Neuronal activity has metabolic consequences. Specifically, energy is required for the maintenance and restoration of neuronal membrane potentials. That energy is not stored within the brain, but must be continuously supplied by the vascular system through the delivery of glucose and oxygen. How can this delivery of oxygen and glucose to active neurons lead to a signal that can be measured by MRI? In this chapter, we will answer this question by describing blood-oxygenation-level dependent (BOLD) fMRI. We consider, in more detail, the changes in the vascular system that are triggered by neuronal activity, and how those changes can (somewhat counterintuitively) alter certain types of MR images. Then, we consider the spatial and temporal properties of BOLD fMRI. Some recent studies have provided striking examples of fMRI's power: mapping activation to specific cellular layers within the cortex, or determining the direction of information flow between active regions. These advances have been made possible both by improvements in scanner hardware and by new approaches to fMRI analysis. Yet, our measurements in fMRI will be ultimately limited by the spatial and temporal concordance of the BOLD signal with the underlying neuronal activity. Thus, researchers using fMRI should understand both its capabilities and its limitations.

History of BOLD fMRI

While investigating the molecular structure of hemoglobin in 1936, the American chemist and Nobel laureate Linus Pauling and his student Charles Coryell discovered a remarkable and (for our purposes) fortuitous fact of nature: the hemoglobin molecule has magnetic properties that differ depending on whether or not it is bound to oxygen. Oxygenated hemoglobin (Hb) is diamagnetic; that is, it has no unpaired electrons and zero magnetic moment. In contrast, deoxygenated hemoglobin (dHb) is paramagnetic; it has both unpaired electrons and a significant magnetic moment. Completely deoxygenated blood has a magnetic susceptibility about 20% greater than fully oxygenated blood. Pauling and Coryell noted wryly that this fact had eluded previous researchers, including the great nineteenth-century physicist Michael Faraday, only because they had not separated arterial blood (which contains only oxygenated hemoglobin) from venous blood (which contains both oxygenated and deoxygenated hemoglobin).
Figure 7.1 Effects of blood deoxygenation on MR relaxation constants. Shown are the differential effects of blood deoxygenation on transverse and longitudinal relaxation times, as expressed by the constants $1/T_2$ (blue circles) and $1/T_1$ (red circles). The x-axis indicates the square of the proportion of deoxygenated blood. Note that oxygenation increases from left to right. Clearly evident is the decrease in $1/T_2$ with increasing oxygenation; that is, the more deoxygenated hemoglobin that is present, the shorter the $T_2$ (which here represents loss of phase due to both spin-spin interactions and local field inhomogeneities). Note that $T_1$ is not affected by blood oxygenation level. (After Thulborn et al., 1982.)

Because paramagnetic substances distort the surrounding magnetic field, nearby protons will experience different field strengths and will thus precess at different frequencies, resulting in the more rapid decay of transverse magnetization (i.e., a shorter $T_2^*$). Thus, MR pulse sequences sensitive to $T_2^*$ should show more MR signal where blood is highly oxygenated and less MR signal where blood is highly deoxygenated. This prediction was verified experimentally in the early 1980s by Thulborn and colleagues, who found that the decay of transverse magnetization depended on the proportion of oxygenated hemoglobin within a test tube of blood (Figure 7.1). They noted that the magnitude of this effect increased with the square of the strength of the static magnetic field. At a low field strength (i.e., less than 0.5 T), there was little difference between the transverse relaxation values for oxygenated and deoxygenated blood, but in higher fields (i.e., 1.5 T or greater), their values differed significantly. So, strong static magnetic fields are necessary for MR imaging of $T_2^*$-weighted contrast in blood. These results showed that changes in blood oxygenation could, in principle, be measured using MRI.

**Discovery of BOLD contrast**

During the late 1980s, Seiji Ogawa, a research scientist at Bell Laboratories, investigated the possibility of examining brain physiology using MRI. Ogawa and his colleagues recognized that MRI was ill-suited for examining physiological processes directly. Since standard MRI contrasts are based on properties of hydrogen, the ubiquity of hydrogen in water throughout the body precludes the detection of the very subtle changes in concentration associated with most metabolic reactions. For MRI to be useful in measuring physiology, it would need to be sensitive to some indirect measure of metabolism. One possibility was blood flow, since metabolic processes require oxygen that is supplied through hemoglobin within red blood cells. Based on the earlier work by Thulborn and colleagues, Ogawa hypothesized that manipulating the proportion of blood oxygen would affect the visibility of blood vessels on $T_2^*$-weighted images.

In a seminal 1990 study, Ogawa and colleagues tested this hypothesis by scanning anesthetized rodents using high-field (7.0 T and greater) MRI. To
Figure 7.2 A schematic illustration of blood-oxygenation-level dependent (BOLD) contrast. Ogawa and colleagues manipulated the amount of oxygen in the blood of rats by adjusting the contents of the air that the rats breathed. (A) When the rats breathed pure oxygen, the cortical surface had a uniform texture on images sensitive to $T_2^*$ contrast. (B) But when the rats breathed normal air, there were areas of signal loss, shown as lines corresponding to blood vessels within the cortex. These lines indicated areas with increased amounts of deoxygenated hemoglobin. This forms the basis for BOLD contrast. (After Ogawa et al., 1990.)

To manipulate blood oxygenation, they changed the proportion of oxygen that the animals breathed. When the rodents breathed pure oxygen, gradient-echo images of their brains showed only structural differences between tissues (Figure 7.2A). But when the rodents breathed normal air (21% oxygen), the images took on a very different character. Thin dark lines became visible throughout the cerebral cortex, usually perpendicular to its surface (Figure 7.2B), and if the oxygen content was further reduced to 0%, the lines became even more prominent. They attributed these thin lines to the magnetic susceptibility effects of paramagnetic deoxygenated hemoglobin in blood vessels. Conversely, when the hemoglobin was bound to oxygen, it was diamagnetic and had little effect on the surrounding magnetic field.

