CHAPTER 9

Limits and Cautions

fMRI is not and will never be a mind reader, as some of the proponents of decoding-based methods suggest, nor is it a worthless and non-informative "neophrenology" that is condemned to fail, as has been occasionally argued.

—Nikos Logothetis[†]

Understanding the brain by looking at images of it strikes me as akin to deciphering a computer by where the casing feels hot. You may locate the hard drive, but you'll never know what the 1s and 0s are.

—Anonymous comment[††]

CHAPTER SUMMARY

This chapter:

- Introduces basic limitations of noninvasive measures of brain function.
- Highlights several principles that can help effectively evaluate and critique functional brain imaging evidence.
- Presents critical perspectives on exaggerated claims about neuroscience.

INTRODUCTION

You have learned about the major noninvasive techniques available to measure human brain function—fMRI and EEG (with QEEG). These methods yield dazzling visual images that are informative to trained scientists and can appear conclusive to laypeople. But having read the previous chapter, you now also know that neuroimages are the product of a multiple-step process involving many statistical and graphical choices.

Because the end result is an image that researchers can use to illustrate patterns of brain activation associated with complex and uniquely human abilities, it is no wonder that we are amazed. When we see a colorful brain image purporting to be ''our brain in love,'' we take a second look. And when we see headlines suggesting that the government can use neuroscience technologies to read our minds and harvest our dreams, we shudder. But are our hopes and our fears about neuroscience justified?

† Nikos Logothetis, *What We Can and Cannot Do with fMRI*, 453 Nature 869 (2008).

†† Excerpted from a comment posted on *wired.com* by ''Docbmac,'' 2008.

This chapter goes beyond the colorful images, and beneath the sensational headlines, to carefully examine what brain imaging neuroscience can — and *cannot* — do for law. The Appendix provides specific details beyond those summarized in this chapter.

A particularly vivid reminder of the need for caution comes from an Atlantic salmon. In 2010 a group of neuroscientists published a paper in which they reported on their measurement of fMRI brain activation in a dead Atlantic salmon. The researchers treated the salmon in the scanner the same way they treated their other human subjects, putting the salmon through a task that asked participants to adopt the point of view of another person shown on the screen. Craig Bennett et al., *Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument for Proper Multiple Comparisons Correction*, 1 J. Serendipitous & Unexpected Results 1 (2010).

But standard analysis of the data found several small regions displaying increased activity during periods when the expired salmon was exposed to pictures of humans in different social-emotional situations (compared to interspersed periods without pictures). The recorded activity was due to ''noise'' in the fMRI signal, and the noise achieved statistical significance only by happenstance.

When modest statistical corrections were applied to quiet this noise and thus limit the presence of false positives (a statistical false alarm that will be discussed in more detail later in this chapter), the differential activity was no longer significant. In other words, when the threshold for statistical significance was appropriately elevated, the dead salmon's brain no longer showed any significant regions of increased activation.

The authors of the study wrote, ''Can we conclude from this data that the salmon is engaging in this perspective-taking task? Certainly not. What we can determine is that random noise in [fMRI data] may yield spurious results if multiple comparisons are not controlled for.''

For the practitioner and consumer of neurolaw, the salmon experiment — though somewhat tongue-in-cheek — illustrates an important point: Brain imaging data must be carefully acquired, analyzed, and interpreted to be meaningful. The salmon story is also, in some ways, an easy case. Dead salmon don't think, so we know that any reported activity associated with the task was not really there. But in typical research experiments or clinical evaluations, we do not have such strong priors. Indeed, the very reason the testing is being undertaken is typically to ascertain whether there is a real difference between two conditions or whether a certain part of the brain is active in a particular way. How, then, can one have confidence in reported findings and proffered brain evidence?

The goal of this chapter is not to diminish the importance or validity of noninvasive brain measurements; indeed, to do so would render a coursebook about law and neuroscience curiously irrelevant. Instead, the goal of this chapter is to inoculate you from being overwhelmed or overawed by evidence provided by these techniques. Although glaring cautionary tales like that of the dead salmon might suggest that noninvasive measurements of brain function are fatally unreliable, the outlook for neuroscientific evidence is not so dire. Let us repeat, this chapter throws a large bucket of cold water on the *uncritical* acceptance of information from noninvasive measures of brain function. But these measures — carefully analyzed and properly understood — can be incredibly informative and

useful. And they will almost certainly increase in their application and their accuracy.

The way forward, then, is not to discard fMRI and EEG measures of brain function, but rather to shape their effective and judicious incorporation into the legal community with knowledge of the limitations of such methods.

The chapter proceeds in three sections. Section A surveys the general and unavoidable limitations and appropriate cautions in the interpretation of fMRI and EEG findings. Section B surveys in more detail particular measurement limitations of fMRI and EEG with QEEG. Section C explores concerns about overreaching claims about what neuroscience can do for law and for society more broadly.

A. GENERAL LIMITS AND CAUTIONS

1. Functional Magnetic Resonance Imaging

Owen D. Jones et al.

Brain Imaging for Legal Thinkers: A Guide for the Perplexed

2009 Stan. Tech. L. Rev. 5

. . . We will now discuss systematically key concepts about brain imaging that legal thinkers should know:

1. Anatomical imaging and functional imaging are different.

Two anatomical images, taken one minute apart, will ordinarily look indistinguishable. Yet two functional images, from data collected one minute apart, will look completely different. This is because brain structure changes slowly while brain activity changes rapidly. Another reason is because fMRI brain images are built statistically, not recorded photographically. In the typical fMRI case, hundreds of samples are made of each voxel in the brain, at slightly different times (e.g., every two seconds). Each recording of each voxel within a given trial is analogous to a single frame in a movie. Learning what happens within each voxel, over time, is akin to watching motion seem to emerge from the observation of successive snapshots that comprise a moving picture. But that metaphor only captures part of the fMRI technique, because there are subsequently many repeated samples of that voxel, under similar conditions, on many consecutive trials — the results of which are typically then averaged across trials. Complicating matters further is that there are about one hundred thousand voxels within the brain, and what typically matters is how neural activity within those voxels is varying over time, in relation to some task the subject(s) undertake while being scanned. Furthermore, within each voxel are millions of neurons of different types, interacting in ways that could be mechanistically different but indistinguishable from the measure of fMRI. In the end, fMRI brain images lay the result of any one of many possible statistical tests overtop of an anatomical image of a selected slice of the brain. That is, an fMRI image is a composite of an anatomical image, of the researcher's choosing, and a statistical representation of the brain activity in that image, also of the researcher's choosing.

2. Functional brain imaging is not mind reading.

There is more to a thought than blood flow and oxygen. fMRI is very good at discovering where brain tissue is active (commonly by highlighting differences between brain activations during different cognitive tasks). But these differences are not thoughts. fMRI can show differences in brain activation across locations, across time, and across tasks. But that often does not enable any reliable conclusion about precisely what a person is thinking.

