Loosemore, R.P.W. & Harley, T.A. (2010). Brains and Minds: On the Usefulness of Localization Data to Cognitive Psychology. In M. Bunzl & S.J. Hanson (Eds.), *Foundational Issues in Human Brain Mapping*. Cambridge, MA: MIT Press.

Brains and Minds: On the Usefulness of Localization Data to Cognitive Psychology

Richard LOOSEMORE
Surfing Samurai Robots, Inc., Genoa NY 13071, USA
rloosemore@susaro.com

Trevor HARLEY
School of Psychology, University of Dundee, Dundee DD1 4HN
t.a.harley@dundee.ac.uk

Abstract. We propose that cognitive science is destined to have three stages in its history, and that the current fashion for brain imaging studies would be entirely appropriate for a cognitive science that had reached Stage 3. This current fashion is unfortunate, because, we argue, cognitive science is currently only in Stage 1, is not yet showing any signs of reaching a mature version of Stage 2, and cannot even begin to move into Stage 3 until the earlier stages are reasonably well developed. Our conclusion is that brain imaging studies are, for the most part, premature, and are taking psychology in a direction it can currently ill afford to go.

Keywords. Cognitive neuroscience, Brain imaging, fMRI, Molecular model.

Introduction

In the early 1990s, one of us wrote a textbook about the cognitive psychology of language; in this book there was hardly a mention of the brain (Harley 1995). There was a great deal of neuropsychology, but it was cognitive neuropsychology; indeed, the book discussed (and tacitly adopted) the position of ultracognitive neuropsychology, which maintains that although we can learn a great deal about how the mind works from looking at the effects of brain damage, we cannot learn very much about the mind from looking at the brain.

How times have changed! The third edition of the same book is full of reports of studies of brain imaging in one form or another (Harley 2008). Furthermore, if you look through a journal such as *Science* or *Nature*, you will find that most articles on psychology contain or refer to imaging. And what now passes as psychology in the popular press is mostly reports on brain imaging studies. The brain is back in cognitive psychology. But is this change for the better? Are we really learning anything new or important about psychology (the science of mental behavior), or rather about the science of the substrate on which the mind is based?

One of us has previously argued (Harley 2004a, b) that psychologists are becoming obsessed with brain imaging, and that this obsession shows no signs of abating; indeed, imaging is now taking over the world. Harley argued that imaging the localization of function of components of anything—parts of cars, computers, or brains—can be described at four levels:

1. The "tokenism" level. In the way that a glossy magazine might print a picture of a luxury kitchen because it looks good, sometimes brain images seem to be included just because they are available, and look good. There's no denying that they do look impressive, but what they add to the science of the paper in *these* cases is either unclear or obviously nothing. We admit that imaging devices are fun to use, are expensive, and require big grant money for their care and feeding, and no doubt all of these characteristics add to their popularity.

2. The "absolute location" level. A car engine is in the middle of the engine compartment. This level of description, of course, assumes that we can identify the engine, or else we are reduced to saying "there is a big box in the middle of the car." The results of knowing just where things are is not really that interesting, for psychologists at least.

3. The "relative location" level. For example, the little fan is above the CPU. This level is more interesting, because it might tell us how things are connected together, but is subject to a new assumption, proximity. This assumption states that adjacent regions of a complex machine do related tasks. While plausible some of the time, this assumption is clearly often wrong (e.g., although proximity between fan and CPU is essential, the fan doesn't carry out computations, it just keeps the CPU cool). Similarly, we cannot assume that just because regions of the brain are near each other, they do similar things. An extension of the proximity assumption is the wiring assumption—if we can spot tracts of fibers, or brain regions mediating between two areas, then they must be working on related tasks.

4. The "functional" level. For example, the engine drives the axles, which in turn drive the wheels. Car manuals provide clear diagrams showing where the components are, what each does, and how they cohere to create a machine that fulfils a particular function. The diagrams explain why components with particular functions are in the places that they are. The trouble here is that you already need to know mostly what each component does before you can begin to make sense of it. In addition, the absolute location of components is rarely critical to the workings of the system; what is more important is the pattern of connectivity.

In an ideal world, brain studies would always be situated at level 4, the functional level. Many of the published reports do indeed make claims that seem to place them near the top of this hierarchy, but in the analysis we present here, we are going to push back on these claims and ask whether the apparent level is the same as the level actually achieved. If we analyze these studies carefully, might we find that those that superficially appear to be at level 3 or 4 are really only at level 2 or 1? We have long suspected that even those brain imaging studies that appear to give us valuable information might actually be suffering from subtle ambiguities or confusions that could invalidate their conclusions.

What we propose to do now, then, is deepen this "levels" analysis by putting six recent brain imaging studies under the microscope.

Before we describe the specific method of analysis to be used on these six studies, here are some of the general background questions we hope to answer:

• Is “where-it-happens” information of any use to present-day psychology? A modern automotive engineer, for example, might find it useful to know that a heavy battery was located in the rear of a car, because that might imply the car was a hybrid; but how useful would this information have been 150 years ago, when internal combustion was barely imagined? At some point in time, when we have an adequate model at the higher levels of analysis, this kind of information undoubtedly becomes useful, but is it useful today?

• Are the specific claims about neural localization internally and theoretically coherent? For example, it would be incoherent if a localization study claimed to have found a particular mechanism, but the putative mechanism could do the job it is supposed to be doing only if every one of the brain’s neurons was directly connected to all of the others: Since we know that neurons are not totally connected, we know the mechanism cannot be right, and so claims about its location would be incoherent.

• Do the localization claims refer to components of the mind that are clearly defined? We would not be interested in an fMRI study that located the id and superego, for example, unless the author could say exactly what the id and superego are supposed to be, how they work, and why we should believe that there are such things.

• Are the inferences researchers make theory-laden? Do they depend on acceptance of a particular functional theory of how the mind works? For example, it would make no sense for someone to establish the location of “consciousness” if later developments in our theory of cognition show that consciousness is not a physical place at all, but a process involving many scattered components.

• Are today’s studies, which give us only crude localization data (both spatially and temporally), just a prelude to later research that will pin down location and function precisely? Or is crude localization all that we can ever expect from brain imaging technology?

Over the last year or so we have kept track of the reports of psychology and imaging we have come across in the popular press. We have made no attempt to be exhaustive, or to make the sample random or representative in any way, so we make no claims to methodological rigor about what is presented in the press; we just wanted to generate a sample of what the popular press finds interesting about psychology and the brain. We certainly have no intention of denigrating these studies just by picking them, or because they have been widely reported in the press. (Indeed, we believe that it is important for scientists to publicize their work, and to explain science to a wider audience.)

Of course, it could be that our sample is biased toward where-it-happens studies because the press is obsessed just with where things are in the brain, rather than any more detailed level of explanation. We don’t believe this for a moment: We have yet to come across an article titled “Hippocampus near amygdala shock horror!” No, the press is interested in behavior, explanations of behavior, and unexpected connections between types of behavior—just the things they should be interested in. Nevertheless, the press has clearly bought the “where it is, is important in explaining how and why” argument.