To verify this interpretation, they placed test tubes with oxygenated or deoxygenated blood into a saline-filled container and created images using both spin-echo and gradient-echo pulse sequences. Recall from Chapter 5 that spin-echo images are largely insensitive to $T_2^*$ effects, whereas gradient-echo images are distorted by $T_2^*$ decay. The tubes containing oxygenated blood appeared as black circles on both types of images, since the blood had shorter $T_2^*$ values than the surrounding saline (Figure 7.3A and B). The spin-echo image of the deoxygenated blood was likewise nearly normal (Figure 7.3C). The largest effect by far was observed for gradient-echo images of the deoxygenated blood, which exhibited a dramatic signal loss that extended well beyond the test tube (Figure 7.3D). These results demonstrated unequivocally that deoxygenated blood decreases the measured MR signal in $T_2^*$ images.

Ogawa and colleagues speculated that this blood-oxygenation-level dependent contrast could identify areas of increased brain activity. They

---

Figure 7.3 Magnetic properties of oxygenated and deoxygenated hemoglobin. To verify that the effects illustrated in Figure 7.2 resulted from changes in blood oxygen level, Ogawa and colleagues compared images of oxygenated and deoxygenated blood collected using both spin- and gradient-echo imaging. The images of oxygenated blood were not distorted, regardless of whether spin-echo (A) or gradient-echo (B) images were acquired. The spin-echo image of the deoxygenated blood (C) was slightly distorted, but the distortion did not extend to the area surrounding the test tube. However, there was substantial signal loss surrounding the gradient-echo image of the deoxygenated blood (D), showing that the presence of deoxygenated hemoglobin reduces the MR signal from water molecules outside of the test tube in adjacent space. (After Ogawa et al., 1990b.)
hypothesized two possible nonexclusive mechanisms. First, neuronal activity might increase oxygen consumption, which would increase the amount of deoxygenated hemoglobin, given a constant blood flow. Alternatively, increased blood flow in the absence of increased metabolic demand might decrease the amount of deoxygenated hemoglobin.

To evaluate the contribution of oxygen consumption to BOLD contrast, Ogawa and colleagues manipulated the gases inhaled by anesthetized rats while collecting $T_2^*$ images and measuring brain activity using concurrent EEG. At a relatively high anesthesia level (3.0% halothane), there was reduced spontaneous brain activity and relatively little BOLD contrast, but at a low anesthesia level (0.75% halothane) there was both increased brain activity and greater BOLD contrast. These results indicated that the reduction in the BOLD signal depended on the metabolic demand for oxygen. To evaluate the effect of blood flow on BOLD contrast, they compared two inhalant conditions: pure (100%) oxygen and a mixture of 90% oxygen and 10% carbon dioxide. Carbon dioxide in the blood does not have significant paramagnetic effects, but it does increase overall blood flow (e.g., it increased velocity in the sagittal sinus by about 30%). While significant BOLD contrast was observed in the pure oxygen condition, the contrast disappeared when the animals breathed the CO$_2$ mixture. The researchers inferred that with greater blood flow in the absence of increased metabolic demand, the deoxygenated hemoglobin is essentially flushed from the venous system, leaving only oxygenated hemoglobin that does not reduce the BOLD signal.

In summary, BOLD contrast depends on the total amount of deoxygenated hemoglobin present in a brain region, which in turn depends on the balance between oxygen consumption and oxygen supply. When neuronal activity increases, one might predict that oxygen consumption will also increase, leading to more deoxygenated hemoglobin and a darker MR image, as Ogawa and colleagues originally hypothesized. Yet, increased neuronal activity actually increases the signal of $T_2^*$ images! How can this be? The answer requires a more complete exploration of the relationships between cerebral blood flow, blood oxygenation level, and metabolism.

**The coupling of metabolism and blood flow**

Glucose is the major energy source for the brain, and oxygen facilitates the most efficient conversion of glucose to ATP. In the 1970s, Sokoloff and colleagues used an invasive imaging technique known as autoradiography to establish that regional brain changes in the rate of glucose metabolism were coupled with an increase in blood flow to that same region. The overall accounting of glucose and oxygen consumption during oxidative metabolism is given by Equation 7.1:

$$C_6H_{12}O_6 + 6O_2 = 6CO_2 + 6H_2O$$

(7.1)

Each glucose molecule to be oxidized requires the consumption of six oxygen molecules, which together lead to six carbon dioxide and six water molecules as products. Thus, the ideal oxygen-to-glucose index (OGI) would be 6:1 if all of the glucose that entered the brain were oxidatively (i.e., aerobically) metabolized. Brain measurements performed under resting conditions have established an OGI of approximately 5.5:1, indicating that while the vast majority of glucose metabolism in the brain is oxidative, a small but significant glucose fraction may be metabolized nonoxidatively (i.e., anaerobically). The rates of oxygen and glucose consumption during local neuronal activity may differ, however, and this fact has engendered considerable controversy that persists today.

In an influential series of PET experiments (see Box 7.1) conducted in 1988 by Fox, Raichle, and their colleagues, cerebral blood flow (CBF), cerebral meta-

**autoradiography** An invasive imaging technique that labels molecules using radioactive isotopes and then measures the concentration of those molecules by exposing slices of tissue to photographic emulsions.
Positron emission tomography, or PET, is a powerful functional imaging technique. At the beginning of a PET study, the researcher uses a particle accelerator called a cyclotron to create a radioactive isotope, or tracer, that can be attached to a molecule to enter a biological pathway of interest. For example, $^{18}\text{F}$, a radioactive isotope of fluorine, can be attached to glucose, creating the molecule fluoro-2-deoxy-D-glucose, or FDG. The fluorine tracer does not prevent FDG from entering into the normal pathways for glucose metabolism. Thus, researchers can inject a bolus of FDG into a vein and then use imaging to determine where it is taken up by cells.