3. Scanners don't create fMRI brain images; people create fMRI brain images.

Images are only as good as the manner in which the researcher designed the specific task or experiment, deployed the machine, collected the data, analyzed the results, and generated the images. It is important to remember that fMRI images are the result of a process about a process. Multiple choices and multiple steps go into determining exactly what data will be collected, as well as how and when it will be collected. More choices are made about how the data will be analyzed and how it will be presented.

4. Group-averaged and individual brain images are importantly different.

Most brain imaging research is directed toward understanding how the average brain, within a subject population, is activated during different tasks. This is not the same as saying either that all brains performing the same task activate in the average way, or saying that the activation of a single brain can tell us anything meaningful about the appearance of the average brain. Consequently:

- Do not assume that the scan of any individual is necessarily representative of any group.
- Do not assume that the averaged scan of any group will resemble any individual.

5. There is no inherent meaning to the color on an fMRI brain image.

fMRI does not detect colors in the brain. fMRI images use colors — from whatever segment of the rainbow the researcher prefers — to signify the result of a statistical test. By convention, the more intense the color (say, yellow compared with orange or light blue to dark blue) the greater the statistical significance of the differences in brain activity between two conditions. Put another way, the brighter the color, the less likely it is that the differences in brain activity in that voxel or region, between two different cognitive tasks, was due to chance alone. As with any color-coded representation, accurate interpretation requires knowing exactly what each color represents in absolute terms. It is important to appreciate that brain activation in one condition can be either greater than or less than brain activation in another condition. Typically, researchers use warm to hot colors to signify greater activation and cool to cold colors to signify less activation. Furthermore, the researcher specifies arbitrarily how certain colors map onto certain levels of statistical significance. For one researcher yellow might mean that there is only one chance in one thousand that the difference between brain activations in this voxel, between condition, is due to random chance. For another researcher yellow might mean that there is one chance in twenty that the difference is due to random chance. Different researchers have different approaches and perspectives on these statistical tests.

6. fMRI brain images do not speak for themselves.

No fMRI brain image has automatic, self-evident significance. Even well-designed, well-executed, properly analyzed, properly generated images must be interpreted in the proper context.

7. Classification of an anatomical or functional feature of the brain as normal or abnormal is not simple.

Because we have learned a great deal about the brain, from dissection, imaging, and the like, we have some confidence about what a typical brain looks like, and how a typical brain functions. But even without full anatomical scans of everyone on the planet, we know there is considerable variation — both anatomically and functionally. That means that it can be (with some exceptions, such as a bullet lodged in the brain) difficult to say with precision how uncommon a given feature or functional pattern may be, even if it appears to be atypical. Base rates for anatomical or functional conditions are often unknown. For example: suppose brain images show that a defendant has an abnormal brain feature. We often do not have any idea how many people with nearly identical abnormalities do not behave as the defendant did.

8. Even when an atypical feature of function is identified, understanding its effect on behavior is complex and uncertain.

Brain images can show unique features and functions of a person's brain. But their meaning is rarely self-evident. Determining which of those are important, and how, depends not only on the legal context for which the images are offered, but also on expert analysis of what the images do and do not mean. For example, suppose that measurement of the fMRI signal during a given cognitive task indicates that a person has less neural activity in a given region than does the average person. Does that mean that the person is somehow cognitively impaired in that region? Or might it alternatively indicate that the person has more expertise or experience than average, requiring less cognitive effort?

9. Correlation is (still) not causation.

The fact that two things vary in parallel tells us little about whether the two are necessarily causally related and, if so, which causes which. For example, suppose brain imaging reveals that seventy percent of inmates on death row for homicide have atypical brain activation in a given region, compared with normal, unincarcerated subjects. That statistic does not mean that the brain activation pattern causes homicidal behavior. It might mean that having murdered affects brain activations, or that being incarcerated for long periods of time affects brain activations, or something else entirely.

10. Today's brain is not yesterday's brain.

In all but the most fanciful of contexts, a brain scan likely takes place long after the behavior (such as criminal activity) that prompts the interest in the brain scan. Drawing causal inferences is therefore further complicated. People's brains change with age and experience. And some proportion of the population will develop atypical anatomical or functional conditions over time. If a defendant is scanned six months or six years after the act in question, and the scan detects an

abnormality, it is not a simple matter to conclude with confidence that the same abnormality was present at the time in question or — even if one assumes so, arguendo — that it would have meaningfully affected behavior.

11. Scanners detect what they are built and programmed to detect.

Scanners are highly complex and individualized pieces of machinery. So (as in other areas of science) are the people who calibrate, program, operate, and interpret the data. It is important to recognize that the product of these intersecting complexities may or may not be reliable, generalizable, and replicable.

12. Inferences about mental states can be supported only indirectly by fMRI brain imaging information.

It is important to remember that fMRI does not provide a direct measure of neuronal activity — as do, for example, invasive techniques that measure the discharges of neurons in the brain. fMRI detects fluctuations in oxygen concentrations thought to be reliably associated with neuronal activity. But the precise relationship between metabolic demands and neuronal function remains poorly understood. Even if regional activations in brain images reflect true neural activity, it should also be kept in mind that our ability to confidently infer the cognitive process that must have led to such regional activation is very limited. This is because neuroscientists still understand so little about what the various regions of the human brain contribute to a particular cognitive function.

---

2. Electroencephalography

Many of the challenges inherent in fMRI data acquisition and analysis exist for EEG and QEEG as well. As you recall, EEG measures cumulative electrical potential on the surface of the head instead of blood oxygenation within the brain. Thus, it is a more direct measure of brain function that does not suffer from the seconds-long time lag of the fMRI blood oxygen signal. The magnitude of the EEG signal, referred to as *polarization*, varies with mental state such as being asleep or alert and in association with responses to stimuli in experimental testing. To facilitate comprehension, we will parallel the previous excerpt as we review key concepts to know about EEG.

1. The anatomical sources of EEG signals are difficult to specify.

EEG does not provide information about brain structure, but the structure of the brain is crucially important for the nature of the EEG signal. The EEG measured on the surface of the head originates in the electrical potentials associated with synaptic processes changing the state of pyramidal neurons in the cerebral cortex. The magnitude of the electrical potential measured on the surface varies with the angle of the neurons in the sulci and gyri of the cortex. The EEG measured on the surface of the head originates from the electrical currents flowing in large areas of the cerebral cortex. However, it is impossible to calculate exactly where in the brain these currents are located. In spite of this, various mathematical approaches and physical and biological intuitions permit some estimates of the source location of EEG signals.

