Depiction of unit in a PDP model in Rogers and McClelland (2014).

One problem that has been frequently cited in equating units in probabilistic/associative models with individual neurons pertains to the changes in connection weights that constitute learning (4). In many probabilistic/associative systems (e.g., Plaut et al, 1996, highlighted in Rogers and McClelland, 2014), the magnitude of the error between the output of the system and an expectation feeds back to a hidden unit layer and alters connection weights. To my knowledge, single cells that are sensitive to difference functions (e.g., dopaminergic cells in VTA; Cohen et al, 2012) do not have unisynaptic reciprocal connections to their afferents. Feedback based on error detection is determined by a complex set of responses of different cells. The implication is that the units in probabilistic/associative models that contribute to property 4 – at least to the extent that changes in connection weights that constitute learning depend on feedback -- must consist of sets of cells, not individual neurons.

A second problem with neurons being the units in probabilistic/associative models pertains to the nature of connection weights. Corresponding to the view that probabilistic/associative units are neurons, connection weights between units are often thought of as synapses. An example comes from Shallice and Cooper (2011, p, 214), who seamlessly move between mathematical and neural levels in describing a connectionist model of semantic processing that "involves neurophysiological assumptions about the operations of synapses within the context of an otherwise standard feedforward connectionist model." Equating connection weights with synapses is an enormous unsustainable oversimplification. There are

multiple types of synapses in virtually all CNS cells, with different structures responsible for different post-synaptic activity. Post-synaptic cells change in multiple ways in response to stimulation, ranging from simple increases or decreases in local membrane potential through post-stimulation tetanic responses and development of new receptors as a result of genetic expression (Kandell, 2009). See Figure 6. One feature of this complex system is that long-term changes in the responsiveness of post-synaptic neurons are due to many mechanisms – gene expression, new dendritic spine formation, etc. Because long-term changes in the responsiveness of post-synaptic neurons involve post-synaptic machinery, synapses do not have property 4 of connection weights – the changes in connection weights that underlie learning and long-term memory involve much more than synapses.