As the radioactive isotope decays, it emits a positron (the antimatter counterpart of an electron). When the emitted positron collides with a nearby electron (Figure 1A), they are mutually annihilated and produce two gamma rays that travel in opposite directions. The gamma rays are subsequently detected by their near-simultaneous impact on opposite sides of a ring of scintillation crystals that surround the subject’s head (Figure 1B). A computer algorithm then evaluates the number and timing of impacts at all of the crystals surrounding the head and traces the paths taken by the gamma rays back to their origins. Through this method, the distribution of the radioactive isotope in the brain can be measured, and changes in glucose uptake in different brain regions caused by sensory, motor, or cognitive activity can be determined (Figure 1C). PET imaging can also be used to study oxygen metabolism or blood flow using $^{15}\text{O}$, a radioactive isotope of oxygen. Certain neurotransmitters can be similarly labeled. For example, $^{18}\text{F}$ can be attached to dopamine to study its distribution in the human brain.

PET scanning provides a relatively direct and easily interpretable measure of brain metabolism, and for many years it was the mainstay of human functional neuroimaging. However, its dependence on high-energy gamma rays makes PET imaging time-consuming and requires shielding to protect the subject from the radiation.

A functional neuroimaging technique that creates images based on the movement of injected radioactive material.

---

Figure 1 Positron emission tomography (PET) imaging. Until the mid-1990s, the most common functional neuroimaging technique was PET, which relies on the injection of a radioactive tracer into the bloodstream. As the tracer decays, it emits positrons, which travel a short distance before colliding with an electron (A). The collision results in a pair of emitted gamma rays that travel in opposite directions. The PET scanner (B) consists of a series of coincidence detectors that record the simultaneous arrival of these gamma rays. Depending on the tracer used, PET can be sensitive to several aspects of brain metabolism, including blood flow or oxygen consumption. The output of a PET scan indicates the number of events measured from each voxel during a long time period. (C) These numbers can be converted to statistical maps, which can then be overlaid on anatomical images, often from MRI. (C courtesy of Dr. David Madden, Duke University.)
radiation (or ionizing radiation, as named for its potential to break chemical bonds) presents problems for human studies. Radiation exposure in human research is carefully regulated, and subjects can participate in only a few PET scans. There are other drawbacks to PET imaging. Its spatial resolution is limited by the distance the positron travels away from the labeled molecule before it collides with an electron. This is dependent on the particular isotope used; for example, positrons emitted by $^{18}$F will travel about 2.6 mm before encountering an electron. Recall that it is the location of the gamma ray emission that is localized and not the location of the molecule of interest. More limiting, however, is the very poor temporal resolution of PET imaging. Many emissions must be detected to produce an image with sufficient signal-to-noise ratio, requiring the collection of data over a long period of time. For example, an image of blood flow based on $^{15}$O may take 90 seconds to acquire, while an image of glucose metabolism based on $^{18}$F may take 30 to 40 minutes to acquire. These acquisition times severely limit the temporal resolution of PET imaging and restrict the types of experimental designs that can be used.

When compared with PET imaging, MRI has several advantages. Because MRI does not involve ionizing radiation, subjects can participate repeatedly without the cumulative health risks of radiation exposure. Images with high signal-to-noise ratios can be acquired in less than a second, and spatial resolution is limited primarily by the motion of the sample and by the signal-to-noise ratios, not by the inherent uncertainty in the measurement technique. And, fMRI can be used in event-related designs to identify processes specific to one phase of a complex experiment, to understand changes in functional connectivity, or to classify data on a trial-by-trial basis. But these advantages should not lead to the incorrect assumption that PET imaging has no relevance for modern neuroscience. Even with the caveats described above, PET can be used to image glucose or oxygen consumption directly. In comparison, BOLD fMRI does not provide any direct information about metabolic processes, but instead measures an indirect correlate of those processes. Moreover, only PET can be used to create images specifically sensitive to a single metabolite or neurotransmitter. Thus, for many important questions about brain physiology and function, PET imaging remains the technology of choice.

ionizing radiation Electromagnetic radiation that has sufficient energy to separate electrons from electrically neutral atoms, turning them into ions.

The main events included in this chapter are metabolic events. The measurements of local brain metabolism include resting cerebral metabolic rate for glucose (CMR$_{glu}$) and cerebral metabolic rate for oxygen (CMRO$_2$) were measured during rest and during visual stimulation. When subjects were exposed to prolonged visual stimulation, CBF in the visual cortex increased by 50% and CMR$_{glu}$ increased by 51%, consistent with Sokoloff’s autoradiographic findings in animals. However, CMRO$_2$ increased by only 5%. The authors concluded that most of the increased uptake of glucose during stimulation was not oxidized but rather was metabolized nonoxidatively through anaerobic glycolysis. Recalling our discussion from the previous chapter, anaerobic glycolysis is relatively inefficient (but fast), yielding only two ATP molecules for each glucose molecule consumed, and thus the energy produced from the increased glucose uptake would be relatively small. Corroborating evidence has come from studies by Prichard and colleagues, who demonstrated that extended visual stimulation results in increased lactate production, which would be expected, since lactate is the primary end product of anaerobic glycolysis.