In each of the studies, our main focus will be on what the article claims to tell us about location, and how that information relates to the background theories in cognitive science. Do we feel better informed after reading that such-and-such a function happens in, near, or connected to a particular brain structure, and if (to anticipate a little) it turns out that we do not feel satisfied, can we be more specific about the source of our dissatisfaction?

Using a Theoretical Framework as a Tool

One persistent feeling we get, when reading the brain imaging literature, is that stated conclusions often seem reasonable if psychological mechanisms are interpreted in a relatively simple or simplistic way, but these same conclusions could easily become unreasonable if the mechanisms are interpreted in other ways, or if they are implemented in a less-than-obvious manner. Brain imaging conclusions, in other words, often seem theory-dependent and vulnerable to any future winds of change that might blow through cognitive science.

In order to test this intuition about theory-dependence, we are going to adopt a rather unusual strategy. In this chapter, we sketch the beginnings of a new, unified framework that describes the overall architecture of the human cognitive system, and we use this framework to ask how well the conclusions of our target brain imaging studies would hold up if this framework should one day become the standard, functional-level model of how the brain works.

The framework has some unusual features, and it is these features in particular that we believe could do some damage to the conclusions some imaging studies have arrived at.

We should be clear about what we are trying to achieve in proposing this new framework. We are not really trying to suggest that we have come up with a new interpretation of each of our target studies, which should then be taken as a challenge to be overcome by new, more cleverly designed imaging studies. Of course, we would be happy if our proposals were taken in this light, because this kind of interaction between theory and experiment is the mark of a healthy scientific paradigm, but this is not our main intention. Our real goal is to see how sensitive the conclusions of these studies might be to a slight change in the theoretical mechanisms whose location is being sought.

Our overall goal, then, is to use this new framework as a tool with which to try to break the conclusions of brain imaging studies that purport to tell us something about localization of cognitive functions.

A Molecular Framework for Cognition

The alternative framework we wish to put forward is a "molecular" model of cognition. It has connections to many previous lines of thought in cognitive science, but it is most closely inspired by Douglas Hofstadter's *Jumbo* model, which was originally intended to explore the cognitive processes at work in an experienced anagram solver (Hofstadter 1995).

The framework is intended to describe only the higher, more abstract, levels of cognition, above the level at which purely data-driven processing occurs.

The Core Concept: Instances

At the heart of this framework lies one key idea: a distinction between *instance* and *generic* versions of the concepts stored in the system. In much of cognitive science the idea of a concept is used as if there were only one entity encoding the concept; so, for example, theorists will talk about *the* [coin] node becoming strongly activated, or about the priming effect this can have on *the* [bank] node. It is tempting to imagine a large network of nodes (or even neurons), with a [coin] node and a [bank] somewhere in the network, and with vast numbers of connections between all the nodes.

But any complete model of a cognitive system must include *instance* nodes that represent the particular entities involved in our thoughts at a given moment: nodes that represent, not coin in general, but the particular instance of the word *coin* that is being witnessed right now. Any realistic model of cognition must make explicit allowance for these instances, and it turns out that this can have a drastic effect on our theorizing. Instance nodes do not sit quietly in a fixed network; they are created on the fly, they have a relatively short lifetime, and the connections between them are extremely volatile. Furthermore, a complete model should explain how the generic concepts are built up from repeated exposure to specific instances.

The primacy of instances is the core concept behind the proposed molecular framework. There is a deep assumption that, in practical terms, the place where these instances are created, interact, and have their effects on the rest of the system is likely to be far more important than the passive network of generic concepts.

In other respects, the framework is little more than a conjunction and distillation of the most common features of many local theories, though with a bias toward the abstract motivations that drove, among others, McClelland and Rumelhart (McClelland et al. 1986).

Foreground and Background

In our framework there is one main type of object, and two main places.

The objects are called *atoms*, and their main purpose is to encode the smallest packets of knowledge (concept, symbol, node, etc.). Atoms come in two sorts: *generics* and *instances*. For each concept, there is only one generic atom, but there can be many instance atoms. In what follows, the term *atom* on its own will usually be understood to mean an instance atom.

The two main "places" in this framework are the *foreground* and the *background*.

The foreground roughly corresponds to working memory, and is the place where instance atoms are to be found. The foreground is an extremely dynamic place: Atoms are continually being created, and while they are active they can move around and form rapidly changing bonds with one another. The sum total of all the atoms in the foreground, together with the bonds between them, constitute what the system is currently thinking about, or aware of.

The background is approximately equivalent to long-term memory, and is just a store of all the generic atoms from which instance atoms can be made. When an instance atom is in the foreground, it maintains a link back to its generic parent in the background. The background is more or less passive; the foreground is where everything happens.

Note that atoms do not necessarily encode concepts that have names. Some of them capture regularities at a subcognitive level, and for this reason the foreground contains some activity that the system is not routinely aware of, or that it does not find easy to introspect or report on (see Harley 1998, for more detail on this point).

Active Representations and Constraints

So far, this is all sufficiently general that it could be the outline of many different theories of how the cognitive system is structured. But now we will make a commitment that distinguishes this framework from many others: The representations in the foreground are not passive tokens of the sort that are meant to be assembled and used by some external mechanisms, they are *active* representations. In other words, although the atoms encode knowledge about the world in just the way you might expect, they also encapsulate a set of mechanisms that implicitly define how this knowledge is used by the system.

How do the foreground atoms do this? Broadly speaking, each atom contains (and continually updates) a set of constraints that it would like to see satisfied by its neighbors in the foreground. For example, the [chair] atom would prefer to see a group of atoms around it that encode the characteristics and components of a typical chair, and these preferences, encoded inside the atom, are what we refer to as the *constraints* it is seeking to satisfy.

An atom will not just passively seek a place where its constraints are satisfied, it will actively try to force its neighbors to comply with its constraints. Its behavior is a mixture of “Do my neighbors suit me?” and “Can I change my neighbors to better suit me?” An atom can engage in several kinds of activity in pursuit of its goals: It can try to activate new atoms that it would like to see in its neighborhood, or deactivate others that it does not want to see, or change its internal state, change the connections it makes, and so on.

Not all of the atoms in the foreground are successful in their attempts to satisfy their constraints. Many get woken up because something thinks they might be relevant, but after doing their best to fit, they fail to establish strong bonds and become deactivated. A substantial number of atoms, however, do find themselves able to connect to an existing structure and thereafter play a role in how that structure develops.

As a general rule, the constraints inside an atom are supposed to encode a “regularity” about the world—say, the regularity that if something is a chair, it must possess constituent parts that could be construed as legs.

Relaxation

The process in which an atom looks around and tries to take actions to satisfy some local constraints can be characterized as *relaxation.* Everything that happens in the foreground is driven by relaxation. If a change occurs somewhere in the foreground, atoms adjacent to the change will try to accommodate to it, and this accommodation will propagate out from the starting point like a wave, or domino effect.