2. EEG measurements are not mind reading, but they do relate to mental states.

Scientists have identified reliable associations between particular mental operations and the magnitude and timecourse of EEG. The frequency content and magnitude of the EEG varies systematically with level of consciousness, being higher frequency and lower magnitude during alert states and lower frequency and higher magnitude during sleep. The form of the EEG is violently dramatic during epileptic seizures. An EEG with no variation of electrical potential is a useful sign of brain death.

In psychological testing conditions, the EEG can be averaged over many testing trials and synchronized on different testing events, such as presentation of a stimulus or initiation of a body movement. These averaged EEG are called *event-related potentials*. For example, the magnitude of average EEG around 200 ms after a visual stimulus is presented is different if the participant is paying attention to that stimulus as compared to simply viewing it. Around 300 ms after a stimulus is presented the magnitude of the EEG is modulated by how unexpected a stimulus is and how much you notice and will remember about it. Preceding actions the EEG slowly increases in polarization until the body movement is made whereupon the EEG changes back to a neutral state. Finally, the magnitude of the average EEG after an action varies according to the consequences of the action. Besides modulating at different times in association with different mental operations, the various event-related potential components vary in magnitude over the surface of the head. Some ERP components have larger magnitudes over the occipital lobe, for instance, and others have larger magnitudes over the frontal lobes. None of these event-related potentials can be detected without averaging over dozens if not hundreds of testing trials. In legally relevant settings it is difficult to obtain such extensive data.

QEEG decomposes EEG into different frequency bands, and maps of the various frequency bands are constructed over the surface of the head. Think of the EEG as a musical chord played on a piano and the QEEG as a measure of which keys (frequencies) were struck with what strength. Some researchers have identified associations between patterns of EEG frequency bands across the head and particular mental states. However, a particular pattern of EEG frequency bands cannot be taken as a certain sign that a particular mental state is occurring.

3. Amplifiers don't create EEG brain images; people create EEG brain images. Consequently, there is no inherent meaning to the color of an EEG map.

One of the alluring aspects of QEEG is that it allows one to produce a color map of the EEG frequency bands across the head. When presented in graphic color, these maps seem to bear some relation to the patterns of activation displayed after fMRI measurements. As noted above, though, the colors used and how they are made to gradually change in these plots is up to the investigator to specify. Accordingly, small absolute differences in the magnitude of EEG frequency bands can be made to appear much larger through the use of colors that we resolve as distinctly different.

4. Group-averaged and individual data are different.

The general timing and spatial pattern of EEG signals associated with different mental states is fairly consistent across individuals. This regularity is the basis of the scientific studies that are performed. However, because the EEG originates

primarily from the pyramidal neurons aligned in the cerebral cortex and because the precise folding pattern of each person's cerebral cortex is different, the timing and magnitude of the EEG measured on each person will have variation. The magnitude of this natural variation has not been quantified systematically.

Further variation in EEG signals can occur when measurements are obtained across days. Some of this variation arises from idiosyncrasies in how the electrodes are placed on the scalp. Some of the variation can also arise from differences in alertness and engagement of the participant across testing sessions.

5. EEG and QEEG maps do not speak for themselves.

As noted above for images derived from fMRI measurements, the maps of EEG magnitude or frequency content cannot be interpreted without the context of the state of the participant (e.g., awake, asleep, dreaming) and what the participant was doing (e.g., simply resting or performing a particular task).

6. Classification of EEG signals as normal or abnormal is not simple.

As noted above for fMRI measurements, the magnitude of variation of EEG and QEEG signal patterns varies across individuals. Some studies have quantified EEG and QEEG measures across healthy normal individuals over time and found less variation across repeated measurements from one individual than across individuals. However, these studies have not thoroughly quantified the nature and magnitude of variation in EEG and QEEG measures across gender, race, and age. Numerous studies have characterized EEG and QEEG differences between individuals with various neurological and psychiatric disorders and healthy counterparts. In most cases, a measure may be significantly different on average between the groups, but the range of variation of that measure overlaps between the groups. Therefore, unless an individual's particular measure is extremely high or low, it can be impossible to specify reliably in which group the individual belongs. In some conditions, like certain types of epilepsy, the EEG signatures are very distinct, but in most conditions, like schizophrenia and depression, the EEG differences are very subtle. Therefore, a conclusion that a particular individual's EEG pattern is meaningfully abnormal must be considered with caution and skepticism.

7. Even when an atypical EEG signal is identified, its relationship to mental states and behavior is uncertain.

As described in Chapter 8 and above, particular measures of EEG have been associated with particular mental states or process such as attention, working memory, volition and sensitivity to consequences. However, scientific findings do not support strong inverse conclusions about disordered mental states from abnormal EEG signals. Furthermore, inferences about relationships between EEG measures and mental states cannot be considered reliable unless suitably controlled behavioral testing procedures are used to independently manipulate or probe the given mental state or process. Scientific studies collect data with a controlled set of testing conditions, e.g., different kinds of stimuli or responses, measured in numerous testing trials to obtain reliable average measurements. As you will see in the chapter on lie detection, the use of an event-related potential measure for lie detection is questionable when the data are collected with too few testing trials with too few and uncontrolled stimuli. In short, without concomitant psychological

testing, EEG measurements can support no strong conclusions about mental states and processes. This is especially problematic for QEEG measurements obtained when individuals are doing nothing in particular.

8. Correlation is (still) not causation.

As noted above for fMRI, the association between abnormal EEG and abnormal behavior may not be one directional. The abnormal EEG could be a consequence of a long period of abnormal behavior.

9. Today's brain is not yesterday's brain.

As noted above for fMRI, brain scans almost never happen during commission of a crime. Even if scientists identify EEG signatures of *mens rea*, it is very unlikely that the brain state can be measured during the planning and commission of the crime.

10. EEG systems measure what they are built to measure.

As noted, the EEG signal bears a more direct relationship to brain function than the blood oxygen signal of fMRI. However, because the measurement is obtained noninvasively, it remains indirect. EEG systems vary in certain ways that may or may not be relevant for a particular case. Some systems sample EEG signals from just a dozen or so locations on the head, while other systems sample EEG signals from 64, 128 or even 256 locations covering the head. More electrodes generally provide better information about potential sources within the brain responsible for a given EEG signal. However, to assess the presence, timing and magnitude of certain mental states and processes, fewer than a dozen electrodes and perhaps even just a couple are needed.