Not all investigators accepted the apparent uncoupling of CMRO$_2$ and CMR$_{glu}$ during stimulation as indicative of anaerobic metabolism. In 1996 Malonek and Grinvald used a high-resolution optical imaging method in which the surface of the visual cortex of a cat was exposed to a light source and the reflected light was analyzed. Because different molecules (e.g., oxygenated and deoxygenated hemoglobin) absorb light of different wavelengths, the spec-
trum of the reflected light can be used to determine the presence of different molecules at any given location in the cortex. The authors selectively activated small, spatially segregated populations of neurons in the visual cortex by presenting the cat with images of line gratings at particular orientations. The expected spatial pattern of neuronal activation across the visual cortex could then be compared with the spatial patterns of oxygenated and deoxygenated hemoglobin accumulation, and the optical imaging method allowed for measurements with high spatial and temporal resolution.

The results of the experiment showed that the evoked changes in hemoglobin and deoxygenated hemoglobin were quite distinct (Figure 7.4). The deoxygenated hemoglobin–time curve showed a rapid increase that peaked about 2 s after the onset of the stimulus and then rapidly declined. By 6 s after the onset of the stimulus, the deoxygenated hemoglobin signal had declined to well below the prestimulus baseline level. In contrast, the oxygenated hemoglobin signal had a slightly delayed onset and then a much slower rise to a peak at about 5 to 6 s after stimulus onset. It was also much greater in amplitude than the deoxygenated hemoglobin signal. Moreover, unlike the weak deoxygenated hemoglobin signal, the spatial pattern of the oxygenated hemoglobin signal did not reflect the expected pattern of neuronal activity. Indeed, it extended into regions where there should have been no neuronal activity.

Several conclusions can be drawn from this experiment. First, the active neurons utilized whatever oxygen was already present to support their initial activity. This suggests that increased metabolism at the onset of activation is oxidative. Second, there is an exquisite spatial correspondence between the initial neuronal activity and the initial increase in the deoxygenated hemoglobin signal. Third, the poor correspondence and greater spatial extent of the later oxygenated hemoglobin response indicates that the regulation of blood flow and oxygen delivery to the cortex is on a coarse spatial scale and mismatched metabolic needs. As Malonek and Grinvald put it, this is analogous to “watering the entire garden for the sake of one thirsty flower.” These data suggest that the uncoupling reported by Fox and Raichle results not from an increase in anaerobic glycolysis but instead from a superfluous perfusion of oxygenated blood that exceeds metabolic need. However, CMR_{glu} was not measured in this experiment, so whether the spatial pattern of the OGI matched the pattern of neuronal activity could not be determined.

Recent studies have indicated that changes in coupling between metabolism and flow may have functional consequences. Neuronal activity may lead to prolonged reductions in glucose metabolism compared with flow, and those changes may persist even after activity ceases. Moreover, the coupling between CBF and CMR_{glu} differs across regions, as shown by an intriguing 2008 paper by Gur and colleagues. These authors found that brain regions differed in whether they exhibited CBF-CMR_{glu} coupling, were hyperperfused (i.e., had greater flow than glucose metabolism), or were hypoperfused (i.e., had reduced flow compared with glucose metabolism). Those regions that were hyperperfused—including the amygdala, basal ganglia, thalamus, and cingulate cortex—might be necessary for rapid action, like detecting and responding to unexpected and important environmental events. In contrast, the hypoperfused regions might be parts of the cortex associated with complex cognition, such as much of the lateral frontal and parietal lobes. These results, though still requiring further elaboration, argue that the function supported by a region may influence its vascular properties. Note that differential coupling between regions may have consequences for BOLD sensitivity; see the 2008 paper by Ances and colleagues for such evidence.
Thought Question

Why might the brain supply blood to a larger region than that immediately surrounding neuronal activity?

Collectively, these studies help explain the paradox we noted previously, namely that the MR signal increases during neuronal activity even though deoxygenated hemoglobin decreases the MR signal. The disparity between oxygen utilization and oxygen delivery means that more oxygen is supplied to a brain region than is consumed. Fox and Raichle note that this is consistent with the experience of neurosurgeons, who have long observed that the cortex becomes more pink in color when it becomes active. As the excess oxygenated blood flows through active regions, it flushes the deoxygenated hemoglobin from the capillaries supporting the active neural tissue and from the downstream venules. So the BOLD contrast following neuronal activity occurs not because the oxygenated hemoglobin increases the MR signal but because it displaces the deoxygenated hemoglobin that had been suppressing the MR signal intensity (Figure 7.5).

Figure 7.5 Summary of BOLD signal generation. (A) Under normal conditions, oxygenated hemoglobin (Hb) is converted to deoxygenated hemoglobin at a constant rate within the capillary bed. (B) But when neurons become active, the vascular system supplies more oxygenated hemoglobin than is needed by the neurons, through an overcompensatory increase in blood flow. This results in a decrease in the amount of deoxygenated hemoglobin and a corresponding decrease in the signal loss due to $T_2^*$ effects, leading to a brighter MR image. (After Mosley and Glover, 1995.)
The Growth of BOLD fMRI

From the work of Ogawa and colleagues, it was clear that changes in blood oxygenation could be measured using MRI. The next step was to demonstrate that BOLD contrast could be used to localize different functions in the human brain. The first functional studies used simple visual and motor tasks, such as watching a flashing checkerboard or squeezing one’s hand repeatedly. Such simple tasks were not intended to provide new information about the organization of the brain; indeed, the locations of the visual and sensorimotor cortices had been known since the end of the nineteenth century! Instead, the first fMRI studies strived to replicate well-established findings, thus validating the capabilities of the new technique. Before describing these early studies, we must resume our historical discussion from Chapter 1 so that the early fMRI studies can be considered in the context of their times.