There does not have to be only one type of relaxation happening at every place; several such processes can occur simultaneously, originating in different parts of the foreground. Almost everything that happens in the foreground can be attributed to changes that occur in a particular place and then send waves of relaxation that propagate across the foreground. Strictly speaking, this is dynamic relaxation, because the foreground never settles into a quiescent state where everything is satisfied; it is always on the boil.

The Foreground Periphery

At the edge of the foreground are a number of areas that connect to structures outside the foreground. One part of the periphery has input lines coming from sensory receptors, while another has output lines that go to motor output effectors. These two patches on the edge of the foreground are responsible for much of the activity among the atoms. When a signal arrives from a low-level sensory detector (or more likely a data-driven system that preprocesses some raw input signals), the result is that one or more atoms are automatically attached to the foreground periphery to represent the signal. These initial atoms then have an effect on nearby atoms.

Another important part of the periphery is a connection to mechanisms that govern the goals, drives, and motivations the system is trying to satisfy. These can be thought of as lying "underneath" the foreground, in the sense that they are more primitive than the cognitive work that goes on in the foreground proper. The work of these motivational/drive systems is not simple, but the effect of their actions is to bias the activity in one direction or another.

Sources of Action

So far, atoms have been characterized as if their role is just to encode knowledge, but in fact, representing the world is only part of what they can do: Some atoms initiate *actions*, and some may encode mixtures of action and representation.

How does the system "do" an action? In the part of the foreground periphery where input arrives from the motivational/drive system, there is a unique spot—the "make-it-so" place—that drives all actions. If an action atom can get itself attached here, this triggers an outward-moving relaxation wave that will (usually) result in signals being sent from the motor output area that cause muscles to do something. The make-it-so place, in the part of the foreground periphery we have called the motivation area, is the place where the buck starts.

Neural Implementation

Nothing has been said, so far, about how this framework is realized in the brain's neural hardware. There are many possibilities here, because this framework is primarily designed to specify high-level mechanisms. The framework itself is neutral with respect to neural implementation.

With that qualification, what follows is a quick sketch of one way it could be realized in the brain. This neural implementation will be assumed in the rest of the chapter.

The cortex could be an overlapping patchwork of "processors," each of which can host one active atom, and the sum total of all these processors is the thing that we have called the *foreground*. Each processor is a large structure with quite complex functionality, and is capable of doing such things as hosting a particular atom for a while, transferring a hosted atom to an adjacent processor if there is pressure for space, and setting up rapidly changing communication links to other atoms located some distance away across the cortex. One possibility is that the central core of each processor corresponds to a cortical column.

The generic, passive atoms that are stored in the background (long-term memory) are colocated with the processors that were just described, but each processor can hold a large number of generics. Each of these passive atoms is encoded in distributed form inside the processor. At the level of the entire cortex, then, a generic atom would seem localized (because it is within one processor), but since each processor is quite extensive, the atom is not at all localized within the processor.

The activation of an instance atom involves a call to the processor that hosts the generic, which causes the processor to find a spare processor that can host the instance atom. The parent processor itself might be able to play host, but if not, then it passes the atom to some other processor (possibly a neighbor) for hosting. One way or another, an activated atom is quickly copied into a processor that can handle it, in much the same way that a computer program might be transferred across a network to a computer that can run it.

In summary, then, the cortex can be viewed in two completely different ways. It can be seen as the foreground, in which case it is effectively a space within which the extremely volatile instance atoms arrive, set up a rapidly changing set of links to other instance atoms, and then depart. Alternatively the cortex can be seen as the place where the contents of long-term memory are stored, since the generic atoms are also located in the processors.

Example 1: Visual Object Recognition

We conclude this summary of the molecular framework with two examples of the kinds of activity that go on in the foreground. The first involves the recognition of an object perceived with the eyes, and the other is the carrying out of an action.

Suppose the system were to start in a thought-free, meditative state, in a darkened room, and then suddenly a light comes on and illuminates a single chair. The foreground would start out relatively empty, and when the light is turned on a sequence of atoms would come into the foreground until, at the end of the process, the [chair] atom would be activated.

This process would begin with the arrival at the foreground periphery of signals along the lines coming from the visual system. When these signals arrive, they trigger the activation of instance atoms representing some low-level features of the visual input. When these atoms appear in the foreground, they attach to the places on the periphery where their features occur, and then they look around at their closest neighbors. If a pair (or perhaps a group) of these atoms recognize that they have occurred together before, they will know about some other atoms that they could call into the foreground, which on those previous occasions represented their co-occurrence. They will activate those second-level atoms, and these in turn will try to attach themselves strongly to the first-level atoms. The ones that succeed in making bonds will then feel confident enough to look around at their neighbors and repeat the process by calling in some even higher-level atoms. An inverted tree of atoms thus develops over the part of the foreground periphery where the chair image came in, and at the top end of this inverted tree, finally, will be the [chair] element.

This picture of the recognition of a chair is extremely simplified, but it gives a rough picture of the kind of activity that occurs: atoms being activated by others, then each atom trying to fit into a growing structure in a way that satisfies its own internal constraints about the roles it can play. As a whole, the system tries to relax into a state in which some atoms have self-consistently interpreted the new information that arrived.

Example 2: Sitting Down

Suppose the person who just recognized the chair next hears the experimenter ask them to sit down. The atoms representing this request will (after an auditory recognition process very similar to the visual recognition event previously described) bump into, and interact with, the large cluster of atoms hovering around the motivation area of the foreground periphery. This cluster represents the mind's complex stack of goals and intentions, all the way from its most nebulous motivations (stay safe, seek warmth, get food, etc.) to its most specific action schemas (accept this experimenter as a trusted friend whose requests should be obeyed). As a result of this interaction between the atoms representing the request and the atoms around the motivation area, a new atom that encodes the sitting action is activated and attached to the make-it-so place on the foreground periphery, and this gives the [perform-a-sitting-action] atom enough strength to cause a cascade of other atoms to be brought in, which then elaborate the sitting plan in the context of where the body currently is, where the chair is, and so on.

In the case of the sitting-down action, a wave of relaxation emerges from the [perform-a-sitting-action] atom and causes a sequence of atoms to spread toward the motor output area. When these atoms arrive at the periphery, muscles move and the sitting action occurs. The only difference between this and visual recognition is that the wave of relaxation does not come from the sensory input area and end with the activation of the [chair] atom, but starts with one atom at the make-it-so place and ends with a broad front of new atoms hitting the motor output area.

Applying the Framework

With this theoretical framework in hand, it is time to examine the claims made in each of our target brain imaging papers, to try to understand those claims in the context of both a regular interpretation of cognition, and the new framework.

Study 1: "'Bottleneck' Slows Brain Activity"

Dux, Ivanoff, Asplund, and Marois (2006) describe a study in which participants were asked to carry out two tasks that were too hard to perform simultaneously. In these circumstances, we would expect (from a wide range of previous cognitive psychological studies) that the tasks would be serially queued, and that this would show up in reaction-time data. Some theories of this effect interpret it as a consequence of a modality-independent "central bottleneck" in task performance.