NOTES AND QUESTIONS

1. Statistical analyses of noisy and variable measurements (like measurements of brain and behavior) have demonstrated two fundamental types of errors that occur. The first, called a *false positive* (also called Type I error), occurs if a statistically significant experimental effect is identified when in fact none actually exists. The second, called a *false negative* (Type II error), occurs when a true effect is not identified. To understand these errors, consider a household smoke detector. Smoke detectors are designed to be quite sensitive, so they can be set off by very small amounts of smoke, such as burnt toast. Assuming that the goal of a smoke detector is to sound alarms for serious house fires, and not minor annoyances like burnt toast, then an alarm triggered by the toast is a false alarm because in fact the house was not on fire. Now imagine a careless homeowner who has unknowingly allowed the batteries in their smoke detector to expire. In this case the alarm would fail to sound even if a significant house fire is burning near the detector. This failure to signal a fire would be a false negative. What legal implications do you see for these two types of errors?
2. Statisticians distinguish *sensitivity* and *specificity*. Sensitivity measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having some illness). The fewer false negatives, the higher the sensitivity. Specificity measures the

proportion of negatives that are correctly identified (e.g., the percentage of healthy people who are correctly identified as not having the condition). The fewer false positives, the higher the specificity. In real world settings like medical diagnosis or airport security, 100 percent sensitivity (i.e., predict all people from the sick group as sick) with 100 percent specificity (i.e., predict no one from the healthy group as sick) is impossible. Consequently, a risk-benefit analysis is undertaken and a tradeoff is made by enacting policies that set categorization thresholds higher or lower. For example, if one wishes to avoid any threat to safety, airport screening scanners can be adjusted to trigger on low-risk items like belt buckles and keys (low specificity) so as to reduce the probability of missing high-risk items like knives and guns (high sensitivity). What risk-benefit policy do you think should govern application of fMRI and EEG measures in criminal law? What about personal injury claims? Social security disability? National security?

B. METHODOLOGICAL AND ANALYTICAL LIMITATIONS OF FMRI

1. Limitations in Measurement

Functional brain imaging with fMRI has proven and will continue to be a very informative tool for research and many applications. However, you must appreciate the various limitations of the technology.[1]

The main advantages of fMRI include noninvasiveness, relatively good spatial and temporal resolution, and the opportunity to obtain a complete view of the entire brain. But as neuroscientist Nikos Logothetis explains that fMRI has profound limitations that are not due to physics or engineering; they are unlikely to be resolved by increasing the sophistication and power of the scanners. These more basic limitations arise from the indirect and coarse nature of the fMRI signal relative to the very dense and precise functional organization of the brain, and also to inappropriate experimental protocols that ignore this organization. Thus, although fMRI cannot be a mind reader, it is a very useful technology that can be used with other complementary technologies to investigate brain function.

The main disadvantage is that fMRI measures a surrogate signal derived from the simultaneous activity of many thousands or tens of thousands of neurons, and the spatial and temporal specificity of this signal is severely limited by physical and biological constraints. Typical voxel sizes have been $3 \times 3 \times 3$ mm, but improved technology is allowing voxels that are fractions of a millimeter ($<0.5 \times 0.5 \times 0.5$ mm). Even so, fMRI voxels contain very many neurons and glia and blood vessels and blood cells. One cubic mm in the human cerebral cortex contains ~25,000 neurons with around 10^9 synapses, 4 km of axons and 0.4 km of dendrites with less than 3 percent of this volume is occupied by blood vessels. This means that a typical fMRI voxel of 55 mm^3 contains ~6 million neurons, ~4×10^{10} synapses, 22 km of dendrites and 220 km of axons. Some authors have compared fMRI to

1. These few paragraphs summarize key points from the article: Nikos Logothetis, *What We Can and Cannot Do with fMRI*, 453 Nature 869 (2008).

measurements of the heat coming from different locations on a computer; perhaps you could deduce something about the operation of the computer but without monitoring the individual circuit elements, you could not learn just how the computer works.

As detailed in Chapter 8, fMRI studies commonly use two basic experimental designs. Logothetis emphasizes a basic limitation of the logic of the *block design*. Recall that in a block design experiment the pattern and magnitude of the BOLD signal during a task condition (e.g., generating verbs associated with a list of nouns) is compared to that during a control condition (e.g., just reading the nouns). The BOLD activation in each voxel during the test condition is subtracted from that in the control condition, revealing voxels with enhanced or reduced BOLD signal. Logothetis notes that a necessary assumption underlying this approach is that a single cognitive process can be inserted into a task without affecting the remainder. Psychologists have concluded many years ago that this assumption is generally unjustified because the human mind does not have discrete modules and distinct operations. He also explains that the brain consists of networks of neurons subserving different functions but overlapping within voxels, so the fMRI signal cannot resolve the subtle differences at the cellular level.

Logothetis also notes the complexity to be considered before drawing conclusions that activation of some brain region means that it is truly involved in the task at hand. Much of the complexity involves biological and physical details that need not concern us here. You should understand this, though. Observing an increase in the BOLD signal in a voxel does not guarantee that the neurons with axons exiting that voxel are necessarily generating more nerve impulses. The BOLD signal is a signature of the metabolic consumption in a part of the brain, and that consumption increases when excitatory neurons produce more impulses but it also increases when inhibitory neurons prevent the excitatory neurons from producing more impulses. Thus, when you read about ''fMRI activation'' in some brain region, remember this does not necessarily mean that more nerve impulses are being sent from that brain region to other parts of the brain.

NOTES AND QUESTIONS

1. Imagine you are given a laptop computer and a variety of devices that can measure things like heat, sound, and vibration. What noninvasive measurements could you make that would provide information about the operation of a computer? What would you be unable to learn about the operation of the computer with such measurements?
2. Imagine you have a device that remotely measures local temperature in a room full of people who are engaged in various conversations. The device is sensitive enough to detect the elevated temperature emitted from speakers as they exhale. What could you infer from the measurements using this device? What limitations of those inferences do you recognize?

2. Limitations of fMRI Analysis Procedures

The previous excerpt surveyed some of the technical and conceptual problems with fMRI. This section summarizes the problems inherent in the various fMRI analysis procedures. One of the most salient points to take away is that what is found in an fMRI study depends very much on how the data are analyzed.

a. The dangers of multiple analysis alternatives

Joshua Carp

On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments

6 Frontiers in Neuroscience 1 (2012)

How likely are published findings in the functional neuroimaging literature to be false? According to a recent mathematical model, the potential for false positives increases with the flexibility of analysis methods. Functional MRI (fMRI) experiments can be analyzed using a large number of commonly used tools, with little consensus on how, when, or whether to apply each one. This situation may lead to substantial variability in analysis outcomes. Thus, the present study sought to estimate the flexibility of neuroimaging analysis by submitting a single event-related fMRI experiment to a large number of unique analysis procedures. Ten analysis steps for which multiple strategies appear in the literature were identified, and two to four strategies were enumerated for each step. Considering all possible combinations of these strategies yielded 6,912 unique analysis pipelines. Activation maps from each pipeline were corrected for multiple comparisons using five thresholding approaches, yielding 34,560 significance maps. While some outcomes were relatively consistent across pipelines, others showed substantial methods-related variability in activation strength, location, and extent. Some analysis decisions contributed to this variability more than others, and different decisions were associated with distinct patterns of variability across the brain. Qualitative outcomes also varied with analysis parameters: many contrasts yielded significant activation under some pipelines but not others. Altogether, these results reveal considerable flexibility in the analysis of fMRI experiments. This observation, when combined with mathematical simulations linking analytic flexibility with elevated false positive rates, suggests that false positive results may be more prevalent than expected in the literature. This risk of inflated false positive rates may be mitigated by constraining the flexibility of analytic choices or by abstaining from selective analysis reporting.