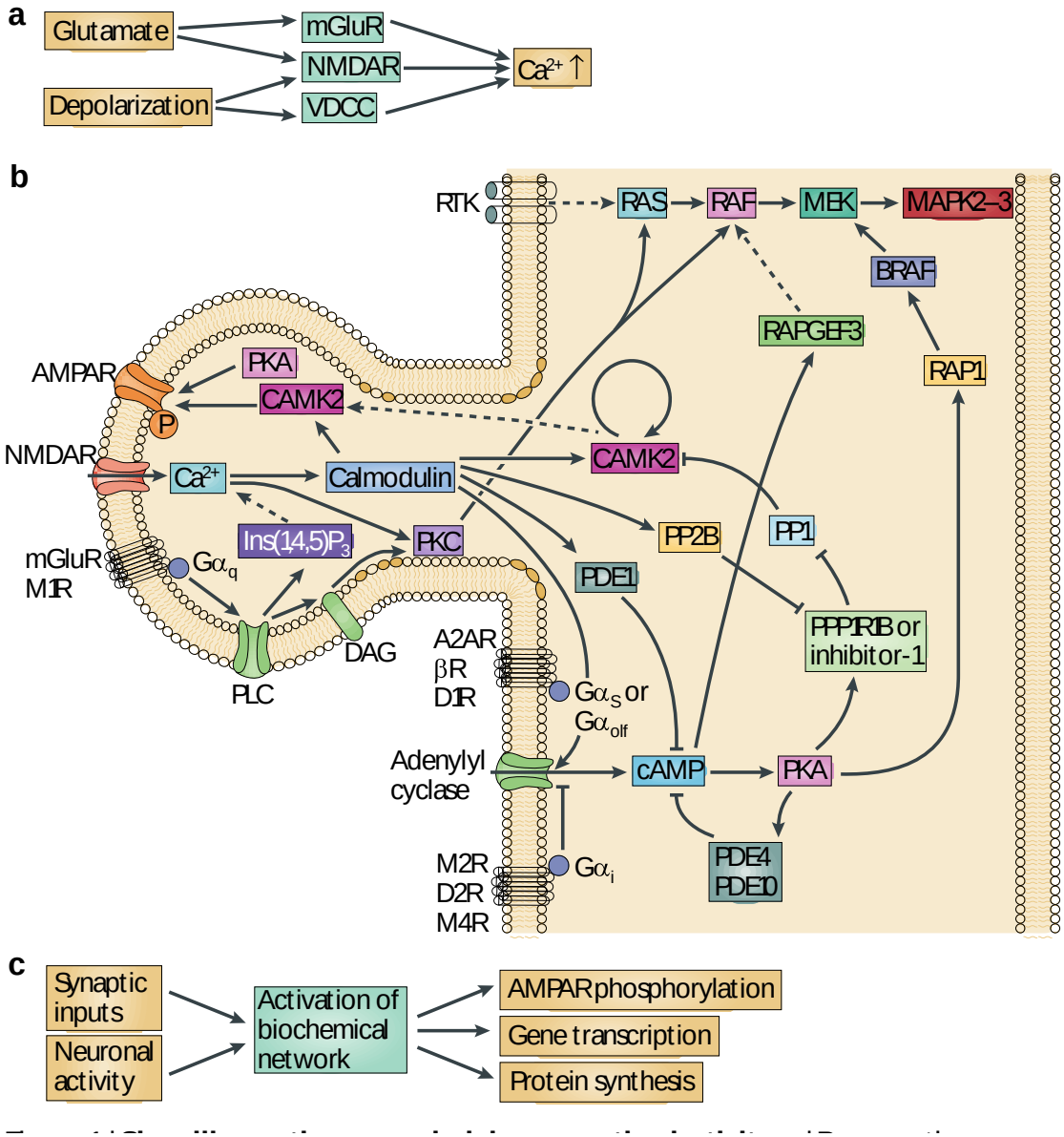

Figure 6: Cartoon of some cellular and molecular events underlying plasticity (Kotaleski and Blackwell, 2010).

There is another problem in equating connection weights with synapses. Connection strength and unit activity are linked in models in ways that do not appear to correspond to properties of synapses. Property 3 – "knowledge is encoded in connection weights" -- is related to property 2 –"active representations are patterns of activation distributed over ensembles of units" -- through property 4 -- learning and long-term memory depend on changes to connection weights. As connection weights change in association with input to a unit, the new connection weight contributes to the representation of knowledge and to activation of that unit by successive input changes. However, these two effects of input are decoupled at synapses. Input to ionotropic receptors leads to fast depolarization or hyperpolarization of the post-synaptic neural membrane, which affects the development of an action potential (the neural activity that corresponds to an "active representation" (property 2)). However, these events do not lead to changes in the responsiveness of the post-synaptic receptor (property 4); i.e., they do not change the post-synaptic receptor in a way that corresponds to the change in connection weight that results from input to a unit in a probabilistic/associative model. The post-synaptic changes that are responsible for long-term changes in the responsiveness of post-synaptic neurons involve other cellular elements, partially listed above.

In addition to being structurally decoupled, properties 2 and 4 occur on different time scales at synapses than the corresponding events in probabilistic/associative models. In models, an input to a unit leads both to that unit becoming active and to a change in its connection weight; however, in neural tissue, membrane potential changes occur on a much faster time scale than the changes that affect a neuron's responsiveness. Postulating "fast" and "slow" connection weights does not solve this problem, as far as I can see, because fast weights both affect immediate activity of the downstream unit and create long term connection weight changes that carry representations. It is not clear that the ensemble of activity in all these cellular elements have properties 2 – 4.

Because of problems like these, advocates of probabilistic/associative models generally do not claim that units are individual neurons. Plaut and McClelland (2010a) say "The computational principles that underlie the probabilistic/associative approach are intended to capture how brain areas learn to represent and process information as patterns of activity over large groups of neurons rather than the detailed operation of the individual neurons themselves. p 286-7." Proposals span a range of sizes, from groups of cells that are very small by neural standards (e.g., Georgopoulos and Massey, 1987) to much larger groups. Cox, Seidenberg, Rogers (2015, 381-2) suggested that they may be as large as a voxel in a BOLD fMRI study: "we take the activation of a single unit to be a model analogue of the mean activity in a population of neighboring neurons, similar to that estimated from changes in the BOLD response at a single voxel using fMRI."