Dux et al. used time-resolved fMRI to show that activity in a particular brain area—the posterior lateral prefrontal cortex (pLPFC)—was consistent with the queuing behavior that would be expected if this place were the locus of the bottleneck responsible for the brain's failure to execute the tasks simultaneously. They also showed that the strength of the response in the pLPFC seemed to be a function of the difficulty of one of the competing tasks, when, in a separate experiment, participants were required to do that task alone. The conclusion Dux et al. drew is that this brain imaging data tell us the location of the bottleneck: It's in the pLPFC. So this study aspires to be level 2, perhaps even level 3: telling us the absolute location of an important psychological process, perhaps telling us how it relates to other psychological processes.

Rather than immediately address the question of whether the pLPFC really is the bottleneck, we would first like to ask whether such a thing as "the bottleneck" exists at all. Should the psychological theory of a bottleneck be taken so literally that we can start looking for it in the brain? And if we have doubts, could imaging data help us to decide that we are justified in taking the idea of a bottleneck literally?

What Is a "Bottleneck"?

Let's start with a simple interpretation of the bottleneck idea. We start with mainstream ideas about cognition, leaving aside our new framework for the moment. There are tasks to be done by the cognitive system, and each task is some kind of package of information that goes to a place in the system and gets itself executed. This leads to a clean theoretical picture: The task is a package moving around the system, and there is a particular place where it can be executed. As a general rule, the “place” has room for more than one package (perhaps), but only if the packages are small, or if the packages have been compiled to make them automatic. In this study, though, the packages (tasks) are so big that there is room for only one at a time.

The difference between this only-room-for-one-package idea and its main rival within conventional cognitive psychology is that the rival theory would allow multiple packages to be executed simultaneously, but with a slowdown in execution speed. Unfortunately for this rival theory, psychology experiments have indicated that no effort is initially expended on a task that arrives later, until the first task is completed. Hence, the bottleneck theory is accepted as the best description of what happens in dual-task studies.

Theory as Metaphor

This pattern of theorizing—first a candidate mechanism, then a rival mechanism that is noticeably different, then some experiments to tell us which is better—is the bread and butter of cognitive science. However, it is one thing to decide between two candidate mechanisms that are sketched in the vaguest of terms (with just enough specificity to allow the two candidates to be distinguished), and making a categorical statement about the precise nature of the mechanism. To be blunt, very few cognitive psychologists would intend the idea of packages drifting through a system and encountering places where there is only room for one, to be taken that literally.

On a scale from metaphor at one end to mechanism blueprint at the other, the idea of a bottleneck is surely nearer to the metaphor end. How many cognitive theorists would say that they are trying to pin down the mechanisms of cognition so precisely that every one of the subsidiary assumptions involved in a theory are supposed to be taken exactly as they come? In the case of the bottleneck theory, for instance, the task packages look suspiciously like symbols being processed by a symbol system, in old-fashioned symbolic-cognition style. But does that mean that connectionist implementations are being explicitly ruled out by the theory? Does the theory buy into all of the explicit representation issues involved in symbol processing, where the semantics of a task package is entirely contained within the package itself, rather than distributed in the surrounding machinery? These and many other questions are begged by the idea of task packages moving around a system and encountering a bottleneck, but would theorists who align themselves with the bottleneck theory want to say that all of these other aspects must be taken literally?

We think not. In fact, it seems more reasonable to suppose that the present state of cognitive psychology involves the search for metaphorlike ideas that are described *as if* they were true mechanisms, but which should not be taken literally by anyone, and especially not by anyone with a brain imaging device who wants to locate those mechanisms in the brain.

Molecular Model of the Bottleneck

How would the molecular framework explain the apparent bottleneck in dual task performance?

When the cognitive system decides to carry out an action—like responding to an aural cue with a finger press—what happens in the foreground is that the cluster of atoms that encode the finger-response action get attached to the make-it-so spot on the edge of the foreground. By design, this spot is not allowed to play host to more than one controlled action sequence, where a “controlled” action is one that requires attention.

But now, what is “attention”? One part of the foreground contents (not the foreground itself, notice, but a subset of the atoms that inhabit the foreground) always has a special status: This is the *attentional patch*. The attentional patch can move around, but it is defined by the fact that atoms in the patch are able to spawn large numbers of associated atom clusters, which means that whatever the attentional patch is representing, it is representing it in exceptional detail. Another way to say the same thing is that this is a region of high *elaboration density*.

Now consider an atom that encodes an action that has only recently been learned (say, the pressing of a button in response to an aural cue). Because this atom is relatively young, it needs to attract the attentional patch to it in order to function; in other words, it cannot be executed unless it is explicitly attended to. If an action atom becomes well learned (like the action of sitting down), it does not need the extra boost of being at the center of the attention patch, so the action is allowed to happen while the attentional patch is elsewhere.

When the first task arrives, in this experiment, the atom encoding the response becomes attached to the make-it-so place, then grabs the attentional patch and does not let go of it until the task is completed. Only when the first task relinquishes control is the second task allowed to become attached to the make-it-so place.

What does this mean for the conclusion of the Dux et al. experiment? One possibility is that the pLPFC lights up when a second task atom arrives, asking to be executed as soon as the first is done. Perhaps the pLPFC is just part of the mechanism that manages a competing task, or perhaps it is a buffer where the atoms encoding the second task await their turn to be executed. Under these circumstances, the pLPFC would not be the “location” of the bottleneck at all, but just a region encoding part of the mechanism related to the bottleneck.

Most important of all, the fact that the pLPFC is involved would tell us nothing about the competing theoretical ideas for explaining the bottleneck: the not-enough-room-for-two-packages mechanism, and the molecular mechanism that involves the management of the attentional patch and the make-it-so attachment point. It is certainly not correct to say that discovering the role of the pLPFC tells us where the bottleneck is, or that there is such a thing as a simple bottleneck. Without having an adequate psychological theory first, the imaging data tells us much less than it first seems to.

It might be worth showing the popular-science interpretation of this study, which appeared on the BBC website on January 29, 2007:

> US researchers have discovered a likely reason why people find it hard to do two things at once. A “bottleneck” occurs in the brain when people attempt to carry out two simultaneous tasks, the research shows. The study found the brain slows down when attempting a second task less than 300 milliseconds after the first. The findings support the case for a complete ban on the use of mobile phones when driving, the team said.

This same conclusion could have been reported on the strength of cognitive psychological studies alone, and has nothing to do with the specific facts reported in this experiment. Did the researchers discover the “reason” why people find it hard to do two tasks at once? Sadly, no.

Study 2: "Love Activates the Same Brain Areas as Cocaine"

Aron, Fisher, Mashek, Strong, Li, and Brown (2005) used fMRI to try to distinguish between two possible interpretations of what happens when a person is afflicted with the early stages of romantic love. They asked if this kind of love is a strong emotion, or an overwhelming desire to achieve an objective. The researchers showed pictures of the object of affection to a number of individuals who claimed to have been recently smitten, and their main finding was that "romantic love uses subcortical reward and motivation systems to focus on a specific individual." They declared that "romantic love engages a motivation system involving neural systems associated with motivation to acquire a reward rather than romantic love being a particular emotion in its own right."