Contributing factors

Few scientific discoveries are made in isolation. Most result from a combination of factors, often including external societal influences, that together allow a nascent idea to flourish. The birth of fMRI was no exception (Figure 7.6). Consider that the paramagnetism of hemoglobin had been known for almost a half century before Thulborn and colleagues examined oxygenated and deoxygenated blood using MRI, and even then it would be another decade before the first fMRI studies were published. This slow pace of progress did not

| 1933 | Rabi uses magnetic resonance to measure nuclear magnetic moment |
| 1935 | Pauling and Coryell study magnetic properties of blood |
| 1940 | |
| 1945 | Purcell and Bloch simultaneously discover nuclear magnetic resonance |
| 1970 | Damadian reports differences in relaxation times for biological tissues |
| 1971 | Lauterbur creates first MR image using magnetic gradients |
| 1975 | Mansfield proposes echo-planar imaging |
| 1980 | |
| 1982 | Thulborn and colleagues report the effects of blood oxygenation on T2* contrast |
| 1985 | GE introduces first 1.5-T scanner |
| 1988 | Insurance reimbursements begin for MRI in United States |
| 1988 | Active gradient shielding developed |
| 1990 | Improved head coil designs developed |
| 1992 | Ogawa proposes BOLD contrast as basis for fMRI |
| 1992 | First fMRI studies published |

Figure 7.6 Milestones in the development of fMRI
reflect a lack of interest in the brain—there was considerable interest in scalp recordings of electrical potentials during the 1960s and 1970s and in PET imaging during the 1980s. The extended gestation and subsequent rapid growth of fMRI shown in the timeline resulted in a large part from two external factors, both related to the clinical use of MRI.

First, improvements in pulse sequence design and scanner hardware reduced image collection time from many seconds to a few tens of milliseconds. Early MR imaging was a slow process. The first MR image, for example, was acquired at a rate of more than two minutes per voxel. In modern imaging, in which the brain may consist of many thousands of voxels, such a slow acquisition rate would correspond to approximately one volume per month! As discussed in Chapter 1, the development of echo-planar imaging (EPI), largely through the work of Peter Mansfield in the late 1970s, allowed an entire image to be collected following a single excitation pulse (Figure 7.7). Yet, limitations in the scanning hardware delayed the practical implementation of EPI. Rapid changes in gradient fields, as required by EPI, could induce currents in metal parts within the scanner hardware, introducing artifacts in recorded images. Rather than change scanner hardware to ameliorate this problem, manufacturers adopted other approaches, such as fast low (flip) angle shot, or FLASH, sequences, that did not tax scanner hardware as significantly as EPI. Not until about 1985 were active gradient-shielding techniques developed, largely through the work of Mansfield and colleagues, that incorporated an outer gradient winding in the opposite direction. The outer winding reduced currents in the scanner hardware but added complexity and increased power requirements. Major manufacturers began adding actively shielded gradients to their standard scanner platforms over the following years. By the early 1990s, fast switching gradient technology proposed by Turner, and high linearity developed by Wong and colleagues, together provided the advances needed for EPI to be practical.

The second key factor that contributed to the growth in fMRI was the increasing clinical applicability of structural MRI. Most of the relatively few MR scanners used in the 1970s were devoted to industrial applications, and almost none were being used in hospital settings. The workhorses of diagnostic imaging were computerized tomography (CT) scanners, which allowed clinicians to assess damage to soft-tissue structures (Figure 7.8). The three-dimensional scans were demanded by doctors and patients alike, serving both to draw new patients to hospitals that possessed the latest equipment and to generate new income from the expensive procedures. Despite the capital commitment required (typically more than $300,000), by the early 1980s more than 5000 CT scanners were in use worldwide.

Around the same time, hospitals began considering the use of MRI as a complement to CT scanning, in part due to the enthusiasm of pioneers like Raymond Damadian (see Chapter 1). Several medical device companies, including Damadian’s FONAR Corporation, General Electric (GE), and Varian, developed high-field MRI scanners that promised image resolution that would far surpass that of CT. The first clinical 1.5-T scanner was installed at Duke University by GE in 1982, and this would remain the most common field strength for both clinical and research purposes for more than two decades.
Figure 7.8 Computerized tomography (CT) imaging. (A) CT uses a moving X-ray source to create a three-dimensional map of underlying tissue. While CT imaging can distinguish tissue types, such as gray/white matter from CSF within the brain (B), it is sensitive to the same limitations on resolution and contrast as conventional X-rays. For comparison, a structural MRI image (C) provides much better contrast between many types of tissue.

By 1985, MR scanning was sufficiently well established that insurance companies in the United States began reimbursing for MRI procedures. The cost of MRI scanners was still very high, often as much as $2 million, but hospitals were now able to recoup those costs through many procedures. As had happened previously with CT, this new clinical demand sparked an explosion in the number of MRI scanners. Just 20 years after the introduction of the first high-field scanner, there were more than 10,000 such scanners worldwide. Although most were devoted to patient care during normal business hours, researchers at many institutions were able to use the scanners at night and on weekends for research into brain function. These research studies were often facilitated by supplemental hardware, such as gradient insert coils, that improved on the hardware provided by the clinical manufacturers. The advances that facilitated the first fMRI studies, therefore, were developed largely to meet the clinical demand for structural MRI.

