CHAPTER 4

CAN fMRI READ MINDS?

One of the films that I remember most vividly from my teenage years is *Brainstorm*, in which Christopher Walken plays a scientist who has developed a device that can record the entirety of one's conscious experience and allow others to replay it. When we think of the concept of "mind reading" it is often in the context of this sort of science fiction, but to listen to some fMRI researchers, it is very close to being science fact. In 2009 the television journalist Lesley Stahl interviewed Marcel Just of Carnegie Mellon University, an early researcher in the field of fMRI decoding:

> STAHL: Do you think one day, who knows how far into the future, there will be a machine that will be able to read very complex thoughts, like "I hate so-and-so," or "I love the ballet because . . . "?
>
> JUST: Definitely, and not in twenty years, I think in three, five years.
>
> STAHL (somewhat incredulous): In three years?
>
> JUST: Well, five [laughs].[1]

Fortunately or unfortunately, depending on your point of view, we are not there yet—but research has nonetheless started to uncover findings that verge on what many people would consider legitimate examples of mind reading.

Deciphering the Language of the Mind

Around the time that I published my critique of reverse inference in 2006, many researchers in our field were becoming very

interested in seeing just how far we could push the limits of fMRI in order to determine what a person is thinking about—which researchers sometimes audaciously call "mind reading" but is more accurately termed decoding. As I presented in the discussion of Haxby's work on face decoding in the previous chapter, the goal is similar in spirit to the kind of reverse inference used in the "This Is Your Brain on Politics" piece that I mentioned in chapter 1—to determine the contents of a person's mind from brain activity. However, the approach is very different because it uses statistical tools to quantify exactly how well we can actually decode what a person is thinking about or experiencing.

One way to think about brain decoding is that it is trying to translate between two languages: The natural language of humans, and the biological "language" of thought in the brain. This translation happens indirectly through a set of sensors (such as an MRI machine), since we can't directly "hear" the brain speak its language. As we will discuss later, this kind of translation is likely to be very difficult, if not impossible, using fMRI alone. However, a potentially more achievable goal is to develop a dictionary that maps between patterns of fMRI signal and particular mental states or experiences. This is how Jack Gallant, one of the leaders in the field of fMRI, thinks of the problem:

> In principle, you can decode any kind of thought that is occurring in the brain at any point in time. … you can think about this like writing a dictionary. If you were, say, an anthropologist, and you went to a new island where people spoke a language that you had never heard before, you might slowly create a dictionary by pointing at a tree and saying the word "tree," and then the person in the other language would say what that tree was in their language, and over time you could build up a sort of a dictionary to translate between your language and this other foreign language. And we essentially play the same game in neuroscience.[2]

Such a dictionary wouldn't give us back complete sentences in the brain's language, but it would at least give us the words—and that's often enough to get one pretty far. Nearly all of the work

done to date in the development of fMRI mind reading can be thought of as working toward such a dictionary, and at this point I would say that our dictionary for simple, common words in the brain's language is pretty good.

A Penny for Your Thoughts

The research by Haxby and his colleagues showed that one can decode the contents of visual perception from fMRI signals with very high accuracy. However, this was not particularly surprising to many researchers, since we already knew that visual objects are processed in the temporal lobe, and that the responses of neurons in these areas to visual objects occur even in animals that are under anesthesia—meaning they don't even require the animal to be conscious of the objects. What about conscious thoughts? This question was taken up by John Dylan Haynes, who in the mid-2000s was a postdoctoral fellow at the Wellcome Trust Centre for Neuroimaging in London, which is known to researchers in the field as "the FIL" (for its former name, the Functional Imaging Laboratory). The FIL was (and remains) one of the top neuroimaging centers in the world, and at the time Haynes was working with a young professor named Geraint Rees who had already made a name for himself studying how we are conscious of visual objects. Together they published a set of studies that showed how neuroimaging could be used to decode the contents of our conscious visual experience. In one of these, they presented volunteers with visual stimuli of a different color to each eye,[3] which results in something called "binocular rivalry" in which the person's conscious perception switches occasionally between the two eyes. They recorded what color the person experienced at every point in time during the fMRI scan. The results showed that they could decode from fMRI data which of the colors the person was experiencing at each point in time with relatively high accuracy.