How common are false positive results in the functional neuroimaging literature? Among functional MRI (fMRI) studies that apply statistical correction for multiple comparisons, most use a nominal false positive rate of 5%. However, [investigators] estimate that between 10 and 40% of fMRI activation results are false positives. Furthermore, recent empirical and mathematical modeling studies argue that the true incidence of false positives may far exceed the nominal rate in the broader scientific literature. Indeed, under certain conditions, research findings are more likely to be false than true.

. . . While some research outcomes were relatively stable across analysis pipelines, others varied widely from one pipeline to another. Given the extent of this variability, a motivated researcher determined to find significant activation in

practically any brain region will very likely succeed — as will another researcher determined to find null results in the same region. To mitigate the effects of this flexibility on the prevalence of false positive results, investigators should either determine analysis pipelines *a priori* or identify optimal pipelines using data-driven metrics. If researchers use multiple pipelines to analyze a single experiment, the results of all pipelines should be reported — including those that yielded unfavorable results. If implemented, these steps could significantly improve the reproducibility of research in the fMRI literature.

NOTES AND QUESTIONS

1. To better understand how conclusions from neuroimaging studies crucially depend on specific analyses used, consider a study published in 2007, in which researchers gathered data and then sent it out to nine different labs for analysis. The authors collected MRI data (static, not functional) from schizophrenic patients and non-schizophrenic controls using a 1.5T scanner. After some standard post-processing of the data, the authors transmitted the unanalyzed data sets to nine different labs for analysis. The labs were only told that the data came from two different groups and that they should use their typical MRI data analysis approach to distinguish the individuals in each group. Would you be surprised to learn that the findings from the different labs overlapped only minimally? This led the authors of the study to conclude that ''just because one method finds a particular difference, it does NOT mean that there were NO other differences — a fact that can be easily overlooked. This study therefore also highlights the extreme difficulty in interpreting differences in reported results obtained by different labs where there may be differences in the subjects recruited, or in the methods chosen to analyze the data, or indeed both.'' D.K. Jones et al., *What Happens When Nine Different Groups Analyze the Same DT-MRI Data Set Using Voxel-Based Methods?*, 15 Proc. Int'l Soc. Mag. Reson. Med. 74 (2007)
2. Some measurements of the human adult are roughly the same over time. Measuring the size of our hands and feet will produce roughly the same measure week-to-week and month-to-month. But researchers have found much more variation in an individual's brain activity pattern across different fMRI scanning sessions. That is, if you scan John doing a particular behavioral task in January, March, and May — or even on Monday, Tuesday, and Wednesday — you might not get the same activation patterns. It is not hard to see why this might be a problem for the types of legal uses of fMRI suggested in this book. What if the defendant's brain scan in January wasn't the same as his brain scan in July, or again in December? Is relying on results from a single scan — as almost all of the cases you will read in this book do — the equivalent of building a case on a lucky (or unlucky) draw from a deck of cards? In light of such possibilities, what should courts do? Keep in mind the high cost of conducting and analyzing data from an fMRI scan. For more on this issue, consider these publications that have compared the replicability of fMRI findings across time and laboratories:

    a. B.J. Casey et al., *Reproducibility of fMRI Results Across Four Institutions Using a Spatial Working Memory Task*, 8 Neuroimage 249 (1998). These investigators

compared fMRI (of that era) obtained at 4 institutions from healthy adults performing a standard task testing people's ability to remember briefly a location in space. They found common general regions of interest in dorsolateral prefrontal and posterior parietal cortex.

b. D.J. McGonigle et al., *Variability in fMRI: An Examination of Intersession Differences*, 11 Neuroimage 708 (2000). These investigators compared fMRI (of that era) obtained from one healthy adult in 99 sessions over 2 months consisting of 33 repetitions of simple motor, visual, and cognitive paradigms. They accounted statistically for the variability across sessions and testing conditions and found that the activation of many voxels was particular to sessions. The authors showed meaningful variability of activation across voxels both within sessions and between sessions and concluded that erroneous conclusions can be drawn from data obtained from a single session from a single subject.

c. S.M. Smith et al., *Variability in fMRI: A Re-Examination of Inter-Session Differences*, 24 Human Brain Mapping 248 (2005). With McGonigle as senior author, these investigators performed new analyses of the 2000 study data. They report that fMRI variability across sessions was of similar magnitude to that within sessions. They also show that the amount of variability observed across scanning sessions is affected by the methods of analysis. They advocate caution in the interpretation of data obtained in a single session.

d. B.B. Zandbelt et al., *Within-Subject Variation in Bold-fMRI Signal Changes Across Repeated Measurements: Quantification and Implications for Sample Size*, 42 NeuroImage 196 (2008). These investigators compared fMRI data collected from ten healthy subjects performing a simple motor task in three sessions, separated by one week. They found large variation in individual activation levels, in individual voxels and in regions of interest. Based on the magnitude of variation, the authors provide sample size estimations needed for repeated measurement studies.

b. The risk of false positives in fMRI data

The dead salmon study reminds us of the risks of statistical multiple comparisons. But even if appropriate corrections for multiple comparisons are applied, the massive quantities of data generated by a typical fMRI study can still create potential analytical pitfalls. One such problem, which was labeled ''voodoo correlations,'' generated a flurry of discussion among researchers using fMRI technology to study the neural substrates of human emotion and personality.

Edward Vul and colleagues (Edward Vul et al., *Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition*, 4 Persp. on Psychol. Sci. 274 (2009)) observed that several high-profile articles in the field of social neuroscience had reported very strong correlations between personality measures (based on self-report questionnaires) and brain activity as measured by fMRI.

Vul and colleagues, intrigued by the suspiciously high correlations, calculated the theoretical upper limit for correlations between personality measure and fMRI data. According to their calculations, many of the correlations reported in social

neuroscience articles exceeded this upper limit.[2] So how did these exceedingly high correlations arise? Vul and colleagues argued that it was the result of what they termed the *nonindependence error*, which occurred during the data analysis portion of the fMRI studies.