The interpretation, then, is that the subjects are not just experiencing a strong feeling, they are wanting to acquire something. If correct, the study is telling us something new at the psychological level, so it is apparently a level-4 (functional) study.

Hence, this study is potentially important and useful. But does it, in fact, tell us something important and new about the mechanisms involved in romantic love?

If the molecular framework is accurate, then the thing that drives a cognitive system to do something is activity coming from deep systems (outside the foreground), which impinge on the motivation part of the foreground periphery, and the way that these drives affect the foreground is through relaxation effects on atoms near the motivation area. Exactly how this works is an open question, but does the Aron et al. study help us to settle this question? Well, it tells us that when reward/motivation systems are active, dopamine is involved, and that romantic love involves activity in those dopamine-rich areas. But does this tell us that dopamine release *causes* the motivational mechanism to kick in? or that dopamine release is a side effect of the motivational mechanism doing something? None of these questions are clarified by the experimental result that some particular areas are activated when the subject looks at a picture of his or her beloved.

The molecular framework could explain romantic love by postulating that there is a specialized slot at the edge of the foreground that has room for precisely one atom that encodes a person, and that when an atom manages to get into that slot it stays there for a long time, kicking in a powerful drive mechanism that tends to force the foreground to engage in certain kinds of thoughts about that person. The original function of this mechanism is to make human beings form a sudden, strong bond to an individual for mating and child-nurturing purposes. Given the amount of activity involved in this unusual mechanism, the framework predicts that there is probably a place in the brain that lights up when this happens, but that prediction by itself is trivial. What matters is exactly how this mechanism exerts its effects.

Is it surprising that a person in early-stage romantic love is experiencing a strong motivation to get various rewards associated with the beloved (wanting to touch, wanting to possess, wanting to receive attention and affection, etc.)? Psychological studies, both formal and informal, tell us that this must surely be the case.

Is there any sign, in this study, that we know more about how the motivational mechanisms work, after finding out that motivational areas are involved in romantic love? As far as we can see, there is no hint of such further information.

Study 3: "Why Your Brain Has a 'Jennifer Aniston Cell'"

Quiroga, Reddy, Kreiman, Koch, and Fried (2005) studied signals coming from an array of several hundred electrodes in the brains of subjects who were undergoing exploratory tests to find an epileptic focus, prior to surgery. When pictures of famous people, landmark buildings, animals, and objects were shown, the experimenters were able to find strong responses to several of the images: On average, fourteen out of ninety-four images elicited a significant response.

Having found some images that caused a response, the experimenters then carried out a testing session in which they showed a number of views of the people or things in those images—and in some cases, they showed only the name of what was in the image. What they found was that the same neurons that responded in the first phase of the experiment also responded strongly to different views of the same subject, and even to the name of the subject written in words. These neurons were very specific: By and large, they did not respond to any other images, only to variants of the one that first triggered them.

The conclusion that Quiroga et al. draw from this is that perhaps grandmother cells (Barlow 1972), which encode single concepts in an abstract way, do exist after all. On the face of it, it seems unlikely that the results could be explained if a distributed representation encodes these images. In the classic type of distributed representation, many units would encode a set of features, and any single image such as Jennifer Aniston's face would be represented only by a pattern of activation across many neurons. Quiroga et al. would have us believe that the brain represents these concepts in a sparse, rather than distributed manner, with a small number of neurons being specifically dedicated to each concept. Although this was not an imaging study, this is clearly an important result, if the preceding interpretation is correct, and could be described as a level-2 account, perhaps being capable of extension to level 3. But are there alternative explanations?

Observations

The first thing to note about this study is the strange fact that the experimenters found some neurons that just happened to respond to the chosen pictures. Who would have thought that when you put several hundred electrodes into the brain, and then show the brain roughly a hundred different images, that some 14% of the images would score a direct hit? If the experimenters' conclusion about sparseness of encoding is correct, the chances of finding the particular neurons that respond *just* to, say, Jennifer Aniston must be very small indeed. Multiply that by 14 (since, on average, fourteen out of ninety-four pictures elicited a significant response in the screening part of the experiment), and we seem to have a problem.

Sparseness of encoding is a conjunction of two ideas. First, there has to be some strong specificity in the response of the neurons: A neuron that fires strongly to Jennifer Aniston's face cannot also respond to the faces of the (superficially similar) Julia Roberts or (thematically related and quite similar) Courtney Cox. Second, sparseness means that there cannot be too many neurons doing the same job. That doesn't necessarily mean there should only be one neuron per job (the mythical grandmother cell that is the only neuron encoding your grandmother's face), but there shouldn't be a million of them, either, or the "sparse" label would start to look inappropriate.

If the experimenters in this study were lucky enough to find neurons that encoded for 14% of the small sample of pictures shown to subjects, then one possibility is that large numbers of duplicate neurons encode each image. This leads to the following problem: If there are so many duplicate neurons encoding image-concepts (enough to enable hits on 14% of the ninety-four pictures), then how much room is there in the brain for the many thousands of other images and concepts to which we can give a name, or that we know, or that we might ever have to distinguish, or might come across in the future? If each neuron represents only one image-concept, and if Jennifer Aniston neurons are so common that a random probe easily finds one, then how much room can there be for other stuff? And what happens when we come across a new face, or object? Do we have a bank of idle neurons waiting to be recruited for the face of the next starlet? Or do we kidnap others that have been doing other jobs that have lapsed into obsolescence?

The simple conclusion the researchers drew in this case is not supportable without further argument. It has significant theoretical ramifications that were not addressed by the authors.

A Molecular Account

What happens when we try to account for this experiment using our molecular framework? Recall that when an instance atom is activated it tries to find a processor, somewhere near where its generic parent lives, where it can start work. We now make a reasonable assumption: Suppose that the atoms tend to be instantiated in roughly the same places each time they come up. So if I see an image of Jennifer Aniston now, and then again in ten minutes, the instance atoms for Jennifer Aniston will tend to be in the same place in my foreground (i.e., hosted by the same processor) on each of the two occasions.

The total number of atom-processors in the foreground is relatively small (perhaps in the thousands, as opposed to the hundred billion neurons in the whole brain), so if Quiroga et al. were looking at a part of the brain that was mostly doing high-level processing (as was indeed the case: they kept well away from the low-level vision areas), it would be reasonable to suppose that a random probe would be relatively likely to score a hit on the processor hosting the instance atom that represents Jennifer Aniston.

This idea could explain the results. If the physical structure occupied by an active atom (what we have referred to as the "processor" that hosts the atom) had a moderately large footprint in the foreground, the chance of an electrode landing somewhere in that processor would be quite reasonable. Then, when different views of the same image are shown in the second part of the experiment, the atom for that image would tend to be instantiated in the same spot, and the same neuron would fire strongly each time. Also, since this is the area where high-level concepts are active, the words "Jennifer Aniston" would be just as likely to invoke the same atom.