Many people might not consider the decoding of visual experience to qualify as full-blown "mind reading," but in Haynes's next study it would become much harder to dispute that label. He wanted to determine whether he could decode

a person's intentions about a future action from fMRI data. To do this, he gave people a task where they had to choose whether to add or subtract pairs of numbers. In all the trials in the experiment, people first saw a cue that told them to decide whether to add or subtract the numbers on that trial, and then a few seconds later the numbers appeared on the screen. They were given some time to do the addition or subtraction, and then shown a set of probe numbers that included the results of both addition and subtraction, and were told to choose the result from their chosen arithmetic operation. Haynes then asked whether he could use the fMRI signal from the initial cue period (when people were simply thinking about what they were going to do) to predict which of the operations each person would actually choose (which Haynes knew based on which number the person chose on the probe). The results were striking: There were several places in the brain where brain activity was predictive of what action the person would make in the future. The prediction was not perfect—it was around 70% accurate, where random guessing would result in 50% accuracy—but nonetheless it provided a powerful example of how fMRI could decode even very private abstract thoughts.

The work by Haxby, Haynes, and others had an important limitation: in each case, the prediction was person specific. That is, the prediction was performed by taking some data from a person and using it to train a statistical model that could then make predictions based on other data from the same person. There was no testing of the ability to generalize from one person to another. In this way, there was a serious mismatch between what the studies actually showed and some of the discussion in the press, which raised concerns about the use of fMRI to predict crimes or other behaviors. After seeing Haynes's results, I became interested in asking the question of whether it was possible to decode mental states from fMRI data, even when we had never seen the particular person's brain. To ask this question, we took data from 130 people who had each participated in one of eight different studies in my laboratory. The studies involved different cognitive tasks ranging from reading words to choosing monetary gambles to learning

new categories of objects, and it occurred to me that we could potentially train a statistical model to predict which of these tasks they were doing. To do this I teamed up with Steve Hanson and Yarick Halchenko from Rutgers University, who are experts in developing these kinds of models. The field that they work in goes by a number of names including "machine learning," "statistical learning," and "pattern classification"—but you can think of it as the science of how to make good predictions from data. In this case, we want to use brain imaging to predict what a person is thinking, but it's the same set of statistical tools that Facebook uses to recognize faces in photos and Google uses to predict which e-mails are spam and which are not.

What we did was to use a method that is standard in machine learning, known as *cross validation*, which had also been used by Haxby and Haynes before us. The goal of cross validation is to let us tell how well our statistical models can generalize to new data. In principle one could test this by collecting another data set and seeing how well the model from the first data set can generalize to the second, but often we just can't collect another data set. The idea behind cross validation is simple but very powerful. First, we break the data into subsets—for simplicity, let's say that the subsets are as small as possible, which would be individuals in the data set. Thus, for the 130 subjects in our study, we would have 130 subsets. Then, we train the statistical model separately, using all of the data *except* those from one left-out subset, and then test the model on the left-out data. For example, on one round we would fit the model on subjects 1–129, and then test it on subject 130, while in another we would fit the model to all subjects except for subject 129 and then test on subject 129, and so on for all possible subsets. This particular technique is called "leave-one-out" cross validation, for obvious reasons. For each left-out data set, we test how well our predictions match with reality. In this case, we know which of the eight possible tasks each person was performing, and the statistical model gives us a prediction for which task the person is doing based on his or her brain activity; we count up how often the prediction is correct, and that is our measure of accuracy. We were able to predict the task for the left-out subject with about 80% accuracy—for comparison, if we were

just guessing we would expect to get it right about 13% of the time. This study was the first to demonstrate conclusively that it is possible to decode mental states from fMRI data, even when our statistical model has been trained on other people, laying the groundwork for later research that would push brain decoding even further.