The nonindependence error is subtle even to scientists in the field and even more elusive to those without a strong statistics background. To illustrate the problem more vividly, Vul and colleagues used an example from another domain. They identified a weather station whose temperature readings predict daily changes in the value of a specific set of stocks with a correlation of $r = -0.87$. They arrived at -0.87 by separately computing the correlation between the readings of the weather station in Adak Island, AK, with each of the 3,315 financial instruments available for the New York Stock Exchange between November 18 and December 3, 2008. They then averaged the correlation values of the stocks whose correlation exceeded a particular threshold, thus yielding the correlation value of -0.87. Should you pay for this investment strategy? Probably not. Of the 3,315 stocks assessed, some correlated with the Adak Island temperature measurements simply by chance — and if we select just those (as our selection process would do), there is no doubt we would find a high average correlation. Thus, the final measure (the average correlation of a subset of stocks) was not independent of the selection criteria (how stocks were chosen). This, in essence, is the nonindependence error.

After identifying this potential error, Vul and colleagues evaluated the results of more than 50 peer-reviewed articles in social neuroscience to determine how often this alleged error was committed. They reported that more than half of the articles contained some form of the nonindependence error, and argued that the presence of this error cast serious doubts on the validity of those studies. Vul and colleagues' claims sparked a barrage of commentary on the nonindependence error, both in the scientific community and in the popular press. In fact, so much discussion and correspondence was generated that the journal *Perspectives on Psychological Science* published several commentaries by other social neuroscientists alongside the original article followed by a response to the commentaries by Vul and colleagues.

Not surprisingly, the social neuroscience community responded primarily with strong critiques of Vul et al.'s arguments and conclusions. The following excerpt highlights several of the most prominent criticisms.

Matthew D. Lieberman et al.

Correlations in Social Neuroscience Aren't Voodoo: A Reply to Vul et al.

4 Persp. on Psychol. Sci. 299 (2009)

. . . Because social neuroscience has garnered a lot of attention in a short period of time, singling it out for criticism may make for better headlines. As this article makes clear, however, Vul et al.'s criticisms rest on shaky ground at best.

2. Vul et al.'s proposed upper limit has been disputed. *See, e.g.*, Craig M. Bennett & Michael M. Miller, *How Reliable Are the Results from Functional Magnetic Resonance Imaging?*, 1191 Ann. N.Y. Acad. Sci. 133 (2010). However, it is undisputed that (1) an upper limit for the correlation of personality measures and fMRI does exist and (2) this limit can be calculated.

Vul et al. describe a two-step inferential procedure that would be bad science if anyone did it, but as far as we know, nobody does. They used a survey to assess which authors use this method, but they did not include any questions that would actually assess whether the nonindependence error had occurred. As long as standard procedures for addressing the issue of multiple comparisons are applied in a reasonable sample size, large correlations will occur by chance only rarely, and most observed effects will reflect true underlying relationships. Vul et al.'s own meta-analysis suggests that the nonindependent correlations are only modestly inflated, calling into question the use of labels such as ''spurious'' and ''untrustworthy.'' Finally, Vul et al. make incorrect assumptions when attempting to use average expected reliabilities to inform on the theoretically possible observed correlations.

Ultimately, we should all be mindful that the effect sizes from whole-brain analyses are likely to be inflated, but confident in the knowledge that such correlations reflect meaningful relationships between psychological and neural variables to the extent that valid multiple comparisons procedures are used. There are various ways to balance the concerns of false positive results and sensitivity to true effects, and social neuroscience correlations use widely accepted practices from cognitive neuroscience. These practices will no doubt continue to evolve. In the meantime, we'll keep doing the science of exploring how the brain interacts with the social and emotional worlds we live in.

NOTES AND QUESTIONS

1. ''A long-standing problem in fMRI research concerns the potential pitfalls of reverse inference. As an example, it is well established that the human amygdala responds more strongly to fear related stimuli than to neutral stimuli, but it does not logically follow that if the amygdala is more active in a given situation that the person is necessarily experiencing fear. If the amygdala's response varies along other dimensions as well, such as the emotional intensity, ambiguity, or predictive value of a stimulus, then it will be difficult to make strong inferences from the level of amygdala activity alone.'' Frank Tong & Michael S. Pratte, *Decoding Patterns of Human Brain Activity*, 63 Ann. Rev. Psychol. 483, 497 (2012). As you read subsequent chapters that explain how fMRI measures have been used to support inferences about, for example, whether a person is not telling the truth, ask yourself whether reverse inference is being applied.
2. How do the technical issues summarized in the last two sections affect your sense of how brain imaging might best be used, if at all, in legal contexts?
3. A brain-imaging study reported greater fMRI activation of the amygdala in response to images of African-American as compared to Caucasian faces that was correlated with the participants' racial evaluation as measured by a psychological test. Would you advocate using such information in the *voir dire* process? What problems can you anticipate in such an application?
4. What are corrected and uncorrected statistics? This example, offered by Daniel Bor (http://www.danielbor.com/dilemma-weak-neuroimaging/), should help to clarify.

Imagine that you are running an experiment to see if corporate bankers have lower empathy than the normal population, by giving them and a control group an empathy questionnaire. Low and behold, the bankers do have a lower average empathy score, but it's only a little bit lower. How can you tell whether this is just some random result, or that bankers really do have lower empathy? This is the point where statistical testing enters the frame.

Classically, a statistical test will churn out a probability that you would have gotten the same result, just by chance. If it is lower than some threshold, commonly probability (or p)=0.05, or a 1 in 20 chance, then because this is really very unlikely, we'd conclude that the test has passed, the result is significant, and that bankers really do have a lower empathy score than normal people. All well and good, but what if you also tested your control group against politicians, estate agents, CEOs and so on? In fact, let's say you tested your control group against 20 different professions, and the banker group was the only one that was ''significant.'' Now we have a problem, because if we rerun a test 20 times, it is likely to be positive (under this p=0.05 threshold at least) one of those times, *just by chance*.

As an analogy, say Joe Superstitious flips a coin four times in a row, willing it with all his might to fall on heads four times in a row (with 1 in 16 odds, so pretty close to p=0.05). But the first time it's just a mix of heads and tails. He tells himself that he was just getting warmed up, so let's ignore this round. So he tries again, and this time it's three heads and a tail—or so nearly there. His mojo must be building! The third time it's almost all tails, well that was because he was a bit distracted by a car horn outside. So he tries again, and again and again. Then, as if by magic, on the 20th attempt, he gets all four heads. Joe Superstitious proudly concludes that he is in fact very skilled at telekinesis, puts the coin in his pocket and saunters off.

Joe Superstitious was obviously flawed in his thinking, but the reason is actually because he was using uncorrected statistics, just as the empathy study would have been if it concluded that bankers are less empathic than normal people. If you do multiple tests, you normally have to apply some mathematical correction to take account of how many tests you ran. One simple yet popular method of correction (known as a Bonferroni correction) involves dividing the probability your statistical test outputs by the number of tests you've done in total. So for the bankers to be significantly lower than the control at a p=0.05 criterion, the statistical test would have had to output a probability of p=0.0025 (p=0.05/20), which only occurs 1 in 400 times by chance.