Early fMRI studies

In the early 1990s, many research laboratories were competing to create the first images of brain function using MRI. Some explored the use of exogenous contrast agents (see Box 7.2). Others investigated the new endogenous BOLD
BOX 7.2 Functional Studies Using Contrast Agents

BOLD contrast depends on the paramagnetic properties of deoxygenated hemoglobin, which causes a loss of phase coherence in nearby protons, and is measurable using $T_2^*$ imaging. Yet, the effects of hemoglobin on nearby protons are tiny; even large BOLD effects result in signal changes of only about one percent. Another method for increasing image contrast is to use exogenous contrast agents, highly paramagnetic substances that can be injected into the bloodstream but do not cross the intact blood–brain barrier. Common contrast agents like gadolinium diethylenetriaminepentaacetic acid (Gd-DTPA) are well tolerated by most people, with mild headache and nausea as the most common side effects. Contrast agents have great importance for clinical imaging, especially in the detection of pathological tissue, including brain tumors. Under normal conditions, the diffusion of a contrast agent like Gd-DTPA through the bloodstream will reduce $T_1$ values of hydrogen protons in the blood, increasing signal within blood vessels but not elsewhere. But if there is damage to the blood–brain barrier due to brain pathology, the contrast agent may escape from the bloodstream and enter the surrounding tissue, resulting in increased signal on $T_1$-weighted images.

This effect on $T_1$ relaxation, though clinically important, provides no information about brain function. Rather, it is the effect of contrast agents on local magnetic field homogeneity that enables functional studies. Because the injected contrast agent is highly paramagnetic and has a very strong magnetic moment, it causes a considerable inhomogeneity between the tissue outside the blood vessel and the contrast-enhanced blood within. Remember from Chapter 5 that sharp gradients in magnetic field homogeneity can cause signal losses known as susceptibility artifacts, due to differential effects of the magnetic gradient on spin precession. Pulse sequences sensitive to $T_2^*$ effects can measure the local concentration of the contrast agent over time. Unlike BOLD contrast, which depends on both blood flow and oxygen extraction, exogenous contrast methods generally rely only on blood volume changes associated with functional activity. In addition, they have a limited lifetime due to their passage through the brain and subsequent dispersal through the vascular system. But, because the exogenous contrast agent is much more paramagnetic than deoxygenated hemoglobin, it causes much larger signal changes that can be extracted from a single pass of the agent through the brain.

The first fMRI study to use exogenous contrast (and the first fMRI study of any form) was reported by Belliveau and colleagues in 1991. They measured visual cortex activation using spin-echo EPI at 1.5 T following injection of a bolus of Gd-DTPA. In the test condition, subjects viewed a visual pattern that flashed at a rate of about 8 Hz, and in the control condition there was no visual stimulus. Based on electrophysiological data, this rate was known to robustly activate the primary visual cortex. The authors hypothesized that, following injection of the contrast agent, the raw MR signal would decrease due to increased magnetic susceptibility. Furthermore, the magnitude of this decrease should be greater for active brain regions due to local increases in blood volume. As shown in Figure 1, there was a transient decrease in MR signal associated with passage of the contrast agent through the primary visual cortex. There was a delay of about 8 to 10 s between the injection of Gd-DTPA and the onset of the signal decrease. This delay reflects the time required for the contrast agent to travel from the inject-
The power of exogenous contrast agents comes from the enormous signal changes they induce. Even in this early experiment, the signal change observed due to the contrast injection was about 30%. The difference between the test and control conditions was on the order of 5%. To appreciate the magnitude of such changes, compare them to the much smaller signal changes observed in BOLD fMRI. Furthermore, the clear differences shown in Figure 2 represent two trials, one test and one control, from a single subject. In comparison, modern BOLD studies usually represent data combined from many subjects with many trials per condition.

Despite this power, exogenous contrast agents are rarely used for fMRI research. One limitation can be seen in the time course of signal change. Because of the substantial transit delay of the agent, it is challenging to accurately measure the timing of brain activity. And since a single bolus is used, one measurement of signal change is obtained per trial. If one wants to study two types of trials, like Belliveau and colleagues, two injections are required. In general, as more conditions are added, more injections are needed. These requirements preclude some types of analyses that are possible with endogenous BOLD contrast, such as the sorting of many trials based on accuracy or reaction time, fast event-related designs, evaluation of brain response as a function of stimulus sequence, and complex parametric studies. In addition, many subjects are less likely to participate in studies that require intravenous injection. So, while a number of researchers are actively investigating the use of contrast agents for fMRI (see Chapter 12), especially for studies in animals, nearly all current fMRI studies use endogenous BOLD contrast.

Contrast mechanism developed by Ogawa and colleagues, culminating in three studies published in 1992. The first study, by Kwong and colleagues, used a gradient-echo EPI sequence at 1.5 T to study activation in the visual cortex. They evoked visual cortex activation by alternating 60 s periods of visual stimulation (e.g., the flashing of an LED pattern) with 60 s baseline periods of darkness. At the onset of the stimulation period, there was a sharp increase in MR signal around the calcarine fissure, increasing by about 3% within 10 s (Figure 7.9). The activation increase was sustained for the duration of the visual stimulation period, then receded to the baseline level once darkness returned. Despite its crudeness in both timing and spatial resolution by today’s standards, this study provided the very first example of BOLD fMRI.
These findings were replicated in a similar study published the following month by Ogawa and colleagues, who also evaluated changes in fMRI gradient-echo signals resulting from long-duration (e.g., 100 s) presentations of visual stimuli. Unlike the Kwong study, however, they used a pulse sequence that was limited to an effective TR of about 10 s, and collected data on a high-field (4.0 T) scanner. They also manipulated the TE to show that the BOLD signal depends on $T_2^*$ effects, and not $T_1$, which should be independent of the TE. At a TE of 40 ms, the stimuli caused changes in the BOLD intensity, whereas at a very short TE of 8 ms, the stimulus-related effects disappeared. Almost simultaneously, a third paper was published by Bandettini and colleagues, describing the use of a motor task in which subjects repeatedly touched their fingers to their thumbs for long blocks of time. Data recorded using gradient-echo EPI at 1.5 T showed significant activation in the primary motor cortex.