Reading the Mind's Eye

The research on fMRI decoding that we have discussed so far has focused on the ability to choose between a small number of possible states—for example, eight different types of images in Haxby's study (chapter 3)—but true mind reading implies the ability to reconstruct any arbitrary thought or image from brain activity. Two studies published in 2008 provided a hint that this might be possible, building on earlier work by a French group led by Bertrand Thirion. The approach that these two studies used was different from the previous work in a particular way: while the previous decoding studies had used generic machine-learning methods (which could have just as easily been used to predict what you are planning to buy on Amazon.com), these newer studies used models that were specifically built to mimic the structure of the human visual system.

Kendrick Kay, working with Jack Gallant at Berkeley, wanted to test whether it was possible to identify natural images amongst a large number of possibilities. Kay and his colleague Thomas Naselaris each spent several hours in an MRI scanner, ultimately viewing almost 2,000 different natural images.[4] They used the brain responses on 1,750 of these images to create a statistical model that was informed by the structure of the human visual cortex. This "quantitative receptive field model" basically tried to figure out, for each voxel in the visual cortex, what parts of the visual world it was sensitive to, which we generally refer to as a "receptive field" (figure 4.1)—you can think of this as a map of which parts of the visual world a particular part of the brain is paying attention to. By combining these across many voxels, they were able to generate a model of the entire visual cortex. Then, using the data from the remaining 120 images, they asked

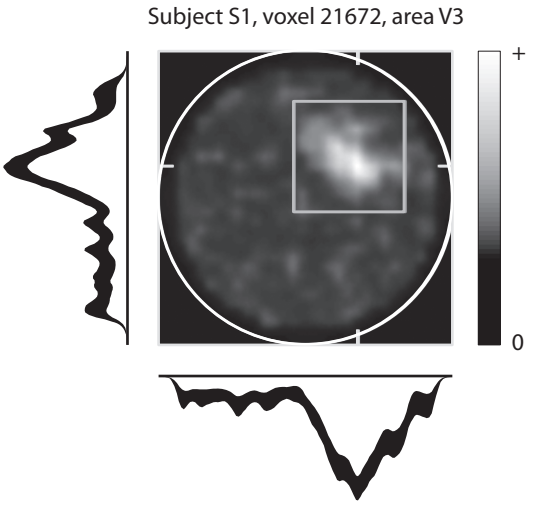

Figure 4.1. A reconstruction of the "receptive field" for a single voxel in Kay's 2008 study. The plot illustrates how different voxels in the brain respond to stimulation in different parts of the visual field—each voxel has its own small area. The bright spot shows the part of the visual scene that this one voxel was sensitive to—in this case, a patch slightly right of the center of the visual field. Unpublished image courtesy of Kendrick Kay.

whether the model could identify which image was being viewed (out of the entire set of 120) using only the fMRI data. To do this, they took the actual brain activity for each image and compared it with the predicted brain activity from the model for each of the 120 images, asking whether the actual brain activity was closest to the predicted activity for the actual image being viewed versus the other 119 possible images. If one were guessing one would get it right only less than 1% of the time, but for both of the subjects Kay and Gallant were able to choose the correct picture with high accuracy (92% for one subject, 72% for the other). This showed that the ability to decode visual image content went far beyond the small number of categories studied in the previous work, but it didn't quite get to full-blown reconstruction of an arbitrary image.

A second study by Yukiyasu Kamitani and his colleagues in Kyoto, Japan, moved a step closer to reconstructing a viewed image.[5] They used an approach similar in spirit to the one used by Kay and Gallant, in which they built a large set of simple

Figure 4.2. An illustration of image reconstruction using fMRI. The *left panel* shows a figure adapted from the 2008 paper by Miyawaki and colleagues. The *top row* shows the actual images that were presented to the subject, and the *bottom row* shows the visual pattern that was reconstructed on the basis of the subject's brain activity. The *right panel* shows images adapted from the 2009 paper by Naselaris and colleagues, with actual images on top and the closest matching actual image on the bottom. The numbers in the bottom-right corners of the lower images are a measure of the accuracy with which the model could predict brain activity. Left panel reprinted from *Neuron* 60, no. 15, Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masaaki Sato, Yusuke Morito, Hiroki C. Tanabe, Norihiro Sadato, and Yukiyasu Kamitani, "Visual Image Reconstruction from Human Brain Activity Using a Combination of Multiscale Local Image Decoders," 915–29, Copyright 2008, with permission from Elsevier. Right panel reprinted from *Neuron* 63, no. 6, Thomas Naselaris, Ryan J. Prenger, Kendrick N. Kay, Michael Oliver, Jack L. Gallant, "Bayesian Reconstruction of Natural Images from Human Brain Activity," 902–15, Copyright 2009, with permission from Elsevier.