5. If scientists can be challenged with reasoning about statistics, what about lawyers and the lay public? Some scholars have called attention to the problem of statistical illiteracy in health care, journalism, politics, and education (Gerd Gigerenzer et al., *Helping Doctors and Patients Make Sense of Health Statistics*, 8 Psychol. Sci. in the Pub. Int. 53 (2008); John Monahan, *Statistical Literacy: A Prerequisite for Evidence-Based Medicine*, 8 Psychol. Sci. Pub. Int. i (2008)). Fortunately, particular training and exercises can improve anyone's statistical reasoning (P. Sedlmeier & G. Gigerenzer, *Teaching Bayesian Reasoning in Less Than Two Hours*, 3 J. of Experimental Psychol. 380 (2001)).

C. CRITICAL PERSPECTIVES

As the amount of neuroscience in law, society, and mainstream media quickly increases, separating wheat from chaff will become increasingly important. Neuroscience is a tremendously powerful science, but it is not science fiction. And maintaining realistic expectations is tremendously important. In this section we present a several critiques, each aimed at a different facet of brain overclaim.

Judith G. Edersheim, Bruce H. Price & Jordan W. Smoller

"Your Honor, My Genes Made Me Do It"

Wall Street Journal A21 (Oct 22, 2012)

Recent high-profile cases of mass shootings have renewed a vigorous debate about the causes of violent behavior. Predicting violence, whether by sentencing judges, parole boards or mental health professionals, has been a perplexing issue as we try to unravel the personal and social forces behind criminal behavior.

Every decade brings a new ''Eureka!'' In the 1970s, we became convinced that urban poverty caused violence. In the '80s, it was all due to mental illness, and in the '90s substance abuse was definitely the culprit. Now it is malfunctioning brains that are to blame. . . .

It has long been understood that if a mental illness or brain injury causes a bad behavior, it might make a defendant less culpable than an offender whose brain theoretically allowed a wider range of lawful choices. But what about someone without an identified injury or illness of the brain? What about someone who claims that because he has a certain genetic and neurological structure he was predisposed to behave badly?

In the hypothetical presented in the Science paper, a make-believe scientific expert testified that psychopathy has a neurobiologic cause — a gene variation (low monoamine oxidase, or MAO-A, activity), atypical brain functioning (in the amygdala) and other neurodevelopmental factors. While this argument is now becoming common in real criminal cases throughout the country, it represents an unfounded exercise in biologic and neurological determinism.

Here are just a few simple errors in this line of reasoning:

You can't isolate a single gene from an individual and claim that it causes a complex human behavior such as violence. Genes don't operate in isolation but interact with social environments and other genes, and we often don't understand how gene variants operate in the working brain. Yet if we have learned anything over the past 100 years of violence research, it is that human violence is a complicated and multifactorial behavior and is not now — and may never be — reduced to a series of genetic variations or mutations.

It is now abundantly clear that the specific genetic argument presented in these cases is incorrect. Although it became all the rage over the last decade to claim that a defendant with the low-activity variant of the MAO-A gene (dubbed the ''warrior gene'') was predisposed to violent or impulsive behavior, we now know that this link lacks proof.

The early and inflated claims for it were based on studies of a single family that didn't have just a variant of the gene and the enzyme it makes, but was missing it

altogether. Such claims were also based on studies investigating an interaction between the gene and an abusive environment — not on genetic makeup in isolation.

Subsequent studies of violence and MAO-A have had inconsistent results, and it is not even clear that having the low-activity variant of the ''warrior gene'' actually results in low enzyme activity in the brain. More than one in three men carry this genetic variant, but the vast majority don't commit violent crimes. At most, this variant may produce a small increase in the risk of antisocial behavior among men with a history of abuse.

The same controversies have emerged regarding brain-scan evidence. Defendants argue that findings on the new functional neuroimages (mostly from fMRIs) show that their brains operate in abnormal ways, causing them to behave impulsively or violently.

The problem lies in the ambiguity of this causation. Although a small number of brain injuries and illnesses have well-researched impacts on behaviors, many don't. In most cases we are not sure how variants in brain structure and function affect behavior, or whether other regions of the brain can compensate for affected circuits. Structural differences seen on a brain scan don't necessarily result in functional changes in individual behavior.

We are making remarkable strides in identifying specific functional brain networks and the genetic and environmental causes for disruptions in these networks. However, until we can make well-founded, scientifically sound and legally relevant links between genes, brains and behaviors, judges, juries and the public should be wary of neuroscience in the courtroom. The criminal-justice system can't function the way it is supposed to if the evidence presented by the defense or the prosecution inside the courtroom brings us no closer to the truth of what happened outside the courtroom.

Teneille Brown & Emily Murphy

Through a Scanner Darkly: Functional Neuroimaging as Evidence of a Criminal Defendant's Past Mental States

62 Stan. L. Rev. 1119 (2010)

. . . The familiar story is one of weak circumstantial evidence and impressive scientific findings. The combination of these elements may be a powerful prescription for injustice: scientific evidence seems so compelling that it could sway even the most skeptical juror and convince him that the elements of a weak case are proved beyond a reasonable doubt. If, on the other hand, the defendant catches the court's sympathies, then the junk science may swing in the opposite direction and make a weak defense appear stronger. This story has played out before with phrenology, the polygraph, and countless other forensic technologies that have since been discredited. Improper reliance on each of these untested and unreliable technologies has led to unjust outcomes.

These older forensic technologies all have the window dressings of science. Each supplies the court with lab coat-wearing experts who will speak to analyses of ''matching'' criteria with confidence that their methods are sound. But despite popular appeal, phrenology, polygraphy, and fingerprint and handwriting analysis

have never had the ringing endorsement of mainstream physical or biological sciences. In addition, empirical studies have confirmed that there is little reliability or validity in many of these methodologies. However, unlike these sensationalized forensic sciences, functional neuroimaging has the imprimatur of the scientific research community. Indeed, it is difficult to open a copy of Nature or Science without eyeing several colorful functional brain images.

Perhaps, then, the once fledgling field of genetics can provide a more appropriate analogy. So long as genetic samples are not contaminated, the ability to exclude someone from a suspect list based on modern DNA testing is fairly robust. Even so, recall that it took many years for DNA evidence to arrive at the presently-understood state of fallible yet scientifically-valid evidence. However, before the lab standards and analytical models were fully vetted, suspects were unfortunately charged based on DNA samples that were later found to have been carelessly analyzed.