These proposals are vague and arbitrary. Though Cox et al pursued their suggestion with several analyses of BOLD signal data that that say support probabilistic/associative models, it seems hard to believe that they are seriously claiming that units are the neurons in a voxel in a BOLD signal study, if for no other reason than that voxel size in functional neuroimaging studies has grown substantially smaller as technology has advanced, but brains have not changed fundamentally.

The fundamental problem with all of the proposals about larger numbers of neurons being the units in probabilistic/associative models is that, although larger groupings of neurons have some of the properties of units in probabilistic/associative models, none of the sets of neurons that have been proposed as the units in these models are known to have all of the properties attributed to units in probabilistic/associative models. For instance, Georgopoulos has shown that information about direction of movement was coded as a population vector over a set of 241 neurons (property 2), but there is no evidence that groups of neurons of about that size have features that correspond to weight changes in probabilistic/associative models (property 3) or have reciprocal connections to neurons that encode error signals (related to property 4).

I have directed my comments towards the lack of correspondence between units in probabilistic/associative models and elements in the brain, but the problem arises the other way around too: there are informationally-relevant physiological events that are not clearly modeled in probabilistic/associative approaches. For instance, theta phase precession -- the time a neuron fires relative to the phase of theta rhythm (6–10 Hz) oscillations in the local field potential -- is exhibited by spatial cells of the rat entorhinal–hippocampal circuit and reduces uncertainty about the position of an animal (Hasselmo et al, 2013). It is not clear how this is modeled in connectionist models.

The problem of what elements in the brain correspond to units in probabilistic/associative models is only one problem probabilistic/associative models have in connecting with neurology. Plaut and McClelland (2010a) say

> "there are clearly many aspects of the standard PDP framework that do not emulate known aspects of neurophysiology: the lack of separate excitatory and inhibitory cell populations, the purely linear integration of inputs with no consideration of dendritic geometry, the use of a real-valued symmetric activation function, no consideration of metabolic constraints, and the propagation of error signals back through forward-going connections, to mention only a few . . . as has repeatedly been emphasized, PDP models are generally not intended to emulate all aspects of the underlying neural substrate." (p 287)

I will return to the implications of this statement in my concluding remarks.

Symbol-based procedural models: Neurological Plausibility

The converse of the belief that probabilistic/associative models are neurally realistic is the belief that symbol-based procedural models are not. What is said to make them neurally unrealistic is that they postulate discrete categories and rules, which, it is claimed, cannot be related to neural elements. I will consider the first of these issues here. The question of what elements in the nervous system might encode discrete, unitary categories postulated in symbol-based procedural models corresponds to the question of what elements in the nervous system might correspond to units in probabilistic/associative models.

The argument that discrete, unitary categories are incompatible with neurology is linked to the view that neural systems encode knowledge as connection weights and activated information as distributed patterns of activity. The contrast is between distributed and localist representations. In distributed models, "A concept is represented by a pattern of activity over a collection of neurons (i.e., more than one neuron is required to represent a concept.) Each neuron participates in the representation of more than one concept." In localist models, "each neuron represents a single concept on a stand-alone basis. The critical distinction is that localist units have "meaning and interpretation" whereas units in distributed representation don't." (Quotations from Roy, 2012, p 551).

The strongest instantiation of localist representations would be single cells, often called "grandmother cells," a term first used, derisively, by Jerry Lettvin and later popularized by work by Barlow. The term "grandmother cell" refers to a neuron that "would respond only to a specific, complex, and meaningful stimulus, that is, to a single percept or even a single concept (Gross, 2002, p. 512)." Barlow says "The concept included invariance of response for changes in some variables as well as selectivity of response for others, together with the idea that these cells are created by processing at a hierarchy of levels (quoted by Roy, 2013)."