decoders, each of which learned to decode the signal in a small part of the visual image based on data from a small portion of the visual cortex. They then used a machine-learning method to learn how to combine these in order to best reconstruct the image that had been presented. The results obtained on simple geometric shapes were impressive (figure 4.2), including a reconstruction of the word "neuron," which was the name of the journal where the work was ultimately published.

The final step toward full reconstruction of a natural image also came from Gallant's group at Berkeley, this time led by Thomas Naselaris.[6] This study made two major advances that allowed these researchers to reconstruct entire natural images. First, they used the idea of Bayesian analysis, in which the data are combined with prior knowledge to determine the

best possible reconstruction of the data. In this case, the prior knowledge consisted of six million images selected at random from the Internet. Essentially what they did was use the fMRI data to create a predicted image, and then ask which of the six million actual images was most similar to that predicted image. The second advance was to include information about the semantic category of the scene (such as animate vs. inanimate or indoor vs. outdoor), which was generated by hand for each of the training images. They then used data from a higher visual area, which is known to respond to object catgories, to detect which category was present. Putting these together they were able to acheive very impressive "reconstructions" (see figure 4.2)—each of which was actually a selection of the closest image from among the six million images. This approach was later used by the same group to reconstruct movies as well. While one can question whether this really counts as "reconstruction," it's likely that any model that is going to successfully decode images from brain activity will have to take advantage of strong prior information about the features of natural images, just as our visual system takes advantage of such information when we recognize images.

Decoding Consciousness in Brain Injury

Of all of the possible things that a neuroscientist fears, a severe traumatic brain injury has to be at the top of the list. A blow to the head from a car accident or a fall can send a perfectly healthy and cognitively intact person into a mental no-man's-land in a split second, and the best one can hope for after such an injury is often basic self-sufficiency; a complete recovery of intellect and personality after a severe brain injury is often out of the question. Advances in emergency treatment have allowed many more people to survive serious brain injuries, but these survivors are usually left in a state of altered consciousness for an extended period of time. The lowest level of consciousness is *coma*, in which people are completely unresponsive (even to painful stimulation) and do not open their eyes. Coma is not the same as "brain death," which represents an even deeper

and irreversible level of damage to the brain; the brain of a comatose person still has electrical activity, though it is highly abnormal. A person in a coma will usually over time begin to show some signs of brain function, such as opening the eyes, but often remain nonresponsive and doesn't show any outward signs of conscious awareness; this is referred to as a *vegetative state* and can sometimes last for years if the person is fed via a tube and otherwise cared for. In other cases, the person begins to show increasing signs of consciousness, and can often exist in what is now called a *minimally conscious state*—in which he or she drifts in and out of lucidity, sometimes being able to interact with others while at other points being unresponsive.

Because people in a vegetative state appear to be unconscious and show abnormal electrical activity in the brain when measured using electroencephalography (EEG), it was long assumed that they did not have conscious awareness. However, we know of cases where a person can be fully conscious yet seem completely unresponsive—this occurs in "locked-in syndrome," a very rare syndrome caused by damage to the brainstem, which leaves patients fully conscious but unable to make any movements other than blinking and moving their eyes. What if some of the people thought to be in a vegetative state were actually aware but unable to express themselves, like the locked-in patients? Adrian Owen has spent the past two decades trying to answer this question. He is a cognitive neuroscientist who started out studying basic cognitive processes using PET imaging, but at some point became obsessed with understanding conscious awareness and using fMRI to identify it in people suffering from disorders of consciousness.