Science can appear to be beyond the reach of human distortion. As a result, the more the scientific evidence relies on complex technologies like computers or imaging devices, the greater the risk that it may be endowed with powers to solve difficult legal questions. Litigants have long used this fact to their advantage, stretching scientific findings in order to retrofit them to legal conclusions. This may be what is happening with fMRI. The device is not yet capable of capturing past mental states, but because the criminal law is sometimes desperate to prove the unprovable, there will almost surely be an increase in proffered evidence and testimony based on this new technology. However, until fMRI is able to reliably capture past mental states, this evidence should not be admissible for such purposes either under FRE 403 or under local standards for admissibility of scientific evidence.

APPENDIX: CHECKLIST FOR JUDGES CONFRONTED WITH FUNCTIONAL NEUROIMAGING EVIDENCE

It is possible that the validity and probative value of fMRI in assessing mental states will improve in the future. It is also possible that the public understanding of the inferential leaps required by functional imaging techniques will progress such that the technology carries less risk of unfair prejudice. In the event that these twin events occur, judges and opposing counsel may appreciate having access to a simple checklist of questions that they can pose when deciding whether to admit fMRI evidence. Although our thesis is addressed to functional brain images used to prove mens rea, this checklist could also be applied to fMRI used as evidence of lie detection and other mental states.

General questions to ask counsel seeking to introduce functional neuroimaging evidence:

(a) Behavioral task. What is the particular behavior assessed during the scan? Why was the particular behavioral task chosen? Is it well supported in the psychological literature as best capturing this type of mental state? Did the subject perform the behavioral task adequately? Is the task vulnerable to manipulation, countermeasures, or malingering? Are the subject's behavioral data within or significantly outside the normal distribution of performance on the task?

(b) Controls. How were the controls selected to be in the control group? Are they the correct reference class? What sort of testing was done on the controls to make sure that they were in fact, ''normal''? Is the sample size large enough to capture normal variance between subjects?
(c) Variance. Can you show us the brain scans of the control group, and are there significant differences among the individuals in this group? How much difference between individuals do we see?
(d) Image construction. Please walk us through the process for developing the image. How did you go from the raw data in the scanner to the color picture of the brain? Can you provide the raw data and exact methodology to an independent party for verification of the image creation process?
(e) Alternative explanations. What are possible alternate explanations for this behavior and corresponding neural activation correlates (i.e., expertise in the task, medication status, drug abuse history, hormonal fluctuations, language or motor limitation, etc.)?
(f) Purpose of fMRI evidence. What justifies the introduction of this brain image over evidence of the accused's behavior at the time of the crime?
(g) Statistical threshold. What statistical threshold was used to create the image? Why was it used?
(h) Causal connection. Is there a known or hypothesized mechanism causally connecting any perceived brain abnormality to a functional deficit? Do we have any data on the incidence of reduced metabolic or hemodynamic activity of this kind resulting in this type of cognitive deficit?

NOTES AND QUESTIONS

1. Although this chapter has focused on neuroscience specifically, it is important to remember that all empirical research necessarily involves various limitations. One interesting observation is that the effect sizes of some published results seem to decline over time. This is known as the ''decline effect'' and it could be the result of regression to the mean, i.e., the first result was particularly large (and hence most subsequent effect sizes are smaller). But as psychologist Jonathan Schooler has pointed out, we remain in the dark on the decline effect because we don't typically have access to unpublished results. Jonathan Schooler, *Unpublished Results Hide the Decline Effect*, 470 Nature 437 (2011). This is not a new problem. And it is not a problem limited to the natural sciences. See, for instance, this particularly jabbing view of the extent to which data can be mined until gems appear: see Edward E. Leamer, *Let's Take the Con Out of Econometrics*, 73 Am. Econ. Rev. 31 (1983).
2. Consider the view of one critic:

> It's not hard to understand why neuroscience is so appealing. We all seek shortcuts to enlightenment. It's reassuring to believe that brain images and machine analysis will reveal the fundamental truth about our minds and their contents. But as the neuro doubters make plain, we may be asking too much of neuroscience, expecting that its explanations will be definitive.

Alissa Quart, *Neuroscience: Under Attack*, N.Y. Times, Nov. 23, 2012. Another critic has similarly written:

> . . . the ''neural'' explanation has become a gold standard of non-fiction exegesis, adding its own brand of computer-assisted lab-coat bling to a whole new industry of intellectual quackery that affects to elucidate even complex sociocultural phenomena.''
>
> Happily, a new branch of the neuroscience explains everything genre may be created at any time by the simple expedient of adding the prefix ''neuro'' to whatever you are talking about. Thus, ''neuroeconomics'' is the latest in a long line of rhetorical attempts to sell the dismal science as a hard one; ''molecular gastronomy'' has now been trumped in the scientised gluttony stakes by ''neuro-gastronomy''; students of Republican and Democratic brains are doing ''neuropolitics''; literature academics practise ''neurocriticism.'' There is ''neurotheology,'' ''neuromagic'' (according to Sleights of Mind, an amusing book about how conjurors exploit perceptual bias) and even ''neuromarketing.'' Hoping it's not too late to jump on the bandwagon, I have decided to announce that I, too, am skilled in the newly minted fields of neuroprocrastination and neuroflâneurship.

Steven Poole, *Your Brain on Pseudoscience: The Rise of Popular Neurobollocks*, New Statesman, Sept. 6, 2012, at 30. Are such criticisms on point? Overstated? Understated? What do and should we expect of neuroscience? Is neuroscience useless to law if it is not definitive?

It would be easy to take away from this chapter the notion that results from brain imaging are so variable, so complicated, and so fraught with uncertainty that such results cannot be employed usefully by law. But that would be a mistake. The key thing to take away is that neuroscience research is most useful to law when it is carefully done, and accurately interpreted. The debates over how best to conduct experiments, and how best to interpret them (both their findings and their limitations) will continue in neuroscience, just as they do in other fields. This chapter has exposed you to some of the major issues in those debates and should equip you to be a cautious and informed consumer of neuroscientific information — neither overskeptical nor overzealous.

Further Reading

General Concerns and Limitations:

Russell Poldrack, *The Role of fMRI in Cognitive Neuroscience: Where Do We Stand?*, 18 Current Opinion in Neurobiology 223 (2008).

John T. Cacioppo et al., *Just Because You're Imaging the Brain Doesn't Mean You Can Stop Using Your Head: A Primer and Set of First Principles*, 85 J. Personality & Soc. Psychol. 650 (2003).

Gregory A. Miller, *Mistreating Psychology in the Decades of the Brain*, 5 Persp. On Psychol. Sci. 716 (2010).

Issues Concerning Risk of False Positives:

Craig M. Bennett & Michael M. Miller, *How Reliable are the Results from Functional Magnetic Resonance Imaging?*, 1191 Ann. N.Y. Acad. Sci. 133 (2010).

Improved Procedures:

Russell A. Poldrack et al., *Guidelines for Reporting an fMRI Study*, 40 NeuroImage 409 (2007).