Now compare this with the conclusion drawn by the experimenters. There are no grandmother cells in this molecular picture; the generic concepts from which the atoms are spawned could be encoded in any way at all, because the electrodes were picking up instances (the active atoms), not the generics. The distinction between generic and instance representations, in fact, is entirely missing from the interpretation of this experiment (a deficit that is shared by many neuroscience studies).

Whichever way we turn, then, there is no evidence for grandmother cells or sparse encoding in this study. If we try to take the sparse encoding idea literally, the results seem strangely improbable, and if we look at our alternative theoretical explanation, the results have no relevance to long-term memory encoding at all.

Study 4: "Subliminal Images Impact on Brain"

Bahrami, Lavie, and Rees (2007) gave participants a visual task to perform, but varied the amount of attention the task demanded. At the same time, participants would see images of tools in peripheral visual areas, but with one eye getting the tool images and the other eye getting a flashing mask in the same place. Because of the masks, the tool images could not be consciously seen, but an analysis of fMRI data from the retinotopic V1 area showed that these tool images were indeed being detected and processed at that stage of the visual pathway. The crucial result was that the amount of activity in V1 associated with the nonconscious tool images was modulated by the amount of attention the main task required: When the attentional load was high, there was less activity in V1.

The immediate conclusion of Bahrami et al. was that unconscious processing of visual stimuli could depend on the degree to which attentional resources were available. This enabled them to say that "These findings challenge previous suggestions that attention and awareness are one and the same (Baars 2005; Mandler 2005) or that attention acts as the gate-keeper to awareness (Block 1996; Lamme 2003)."

If this interpretation is correct, this study should count as level 4 (functional), in that it uses localization data to relate brain function and location and psychological processes, and furthermore enabling us to distinguish between theories at the cognitive level of theorization. Although these conclusions were among the most robust of those we studied in our brief survey, the molecular framework would nevertheless give a slightly different interpretation of the results—and the difference might be enough for those advocating the two rival accounts listed by Bahrami et al. (Baars 2005 and Mandler 2005; Block 1996 and Lamme 2003) to claim that their theories were not necessarily inconsistent with the results after all.

To see why, consider what might be happening if the foreground zone encompasses visual area V1. As described earlier, there is an attentional area, which is a moving patch of high "elaboration" density in the atoms that inhabit the foreground—a subset of the atoms that are able to call up large numbers of others, so as to build a more detailed representation than would otherwise be the case. But when the attentional patch becomes large (as would happen when a task is attentionally demanding), it causes a corresponding thinness in the density of atoms available elsewhere. This thinning of the atoms could stretch out as far as V1. This, in turn, would mean that the atoms being activated in V1 to represent the peripheral tool images would be struggling to form a strong, coherent representation, because strength is partly governed by weight of numbers.

This is straightforward enough, and it gives an interpretation of why there might be clusters of atoms in V1, representing the tool images, that were stronger and more noticeable to the fMRI scan when the attentional load was not as high.

But what determines whether these tool images make it to conscious awareness? This can happen only if the foreground atoms can switch from their current mode (in which the foreground is dominated by atoms related to the primary task) to a new mode in which the system tries to recall and reassemble the atoms that, a few moments ago, were trying to represent the peripheral tools. If the foreground brings back enough of those atoms (which will include those representing both the tool and the intrusive flashing mask), there is a chance that they can cohere enough to form a representation of the tool, at which point the system would move toward a valid conclusion about where the tool image was located. But if the mask created enough noise (in the form of spurious atoms not related to the tool image), then this reconstruction may fail, and when the participants do this introspective examination of their awareness, they may come up with nothing. The tool images would be invisible, not because they caused nothing to happen in the foreground, but because when the foreground attempts to give attention to the place where the tool images might be, it comes up with nothing but noise.

What is interesting about this molecular account is that attention and awareness are *processes*, not places, and they are extremely closely coupled. Bahrami et al. are correct to reject the claim that “attention and awareness are one and the same,” but the fact remains that terms like *attention* and *awareness* are used in the literature in widely differing ways, and so this study’s conclusion is not as clear as it might seem. If someone were to interpret the Bahrami et al. result to mean that there are two distinct places that control attention and awareness, that conclusion would be false if the molecular account turns out to be correct, because the latter predicts that the two are almost completely enmeshed in one another and separate only under special circumstances.

These questions beg for further theoretical and experimental clarification, but while this particular study makes an interesting and (within limits) valid point about a separation between attention and awareness, the knowledge of that separation does not do much, if anything, to illuminate the detailed differences between the molecular framework and other possible models of attention. Again, we find that the usefulness of the imaging data is circumscribed by the lack of a sufficiently detailed psychological theory.

Study 5: “Brain Scans Can ‘Read Your Mind’”

Haynes, Sakai, Rees, Gilbert, Frith, and Passingham (2007) wanted to know if they could decode their subjects’ *intentions* (not their explicit motor activity, or preparation to perform a motor activity, but their intention to carry out an abstract idea) from the spatial layout of brain activity in the prefrontal cortex. The subjects were intending to perform either an addition operation or a subtraction, and Haynes et al. did indeed find that they could recognize distinct patterns corresponding to the two intentions.

Two conclusions emerged. One was that the intention manifested itself in a spatial pattern of activity, rather than in the overall level of activity in a particular area. The second conclusion was that the location of this pattern differed in the two cases of (a) thinking about the intention and (b) carrying out the intention: During task execution, a more posterior region of prefrontal cortex was involved, whereas during the intention phase the medial prefrontal cortex showed a clearer pattern. So, reflecting on an intention and carrying out the intention might happen in two different places. These conclusions appear to place this study at level 3, possibly even level 4.

Does this result help us discriminate between any functional-level accounts of cognition? It does tell us that an intention like “I am going to do an addition” is encoded in such a way that it causes changes across a large area of neural circuitry, rather than just in one small patch below the resolution of the scanner. After all, the intention could have been encoded in just a handful of neurons in the prefrontal cortex, with the rest doing unrelated processing, so that the difference would have been undetectable.

Notice, however, that this study, like many of the others, gives us information that seems to be locked in at the neural level alone, without coming up to the functional level and telling us something about how the mechanism of “intending to do an action” actually works. Both empirical conclusions—about the distributed spatial pattern and the change of location between intention and execution—are just giving us different kinds of location data without telling us what kind of mechanism is operating and how it is doing its work.

This study is straightforwardly consistent with our molecular framework. The spatially distributed pattern of atoms that encode the intention to perform an action would be very similar each time that same intention occurred (for the same reason that, earlier, we argued that the Jennifer Aniston atom would likely appear in the same place each time it was activated). If a particular spatial distribution of atoms gave rise to a particular spatial distribution of brain activity (not a foregone conclusion, but quite plausible nonetheless), then a brain scan could distinguish the patterns resulting from two different intentions.

Study 6: "Scientists Discover Brain Trigger for Selfish Behavior"

This study is somewhat different from the others. Knoch, Pascual-Leone, Meyer, Treyer, and Fehr (2006) used low-frequency repetitive transcranial magnetic stimulation (rTMS) to disrupt the dorsolateral prefrontal cortex (DLPFC) in subjects who were trying to play something called the "Ultimatum Game."