Owen's solution was surprisingly simple. He placed individuals in the MRI scanner and then asked them to do one of two different things: either imagine playing tennis, or imagine walking through their house. He chose these two different tasks because he knew from studies of healthy people that the two tasks should evoke very different patterns of activity if the person performed the task properly. They tested the method on a 23-year-old woman who had suffered a severe brain injury in a car accident five months earlier and who remained in a vegetative

state, completely unresponsive to stimulation. While her unresponsiveness suggested that she did not have intact cognitive function, fMRI showed a different story. When she was told to imagine playing tennis there was activity in her premotor cortex, while there was activity throughout the network that is engaged during spatial navigation in healthy people when she was told to imagine navigating her house (see color plate 6).[7] This landmark finding inspired a much broader analysis, which has shown that the proportion of people in a vegetative state who pass the test for conscious awareness is relatively low—in a subsequent study led by Martin Monti, 5 out of 54 individuals in a vegetative state showed evidence of awareness.[8]

The use of neuroimaging to detect conscious awareness in people with brain injuries is a major advance that shows the real-world utility of fMRI decoding. At the same time, these studies have raised some difficult ethical and medical questions. Foremost, the ability of people in a vegetative state to answer questions means that they could in principle be asked the most difficult possible question of all: Do they wish to continue living? These individuals are not able to feed themselves and thus require feeding through a tube, which they could in theory ask to be stopped. How would we decide whether they have sufficient reasoning ability to make this judgment, and how would we respond? It's worth noting that while healthy people would assume that many individuals in such a state would choose to end their own lives, there is some evidence to suggest the contrary. In particular, a study of individuals with locked-in syndrome that assessed their subjective well-being showed that the large majority claimed to be happy with their life, and only 7% expressed the desire to be allowed to die.[9] Caregivers and physicians also need to think deeply about how the knowledge of a person's state of conscious awareness and the ability to ask him or her questions using fMRI would change the way that they treat the person. Would they try to seek consent for a risky medical procedure that has the potential for negative outcomes? We also need to know whether, and how well, these markers of consciousness predict recovery in the future, and to decide whether knowing about a person's cognitive status

is actually useful if it does not provide any useful clinical guidance.

Are You Really in Pain?

I once attended a lecture by an eminent pain researcher, who started his talk by making what may seem like an obvious point: pain is in the brain. What he meant was that the aversive nature of pain depends on how our brain responds to the input from our peripheral nerves that carry impulses to the brain. When I accidentally smash my thumb with a hammer, specialized nerve receptors in my thumb send a message to my brain telling it that something bad has happened, and I experience those impulses as the aversive experience of pain. Pain is annoying but essential, as it alerts us to the need to protect the injured area to prevent further injury, and to seek treatment if needed. The utter importance of pain is evident in the plight of people who suffer from *congenital analgesia*—that is, they are born without the ability to feel pain. Mo Costandi described the case of Ashley, a teenager born with this disorder:

> As a newborn, she barely made a sound, and when her milk teeth started coming out, she nearly chewed off part of her tongue. Growing up, she burnt the skin off the palm of her hands on a pressure washer that her father had left running, and once ran around on a broken ankle for two whole days before her parents noticed the injury. She was once swarmed and bitten by hundreds of fire ants, has dipped her hands into boiling water, and injured herself in countless other ways, without ever feeling a thing.[10]

The experience of acute pain has been studied extensively using fMRI, generally by subjecting subjects to painful stimulation using a heat probe (or, in studies of visceral pain, a rectal balloon). This kind of acute pain results in activity in a broad set of brain areas that has come to be known as the "pain matrix"—which includes the somatosensory areas that receive sensory inputs from the body, as well as areas such as the insula and anterior cingulate cortex. You have already heard about

these latter areas at many points throughout the book, so it hopefully occurs to you to ask whether activity in these areas is specific to pain, which in fact it is not. However, with the development of machine-learning techniques, researchers have begun to ask whether the experience of pain can be decoded from brain activity. Tor Wager is a researcher at the University of Colorado–Boulder who has led the charge to develop what he calls a "neurological signature of pain." In a landmark paper published in 2013, his group demonstrated that they could predict levels of reported pain with a high degree of accuracy using fMRI combined with a machine-learning technique that combined data in an optimal way across many different regions.[11] Given a person's brain image, they were able to predict that person's rating on a nine-point pain scale, with an error of about one point, and they were able to tell whether or not a person was in pain with an accuracy of greater than 90%.