There are many examples of single cell localist coding in the CNS. The first I came in contact with, as a graduate student, are the type 2 neurons in the frog's retina that Lettvin, Maturana and colleagues (1959) found responded to stimuli with the characteristics of small moving bugs. Many more examples of single neurons that respond to complex biologically salient events or code for complex motor actions are given by Bowers (2009, 2010a, 2010b).

Localist representations have been documented at high neurological levels. Page (2000), Bowers (2009) and Roy (2012, 2013) review results. Striking and much discussed examples are the cells reported by Quian Quiroga et al. (2009) that respond selectively to faces of individuals presented from a variety of angles and with a variety of features. For instance, one cell in medial temporal lobe responded to pictures of the actress Halle Barry even when she was presented as "Catwoman," one of her roles. Some of these cells respond to both pictures and names; e.g., Quian Quiroga et al. (2009) found (separate) cells that responded to both pictures of the actress Jennifer Aniston and the Iraqi president Sadaam Hussein and to their names.

Cells with localist properties are often modeled as the terminals of hierarchies of cells that respond to increasingly more complex sets of features, with categorized entities at the top of the hierarchcy. Hierarchical organization of neural responsiveness was first reported by Hubel and Weisel (1962) in their work on responses of cells in V1 – V3. Poggio and Bizzi (2004) present models that contain cells with broad tuning for elementary features and narrow receptive fields that project to cells with broader receptive fields that respond to combinations of features and then to cells with even larger receptive fields that respond to shapes in a viewpoint invariant way after learning. An example is Riesenhuber and Poggio (1999) who modeled responses of cells in IT to complex shapes under rotation (Figure 7).

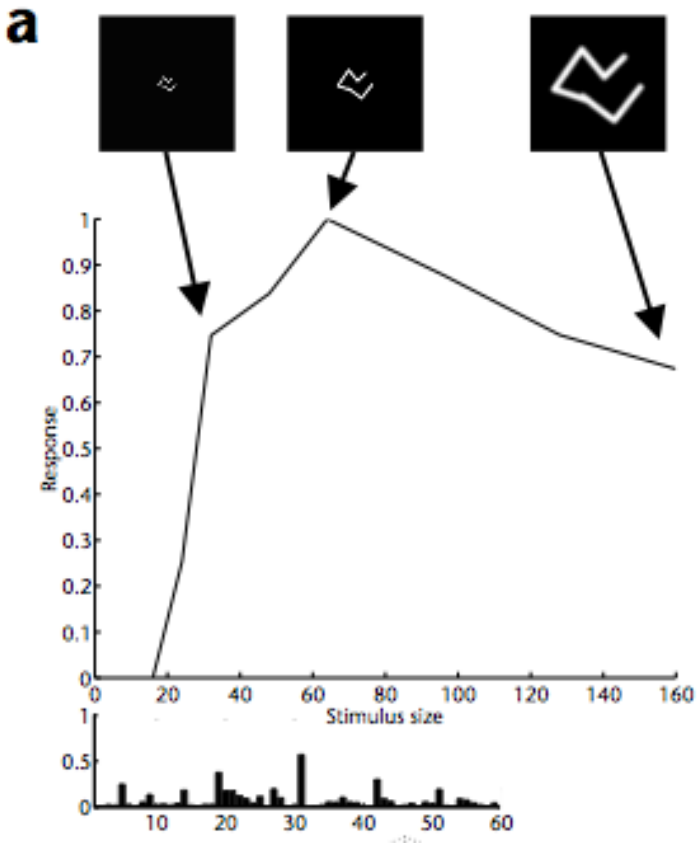

Figure 7: Responses of cells in IT to complex shapes under rotation (Riesenhuber and Poggio, 1999).

Poggio and Bizzi extend the modeling from higher order visual perceptual categorization to motor planning.