This game is a test of the subject's willingness to make a tradeoff between accepting an unfair offer of money (it is unfair because the person making the offer will get a bigger cut than the subject) and rejecting the offer (in which case neither person will get anything). If the unfair offer is accepted, then this indicates that the selfish motive of just taking the money is the one that dominates. If the offer is rejected, this would show that the person has given greater weight to the need to maintain reciprocity—the social custom of rewarding fairness and showing disapproval of unfairness.

The researchers had reason to believe that the DLPFC was involved in the decision-making process here, but there was a question about whether (a) an impulse to reject the unfair offer was coming from somewhere, and the role of the DLPFC was to control that impulse, or (b) an impulse to be selfish and accept the money was coming from elsewhere, with the DLPFC moderating that impulse. The way to decide, according to Knoch et al., was to disrupt the DLPFC during the decision making, and see what happened. More acceptances of unfair offers would imply that this region had previously been acting as a brake on selfish impulses, but if the acceptance rate dropped, this would indicate that the usual role of the DLPFC was to moderate the unfairness motive.

The experimental results indicated that the right DLPFC (but not the left) was involved in suppressing selfish impulses, because the acceptance rate went up when it was disrupted. Subjects still said that they judged the offers to be unfair, when asked, but they felt less inclined to reject them. We consider this study to be concerned with locating where in the brain functions happen, and is therefore level 2.

Observations

Two observations can be made about this experiment. First, the role of the DLPFC might not be to act as a "gate" on the signals coming from the source of selfish impulses: There may be a separate structure that adds up the motives coming from various sources, with the DLPFC sending a vote for reciprocity and another structure sending a vote for selfishness. This kind of architecture would look very different from one in which the DLPFC was specifically designed to gate the signals coming from a selfishness module.

The second observation is about whether the DLPFC is specialized to do the job of enforcing fairness (as Knoch et al. imply), or whether it might simply be part of a mechanism for considering complex motivational issues.

The easiest way to see how this could be so is to go back to the molecular framework again. In the part of the foreground periphery that we have called the "motivation" area substantial numbers of atoms are building representations of the system's goals, drives, and desires. In times of simplicity (I am hungry, there is a cream puff in front of me, and it's mine), there is not much complexity in the structures hanging around the motivation area. But when difficult decisions have to be made, as in the tradeoffs of the Ultimatum Game, a good deal of activity may occur, during which large complexes of atoms must represent complex, abstract ideas and decide between rival impulses coming from outside the foreground. All this activity takes up space in the foreground, so these complex decisions might require larger amounts of cortical real estate.

Now consider one more feature that might be built into the design of the foreground: As far as possible, it needs to be robust against dithering. It must have default plans ready to go if more complex decision-making fails. So, in trying to make a decision to follow one impulse or another, the foreground probably builds representations for several different options, in parallel.

In the present case, the option to obey a selfish impulse is fairly simple, not involving much thought or emotion, so the atoms representing the "take the money" plan may be quite compact and easily assembled. But the processing of fairness and reciprocity concepts is likely to be more extensive, and it may also trigger some strong emotions that trigger yet more action plans. This combination of abstractness and a strong cluster of emotional responses could mean that a larger amount of the foreground needs to be taken up by the processing of the impulses coming from an "unfairness" signal.

If the role of the DLPFC is to accommodate large clusters of atoms involved in difficult decisions, and if simple, default decisions (like just going with the selfish motive) are handled elsewhere, then a disruption of the DLPFC might cause the foreground to go for the simple, selfish option; not because the DLPFC was specialized for fairness, but because its job was to act as an overspill area for complex motivational decisions.

This idea could be consistent with the observation that subjects still say the offer is unfair even when, with their DLPFC disrupted, they decide not to do anything about it. This would happen because the DLPFC is located near the motivational area of the foreground periphery, and is primarily involved in hosting atoms that deal with the flood of signals coming from the "drives"—the cluster of lower-level brain mechanisms that push the foreground to attend to different priorities such as food, comfort, sex, stimulation, and threat. The abstract representation of the idea [this offer is unfair] will take place in the main part of the foreground, where the recognition of other abstract objects occurs. But having abstract knowledge of unfairness is not the same as using that knowledge as an ingredient in the complicated process of weighing the relative merits of different drives, and it is the latter process that the DLPFC might be specifically responsible for. So the DLPFC would get the information that "this offer is unfair" from the main part of the foreground, but if it were disrupted it might not host the complex set of motivation-related atoms triggered by the "unfairness" signal coming from below, and so the default "selfishness" signal wins the day.

Summary

This study is a good illustration of how the conclusion of a brain imaging experiment can depend on subtle aspects of the theoretical mechanisms that the experiment is supposed to be localizing. A simple interpretation of the result, in this case, might lead us to believe that the DLPFC is responsible for implementing fairness or reciprocity behaviors. Our alternative molecular framework, however, might give the same structure the role of considering all kinds of complex problems related to the resolution of drives and motivations, not just fairness.

Discussion

In this analysis of the theoretical integrity of six randomly chosen, influential brain imaging experiments, we have found that in no case did the experimental conclusions give us new information about the structure or function of any mechanisms in the human cognitive system. Where we learned anything, we learned only that some known mechanisms are located in a particular place. In no case did a location-fact noncontentiously give us a new functional-fact. So, although at first sight three of these studies aspired to at least level 2, and three even to level 4, in fact we are forced to conclude that when we put them under the microscope they all end up at level 2—localization studies. Even then, we have argued that, because there are alternative accounts of what is happening during the cognitive processing involved in these tasks, we are not sure what exactly is being localized in these brain regions. The strongest conclusion that can be justified in each of the studies appears to be that brain region X "has something to do with" cognitive function Y. There is no reason to assume that these particularly famous studies are other than representative of the entire field.

In the Quiroga et al. experiment (study 3) we also found that the conclusion about sparse encoding in neurons was simply not theoretically coherent. Our alternative "molecular" framework could explain the results of that experiment, but the explanation was orthogonal to the one the authors gave.

In two cases where the authors made strong claims about the location of a functionally defined mechanism—study 1 (claiming that the pLPFC is the location of the dual-task performance bottleneck) and study 6 (claiming that the right DLPFC is specialized for enforcing the fairness motive)—we were able to nullify the reported conclusion by shifting to our molecular framework. In another case, study 4, we were able to at least reduce the clarity of the conclusion by giving a molecular-framework interpretation of what was happening.

Overall, we believe these six studies showed an alarming sensitivity to the theories in cognitive psychology that generated the mechanisms these authors tried to locate. By changing the theoretical framework to one that was slightly out of the mainstream, we were able to show that the conclusions changed. Indeed, in those cases where the conclusions were not perturbed (studies 2 and 5), this may only have been because the claims were too dilute to be susceptible to attack.

If it was this easy for us to propose a framework that made some of the localization results seem less secure, then how vulnerable might these localization studies be to other, as yet unheard-of theoretical frameworks?