What about different types of pain? Anyone who has experienced the heartbreak of a failed relationship will know that while the sensation is different from acute physical pain, it hurts nonetheless. In order to test whether their model could distinguish between physical pain and heartache, Wager and his colleagues recruited a set of subjects who had recently been broken up with. During scanning, they showed them pictures of their former lover as well as pictures of other friends, and also subjected them to physical pain in a separate scan. Although seeing pictures of the person who had rejected them caused activity in many of the same areas that are activated during the experience of physical pain, the algorithm was able to distinguish the two with a high degree of accuracy.

While acute pain is a useful signal to help prevent and limit injury, when it becomes chronic it can lead to misery, disability, and sometimes suicide. The community of cognitive neuroscience lost one of our brightest stars to chronic pain in 2012, when Jon Driver, a professor at University College London, committed suicide after suffering from chronic pain resulting from an injury sustained in a motorcyle accident the year before. Chronic pain is also a central feature in many civil law suits, in which individuals pursue damages related to their professed

suffering. A challenge in these suits has always been that it was impossible to know whether the person is truly suffering from the claimed pain or is falsely claiming to be in pain in order to obtain a financial judgment or settlement. Could these new fMRI pain signatures provide a better way to validate claims of chronic pain? Potentially, but we are not there yet. Most importantly, chronic pain appears to have a different basis in the brain than acute pain. Research by Vania Apkarian and his colleagues has shown that chronic pain engages a different set of brain areas than acute pain, such that chronic pain involves areas that are more involved in emotional processing compared with the areas involved in acute pain.[12] This means that the neural pain signature developed by Wager's group would likely not work to detect many forms of chronic pain; a different tool would be necessary.

Just as in fMRI lie detection, the lack of a solid scientific background has not stopped people from trying to commercialize the techniques and use them in court. Unlike fMRI lie detection, pain detection *has* been allowed as evidence: in the case of *Koch v. Western Emulsions Inc.*, Carl Koch sued his employer over chronic pain resulting from an accident on the job, presenting fMRI evidence for the reality of his pain.[13] There are a number of companies that now market fMRI pain detection for the purpose of civil lawsuits in the United States. Most of these companies use secret methods, but one of them—Chronic Pain Diagnostics—has published its work in a peer-reviewed journal,[14] for which it should be applauded. We often think that publication in a peer-reviewed journal is a stamp of approval for the quality of a particular piece of research, but that's not always the case, as it depends on the peer reviewers having the right expertise to find flaws if they exist, which is especially tricky with new methods like machine-learning techniques. In the case of the work from the Chronic Pain Diagnostics team, the glaring flaw is that the sample size was far too small to make any meaningful inferences—only 13 people in each of the groups being studied. The study claimed to be able to detect chronic pain with 92% accuracy, but work by French machine-learning expert Gaël Varoquaux has shown small samples can lead to highly

inflated estimates of accuracy on just this kind of test.[15] In addition to these concerns, questions remain about the potential ability for subjects to intentionally trick the scanner (which will also appear in the discussion of fMRI lie detection in chapter 6)—though research from Tor Wager's group has shown that the neural pain signature is not affected by imagined pain.[16] I certainly hope that one day fMRI will be able to help people in pain obtain the justice they deserve and prevent abuses of the legal system, but I think there is a substantial amount of hard work to do before we get there.

It is clear that fMRI decoding is already very powerful, and becoming more so as machine-learning techniques become increasingly powerful. In this chapter we have already seen a number of applications of fMRI decoding to real-world problems. In chapters 6, 7 and 8, we will dig further into the ways in which the use of decoding has started to impact business, medicine, and the law.