Several issues have led to confusion about localist and distributed representations. I will touch on three. One is that information about complex stimuli that may be coded locally at higher levels of the nervous system is distributed at lower levels. For instance, information about the distal visual stimulus that triggers recognition of a face is distributed over retinal neurons. Bowers (2009) insightfully points out, however, that this information is necessary but not sufficient to recognize a face; the information that the distal object is a face, and whose face it is, is not present in the retina. Localist models apply to the representation of this latter type of information. A second point of confusion is that many neurons that selectively respond above a threshold also respond to some degree to other similar stimuli; i.e., their tuning curves are broad. Bowers (2009) points out that there is a difference between what a neuron responds to and what it encodes. Unless the subthreshold activity of a neuron is required to recognize a second item, the representation is localist. The third issue is empirical – have enough cells been sampled in work that documents localist encoding to rule out extremely sparse coding (i.e., might a small number of cells contribute to representing the information that appears to be encoded in one cell)? The answer is unknown, and may not be knowable; if it turns out that information is encoded very sparsely rather than in single neurons, whether this would constitute a neural correlate of a unitary symbolic representation or of a unit in a probabilistic/associative model depends on whether these neurons participate in encoding other pieces of information. Localist neurons have also been argued to have several limitations; one, for instance, is the ability to represent both instances and categories (i.e, to deal with the type/token issue). Bowers (2009, 2010a, b) and Poggio and Bizzi (2004) discuss possible solutions to this problem, which arises in distributed models as well, as far as I can see.

All told, the evidence for localist coding of categorical information is reasonably strong, and provides *prime facie* validity of one important aspect of symbol-based procedural models.

Concluding remarks

I began this presentation by outlining several basic psychological phenomena that have been modeled within two different frameworks – a symbolic/procedural framework and a probabilistic/associative framework. I have argued that there are significant problems within both frameworks in accounting for the nature of concepts, and that the probabilistic/associative framework has not, to date, described or explained combinatorial relations of categories and concepts. The obvious way to proceed, I would think, would be to pursue both approaches, which is, in fact, what scientists are doing. However, there is a sense in at least certain quarters in cognitive psychology that it is a mistake to pursue symbolic/procedural models because they are inconsistent with neurobiology and therefore cannot be correct, and that neural considerations support probabilistic/associative models. In this last section, I will elaborate a bit on this view.

As the quote from Plaut and McCelland above indicates – “as has repeatedly been emphasized, PDP models are generally not intended to emulate all aspects of the underlying neural substrate" – advocates of probablilistic/associative models are well aware of the extent to which they are abstractions and models of neural events. The question that I think needs to be addressed is not whether this approach has any relation to neural events – it clearly is a productive bridging model worth considering. The question is whether it corresponds sufficiently well to invoke neural realism as a reason to accept it as the sole theoretical framework to use to understand human cognition; i.e., is it so well founded neurologically, and the alternative so poorly neurologically connected, that we should discount models developed in the symbolic/procedural framework or invoke neural considerations in adjudication between those models and ones developed in the probabilistic/associative framework (with inevitable advantage to probabilistic/associative models)? I have argued that we should not. On the one hand, probabilistic/associative models models encounter significant problems in modeling neural phenomena. Conversely, there is not-unreasonable evidence that a key feature of symbolic/procedural models has a

neural correlate. Probabilistic/associative models have been *inspired* by results in the neuroscience to a much greater extent than symbolic/procedural have. Whether they are actually *neurally more realistic* is another question, and an open one. Neural facts do not rule out either class of models and, as far as I can see, do not favor one over the other.

A corollary of this conclusion is that the fact that the neural basis for a feature of a model is unclear is not a strong argument that that feature is incorrectly postulated in a theory of cognition; in my view, it is an extremely weak argument against the existence of that feature. Specifically, the fact noted above, that what aspects of neural tissue might encode rules remains very unclear, has little-to-no value in adjudicating between models that do and do not postulate rules. These models need to be evaluated for their descriptive and explanatory value regarding the phenomena that constitute their domain of science. Adding rules incurs costs in terms of degrees of freedom and numbers of variables in a model, and perhaps incurs a special cost because of the power of rules. Whether a theory containing rules should be accepted requires balancing these costs against its descriptive and explanatory successes. This is not an easy task, but the absence of neural correlates of rules does not weigh in the process. At this time, arguably, critical elements of all theories remain in search of neural correlates; accordingly, if neural correlates are required for a theory to be considered, we might reject all existing models. Clearly, this would be a mistake. There was no known biological mechanism that could be the basis of Mendel's laws at the time he developed them, but those laws described and explained important phenomena.