While most early fMRI studies used long stimulus durations, work reported later in 1992 by Blamire and colleagues examined responses to individual visual stimuli using a spin-echo EPI sequence at 2.1 T. They found that lengthy visual stimuli (10 s to 90 s) generated long-duration and large (about 10%) increases in signal in the visual cortex, similar to the findings of Kwong and colleagues. More remarkable were the results from much shorter-duration stimuli. Even the shortest stimulus (2 s) evoked a significant signal change within the visual cortex (Figure 7.10). The authors noted that there was a short but measurable delay between the stimulus presentation and the MR signal change. On average, the first observable fMRI change in the primary visual
cortex occurred about 3.5 s after the onset of the stimulus. This was the first demonstration of the time course of the BOLD hemodynamic response evoked by a single stimulus event.

**Thought Question**

Why was it critical to observe measurable BOLD activation to short-duration stimuli? What implications does this have for fMRI experimentation?

It is instructive to compare these early studies with current fMRI practices. At first glance, the procedures and equipments seem very similar. The field strengths reported above (i.e., 1.5 T, 2.1 T, 4.0 T) are similar to those of MRI scanners used now, more than fifteen years later. EPI pulse sequences are also still used commonly, though other sequences such as spiral imaging have grown in popularity. Yet, there are subtle differences between the studies described above and modern studies. The gradient coils on early scanners generated only weak magnetic gradients that could not be changed rapidly, constraining the rate of data acquisition (i.e., early studies collected data from only one or a few slices, despite having relatively long TRs). Now, the collection of 20 or more slices per second is commonplace. Even more important are differences in the approach to data analysis. None of these early studies used pre-processing techniques, which will be introduced in Chapter 8, to correct for head motion or physiological variability. (Blamire and colleagues did note that voxels on the edge of the brain showed systematic oscillations in signal intensity, which they attributed to pulsatile motion of the brain associated with the cardiac cycle.) Nor did they use modern statistical approaches based on the general linear model, which will be introduced in Chapter 10. Instead, they used simple statistics to evaluate whether activation in a region of interest was greater during a task condition than during a control condition. While this approach provided adequate power for evaluating very simple visual or motor tasks, answering more complex experimental questions would require more complex approaches. Nevertheless, the basic elements of modern fMRI practice can be traced back to these early studies. They hinted at the future capabilities of fMRI, setting the stage for later research.
Figure 7.11 Schematic representations of the BOLD hemodynamic response. Shown are representative waveforms for the hemodynamic response to a single short-duration event (A) and to a block of multiple consecutive events (B).

**The BOLD Hemodynamic Response**

The change in the MR signal triggered by neuronal activity is known as the hemodynamic response, or HDR (Figure 7.11). Referring to the hemodynamic response is, however, a bit misleading, as its shape varies with the properties of the stimulus and, presumably, with the underlying neuronal activity. We might expect, therefore, that increasing the rate of neuronal activity would increase the amplitude of a hemodynamic response, whereas increasing the duration of neuronal activity would increase the width of the hemodynamic response. Determining the exact relationship between neuronal activity and fMRI activation, however, is complicated by the different dynamics of the two processes (Box 7.3). Cortical neuronal responses occur within tens of milliseconds following a sensory stimulus, but the first observable hemodynamic changes do not occur until 1 s to 2 s later. Thus, the hemodynamic response is said to lag the neuronal events that initiate it. Throughout the remainder of the book, we will make frequent references to different aspects of the hemodynamic response waveform, and so here we define some terms.

Remember that BOLD fMRI, defined simply, measures changes in the total amount of deoxygenated hemoglobin in a voxel over time. Yet, the quantity of deoxygenated hemoglobin depends not just on the extraction of oxygen by active neurons, but also (and more importantly) on changes in blood flow and blood volume that together shape the BOLD hemodynamic response (Figure 7.12). We can summarize the BOLD response as a series of phases. As described below, some studies have reported an initial dip of 1 to 2 s duration in the amount of deoxygenated hemoglobin in the voxel. After a short latency, the metabolic demands due to the increased neuronal activity over baseline levels

Figure 7.12 Relative changes in cerebral blood flow and cerebral blood volume following neuronal activity. This figure shows data from an experiment in which the forepaw of a rat was stimulated for a period of 30 s and the resulting changes in cerebral blood flow (CBF) and cerebral blood volume (CBV) were measured. Note that following the stimulus offset, CBF returns quickly to baseline levels but CBV returns slowly. Elevated CBV relative to CBF causes an increase in the total amount of deoxygenated hemoglobin that is present, which may lead to the poststimulus undershoot in the BOLD signal. (After from Maneville et al., 1999.)
Box 7.3 Neuronal Activity and BOLD fMRI

Both single-unit firing and field potentials provide direct, albeit different, measures of the informational transactions of neurons. While we have emphasized so far in this chapter that these measures complement fMRI data, it is important to recognize that the relationship between neuronal activity and the BOLD response remains unclear. Is there a strong correlation between BOLD fMRI measurements and these direct measures of neuronal activity? How might the integrative and signaling aspects of neuronal activity individually contribute to the BOLD signal? One might expect that both electrophysiological measures would have similar influences on the BOLD response, since the EPSPs that trigger action potentials also generate field potentials. However, summated IPSPs generate field potentials as well, but inhibit the firing of action potentials, so there are clear circumstances in which different electrophysiological measures could have different contributions. We will thus consider signaling and integrative aspects of neuronal activity separately.

Few studies have directly compared electrophysiological and BOLD fMRI measurements. In a landmark 2001 report, Logothetis and colleagues simultaneously recorded fMRI and electrophysiological data from the primary visual cortex.