Looking Back

Let us return to the questions we raised in the introduction and try to fill in some answers:

- Is where-it-happens information of any use to present-day psychology? Not at the moment, because our cognitive models are insufficiently specified.
- Are the claims about neural localization internally and theoretically coherent? Not always: In one of the six cases we examined, the claim was theoretically incoherent.
- Do the localization claims refer to components of the mind that are clearly defined? No. Many of our current models we consider too constrained by the language of description, so such lay labels as "attention" are sometimes treated as if they correspond in a simple way to cognitive mechanisms. For this reason, we advocate computational modeling, where cognitive processes may have no simple correspondence with our intuitions and labels, as the way forward. Note, we are not saying that where-it-happens is never going to be interesting or useful!
- Are the inferences made by researchers theory-laden? A definite yes to this one.
- Are today's studies, which give us only crude localization data (both spatially and temporally), just a prelude to later research that will pin down location and function precisely, or is crude localization all we can ever expect from brain imaging technology? This question is an important one to which we return in the following sections.

By pointing out that our new framework often conflicts with these localization results, we are implying that more work needs to be done, somewhere. But in any of the cases where our framework raises new questions, should those questions be addressed by more studies of the localization of function? Is the solution really just more brain imaging?

The Stages of Cognitive Science

Cognitive science is destined to go through three phases in its history. In phase 1 we do our best to produce metaphorlike descriptions of functional-level mechanisms. The language we use to articulate theories at this level will contain descriptions of things that sometimes look as if they could be mechanisms at the implementational level, but this is often an illusion.

In the future (and perhaps starting already) we would hope to move toward a complete outline theory of the human cognitive system. At this stage, we would expect the basic processes and structures to be clear enough that no drastic changes would be arriving to disrupt the outline theory in the future. This then would be phase 2; but this stage would still be only a functional-level description of the mind.

In phase 3, we would commit to how the complete functional-level theory was implemented in the particular neural hardware we find in our brains. Instead of just completely describing how the "atoms" of our framework interacted with one another to give rise to all known psychological data, for example, we would go on to say how those atoms were implemented in specific neural circuits. These three stages are not expected to be completely separate, of course, but we nevertheless believe that extensive phase 3 work is not very useful or appropriate when we are still struggling to move from phase 1 to phase 2.

Are all studies of the brain a waste of time? Certainly not, but a great deal hinges on the granularity of the information being gathered. If today's brain imaging studies were just a warm-up for new types of investigation that promise to yield detailed circuit diagrams and real-time behavior of large networks of human neurons, with such things as precise tracking of synapse strengths and dendritic tree layouts, then we could perhaps see how today's crude localization studies might be laying the groundwork for future scientific cornucopias.

But nothing remotely like such a level of neural detail is on the horizon, and so we are in a bind. On the one hand, the resolution of these brain imaging studies is not enough to tell us useful things about the functional level, and future improvements in the technology do not appear to offer the granularity of information that we need. On the other hand, the level of specificity of the cognitive theories is currently not good enough to make coarse-grained localization theories useful.

The Way Forward

Would it be reasonable for someone working in imaging to say "so we're in this pickle because of psychology—how are they going to get us out of it?" We don't think that psychology, in the sense of being just an experimental science, can solve the problem by itself. In all the preceding cases, where our molecular framework gives an idiosyncratic view of what might be happening at the functional level, there are many questions we could ask about what might be going on, but the best way to answer those questions would be to increase the sophistication of our computer simulations of the functional-level mechanisms (Loosemore 2007), and to perform human behavioral experiments to test the predictions those simulations made. We can see how answers to these kinds of questions would advance our understanding of psychology immeasurably. But right now, we are being flooded with accurate answers to questions about the brain location of mechanisms that we do not believe in and inaccurate answers to questions about the brain location of mechanisms that are currently not terribly interesting. This state of affairs seems to us to be a great leap backwards.

References

Aron, A., Fisher, H., Mashek, D. J., Strong, G., Li, H., and Brown, L. L. (2005). Reward, motivation, and emotion systems associated with early-stage intense romantic love. Journal of Neurophysiology, 94, 327-337.

Baars, B.J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. Prog. Brain Research, 150, 45–53.

Bahrami, B., Lavie, N. and Rees, G. (2007). Attentional load modulates responses of human primary visual cortex to invisible stimuli. Current Biology, 17, 509-513.

Barlow, H. (1972). Single units and sensation: a neuron doctrine for perception. Perception, 1, 371-394.

Block, N. (1996). How can we find the neural correlate of consciousness? Trends in Neurosciences. 19, 456–459.

Dux, P. E., Ivanoff, J. G., Asplund, C. L., & Marois, R. (2006). Isolation of a central bottleneck of information processing with time-resolved fMRI. Neuron, 52, 1109-1120.

Harley, T.A. (1995). The psychology of language (1st ed.). Hove: Psychology Press.

Harley, T. A. (1998). The semantic deficit in dementia: Connectionist approaches to what goes wrong in picture naming. Aphasiology, 12, 299-308.

Harley, T.A. (2004a). Does cognitive neuropsychology have a future? Cognitive Neuropsychology, 21, 3-16.

Harley, T.A. (2004b). Promises, promises. Reply to commentators. Cognitive Neuropsychology, 21, 51-56.

Harley, T.A. (2007). The psychology of language (3rd ed.). Hove: Psychology Press

Haynes, J-D., Sakai, K., Rees, G. Gilbert, S., Frith, C. & Passingham, E. (2007). Reading hidden intentions in the human brain. Current Biology, 17, 1-6.

Hofstadter, D. R. (1995). The architecture of Jumbo. In D. R. Hofstadter, Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought. New York: Basic Books.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V. & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science, 314, 829-832.

Lamme, V.A. (2003). Why visual attention and awareness are different. Trends in Cognitive Science, 7, 12–18.

Loosemore, R.P.W. (2007). Complex Systems, Artificial Intelligence and Theoretical Psychology. In B. Goertzel & P. Wang, Proceedings of the 2006 AGI Workshop. Amsterdam: IOS Press.

Mandler, G. (2005). The consciousness continuum: From ''qualia'' to ''free will.'' Psychological Research, 69, 330–337.

McClelland, J.L., Rumelhart, D.E. & Hinton, G.E. (1986). The appeal of parallel distributed processing. In D.E. Rumelhart, J.L. McClelland & G.E. Hinton and the PDP Research Group, Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1. MIT Press: Cambridge, MA.

Ng, M., Ciaramitaro, V. M., Anstis, S., Boynton, G. M. & Fine, I. (2006). Selectivity for the configural cues that identify the gender, ethnicity, and identity of faces in human cortex. Proceedings National Academy of Science USA, 103, 19552-7.

Owen, A. M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S., Pickard, J.D. (2006). Detecting awareness in the vegetative state. Science, 313, 1402.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. (2005). Invariant visual representation by single-neurons in the human brain. Nature, 435, 1102-1107.

Tobler, P., Fletcher, P., Bullmore, E. & Schultz, W. (2007). Learning-related human brain activations reflecting individual finances. Neuron, 54, 167-175.