On rare occasions, advocates of probabilistic/associative admit that, at their core, they take the position I am advocating about what data theories are responsible for. For instance, Plaut and McClelland (2010b) say:

> neural verisimilitude per se has not been our primary goal; rather, the [PDP] approach is directed first and foremost at accounting for performance on cognitive tasks as it occurs in real time, how performance changes over the course of normal and abnormal development and in adulthood, as well as addressing individual differences and the consequences of brain damage. (p 289)

Despite our apparent agreement about the data that models are responsible for, Plaut and McClelland and I disagree about the relevance of neural correlates in adjudicating between theories. On the next page of their article, Plaut and McClelland appeal to neurological plausibility in arguing about probabilistic/associative and symbolic/procedural solutions to the type/token problem:

> One common variant of localist theory [allocates] a localist unit to each separate experience with each instance of every different type of entity and then [allows] partial activation of each instance to play a role in determining the output of the system . . . To us, this idea really does not seem biologically plausible (p 290).

I have no idea why it is any more or less plausible than other ideas that have been suggested about neural correlates of types and tokens.

My sense is that advocates of probabilistic/associative models invoke neural considerations in arguing for their models because they believe deeply that their models are neurally realistic. Consider the following astonishing statement by Rogers and McClelland (2014). They say the brain "contains 10–100 billion neuron-like processing units (p 1063)." The significance of this statement is likely to escape the reader on an initial superficially-to-moderately deep reading (as I suspect it did the authors). On reflection, it is a reification of their model, on steroids. The last I looked, brains contained actual neurons, not "neuron-like processing units." The *model* contains "processing units." Whether they are "neuron-like" – and, more broadly, whether the ensemble of features of these models corresponds to neural structures and neurophysiological functions -- are open questions.

A more reasoned, and reasonable, statement is found in Plaut and McClelland (2010b): "The PDP approach, for us, is grounded in the belief that certain computational principles of neural systems are fundamental to explaining human cognitive performance. p 289)" Accepting this, we need to ask what the "computational principles" of neural systems are and which are "fundamental to explaining human cognitive performance." Distributed representations? Back propagation with unsupervised learning? Or feedforward processing to localist neurons that encode combinations of features from lower levels? I will end with the charitable and ecumenical, if theoretically tepid, thought that currently available neurological data are compatible with both types of models.

References

Barlow HB. 1972. Single units and sensation: a neuron doctrine for perceptual psychology. Perception 1:371–94.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116,* 220–251.

Bowers, J. S. (2010a). More on grandmother cells and the biological implau- sibility of PDP models of cognition: A reply to Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010). *Psychological Review, 117,* 300– 308

Bowers (2010b) Postscript: Some Final Thoughts on Grandmother Cells, Distributed Representations, and PDP Models of Cognition, *Psychological Review, 117,* 306-308.

Cohen J. Y., Haesler S., Vong L., Lowell B. B., Uchida N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482, 85–88

Cox, C. Seidenberg, M. and Rogers, T. (2015) Connecting functional brain imaging and Parallel Distributed Processing, Language, Cognition and Neuroscience, 30, \ 380–394,

Chomsky (1994) *The Minimalist Program*, MIT press

Croft, W. (1990) *Typology and Universals,* Cambridge University Press, Cambridge, UK

McDonald, M-E and Christiansen, M (2002), Reassessing Working Memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996), Psychological Review, 109, 35-54.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning, 7, 194–220.

Fodor, J (1998) Concepts: Where Cognitive Science Went Wrong, Oxford University Press

Fodor, J (2000) *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Fodor, J (2008) LOT 2: The Language of Thought Revisited, Oxford University Press.

Frege, G (1892) Über Sinn und Bedeutung, *Zeitschrift für Philosophie und philosophische Kritik*, 100: 25–50. Translated as 'On Sense and Reference' by M. Black in *Translations from the Philosophical Writings*

of Gottlob Frege, P. Geach and M. Black (eds. and trans.), Oxford: Blackwell, third edition, 1980

Hasselmo, M E , Newman E L , Climer, J R (2013) Phase coding by grid cells in unconstrained environments: two-dimensional phase precession, *Eur. J. Neurosci.* 38, 2526-2541

Hubel, D. H. & Wiesel, T. N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160,** 106–154.

Georgopoulos AP, Massey JT (1987): Cognitive spatial-motor processes. 1. The making of movements at various angles from a stimulus direction. *Exp Brain Res*, 65:361-370.

Gross, C. (2002) Genealogy of the "Grandmother Cell", NEUROSCIENTIST 8, 512–518

Jackendoff, Ray (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation, and processing. *WIREs Cognitive Science.* doi: 10.1002/wcs.1340

Lettvin JY, Maturana HR, McCulloch WS, Pitts WH. 1959. What the frog's eye tells the frog's brain. Proc Inst Radio Engin 47:1940–51.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29:375-419.

Kandel, E.R. (2009) The Biology of Memory: A Forty-Year Perspective. J Neurosci 29: 12748-12756

Kotaleski JH, Blackwell KT (**2010**) Modelling the molecular mechanisms of synaptic plasticity using systems biology approaches. *Nat Rev Neurosci.* 11:239

Margolis, E. and Laurence, S (eds) (1999) *Concepts*, Chapter 1, MIT Press; Cambridge MA

McClelland, J. L., Rumelhart, D. E. & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models*. Cambridge, MA: MIT Press.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9, 35-58.

Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23,* 443–512.

Pinker, S (1994). *The Language Instinct*, William Morrow (New York)

Plaut, D. C., & McClelland, J. L. (2010a). Locating object knowledge in the brain: Comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review, 117,* 284–290.

Plaut, D. C., & McClelland, J. L. (2010b). Postscript: Parallel distributed processing in localist models without thresholds. *Psychological Review, 117,* 284–290.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review,* 103, 56–115.

Poggio, T and Bizzi, E (2004) Generalization in vision and motor control, *Nature*, *431*, 768-774

Putnam, H (1975) *Mind, Language and Reality*, chapters 8 and 12, Cambridge University Press; Cambridge, UK

Quine, W.V.O. (1951), Two Dogmas of Empiricism, *The Philosophical Review* 60: 20–43.

Quian Quiroga, R., Kraskov, A., Koch, C., & Fried, I. (2009). Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology, 19,* 1308–1313.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025.

Rosch, E (1978) "Principles of Categorization", in Rosch, E. & Lloyd, B.B. (eds), *Cognition and Categorization*, Lawrence Erlbaum Associates, Publishers, (Hillsdale), pp. 27–48

Rogers T. and McClelland, J. (2014) Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition, Cognitive Science 38 ,1024–1077

Roy, A. (2012). A theory of the brain: localist representation is used widely in the brain. *Front. Psychol*. 3:551.

Roy (2013) An extension of the localist representation theory: grandmother cells are also widely used in the brain, Front. Psychol., 4: 300.

Rumelhart, D. E., McClelland, J. L., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: MIT Press.

Seidenberg, M. S. & Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), *From Inkmarks to Ideas: Current Issues in Lexical Processing*. Hove, UK: Psychology Press

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist networks. Artificial Intelligence, 46, 159–216.

Smolensky, P., Matthew Goldrick, M., Mathis, D (2014) Optimization and Quantization in Gradient Symbol Systems: A Framework for Integrating the Continuous and the Discrete in Cognition, Cognitive Science 38, 1102–1138

Shallice, T. (1988). From neuropsychology to mental structure. Oxford, UK: Oxford University Press.

Shallice, T., & Cooper, R. P. (2011). The organization of mind. Oxford, UK: Oxford University Press.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631–1642). Seattle, WA:

Association for Computational Linguistics.

Spivey, M (2007) The Continuity of Mind, Oxford University Press

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. Blackwell Publishing (2001).

Zadeh, L (1965) Fuzzy sets. *Information and Control*. 1965; 8: 338–353.

Zadeh, L (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets and Systems*, 90, 111–127