Anesthetized monkeys viewed a rotating visual checkerboard pattern while being scanned in a 4.7-T scanner using gradient-echo echo-planar imaging. The authors simultaneously recorded three types of electrophysiological activity: single-unit firing of an individual neuron close to the electrode, multi-unit (the collective firing rate of neurons, particularly large pyramidal cells, within a few hundred microns), and local field potentials (synchronous changes in integrative activity from cells within a few millimeters). The results are shown in Figure 1. Single-unit and multi-unit activity occurred transiently at the onset of the stimulus and did not persist over time, while the local field potentials showed both transient and persistent activity. The local field potential activity—which includes both postsynaptic potentials and integrative activity occurring at the soma—better predicted the BOLD signal change than did the multi-unit activity, although the latter

(continued on next page)

Figure 1 The relationship between BOLD activation and neuronal activity. Simultaneous electrophysiological and fMRI data were recorded in monkeys during the presentation of visual stimuli. The time course of BOLD activation evoked by visual stimulation is shown as a solid green histogram, while the time course of multi-unit activity (MUA) is shown in purple, the time course of single-unit activity (SUA) is shown in blue, and the time course of local field potentials (LFP) is shown in red. The duration of visual stimulation, indicated by vertical black bars, varied from 24 to 12 to 4 s, shown in the top, middle, and bottom panels, respectively. Note that the BOLD activation and the local field potential activity are extended in time throughout stimulus presentation, while the single- and multi-unit activities rapidly return to baseline. These results suggest that postsynaptic activity that generates local field potentials may be a primary contributor to the BOLD response. (After Logothetis et al., 2001.)
still provided some predictive information. Moreover, as shown in a 2008 study by Goense and Logothetis, these results hold when measured in awake monkeys viewing simple visual stimuli: measured hemodynamic responses tracked the sustained changes in local field potentials, but not the more transient changes in single- and multi-unit activity. Data from several studies indicate that local field potential activity in the frequency range of about 20–60Hz best predicts the BOLD signal.

These and other results have led to a simple conclusion: the BOLD contrast mechanism reflects primarily the input and intracortical processing in a given area, that which we have characterized as the integrative aspect of neuronal processing, rather than the output reflected in action potential firing. Given this result, recall from the previous chapter our discussion of Attwell and Laughlin’s energy budget for the brain. On the basis of the large number of synapses per neuron in primates, they hypothesized that compared with lower animals, a greater proportion of the brain’s energy budget would be required to restore postsynaptic concentration gradients. Integrative activity may be the best predictor of the current and future metabolic demands of a region. However, it must be emphasized that post-synaptic activity may represent, to a large extent, the inputs to a given brain region. Thus, its amplitude may depend on computations elsewhere in the brain, not necessarily local information processing.

Due to technical limitations, notably the invasiveness of electrode recording, relatively little data of this type has been collected from human subjects. One notable exception comes from a 2005 paper by Mukamel and colleagues, who recorded single-unit activity from patients with implanted electrodes and fMRI data from neurologically normal subjects, while each individual watched and listened to an extended clip from a movie (see Box 11.1 for a discussion of this approach). From the electrode recordings, the authors measured changes in local field potentials and firing rate over time in the primary auditory cortex. (Note that because of the complexity of their auditory stimulus, the firing rate and the local field potentials were highly correlated, precluding the authors from separating these as individual contributors to the BOLD response.) Then, they convolved the neuronal activity with a standard hemodynamic response function to obtain a predicted BOLD timecourse, which they used as a regressor in the analysis of their fMRI data from the normal subject group. The resulting activation map revealed activation localized to the primary auditory cortex, exactly replicating the location of electrophysiological activity in the other subject group. Another study, reported in 2004 by Huettel and colleagues, examined simpler visual stimuli that varied in duration, also in separate groups of subjects: samples of patients with implanted electrodes, and neurologically normal fMRI subjects. Their results suggested that stimulus duration had corresponding effects on electrophysiological and fMRI signals, at least within the primary visual cortex, although somewhat different effects were reported for other brain regions.

Despite the data showing the remarkable correspondence between neuronal activity and BOLD fMRI, largely from simultaneous recordings in monkeys, some key challenges remain for future work. The slow temporal resolution of the BOLD response precludes the separation of feed-forward and feedback processing within a region. If the BOLD signal is indeed most sensitive to input-related activity, then complex and temporally sustained feed-back processing could contribute more to the signal than initial first-pass information processing. Moreover, both excitatory and inhibitory post-synaptic activity can contribute to the BOLD signal. In principle, a large BOLD signal could be observed in a region in which these two types of activity are relatively balanced, such that the region’s output (and thus its contribution to thought and behavior) remains unchanged. Finally, there are likely to be situations where the amount of information processing might not be correlated with metabolic demand. Suppose that you observe that BOLD activation in a region (e.g., the prefrontal cortex) is evoked by a complex cognitive task (e.g., maintaining information in working memory), and furthermore the amplitude of that activation correlates positively with task performance. You might then conclude that individuals have good working memory if they engage the prefrontal cortex. But, what if the opposite relationship were observed: less activation in individuals with the best performance? At first consideration, this may seem implausible, given the concepts raised throughout this chapter. Yet, strong support exists for this possibility; for example, activation within a brain region tends to decrease with task practice, while the practice results in improvements in performance. One plausible hypothesis for such effects is that the brain represents the needed computations more efficiently, requiring fewer neurons to accomplish the same processing. Future research will be necessary to extend the results obtained so far, in order to clarify our understanding of the relationships between BOLD fMRI and neuronal activity in different brain regions, forms of information processing, and subject populations.
result in an increased inflow of oxygenated blood. More oxygen is supplied to the area than is extracted, and this results in a decrease in the amount of deoxygenated hemoglobin within the voxel. If we monitor the voxel’s activation using BOLD fMRI, we find that its signal increases above baseline at about 2 s following the onset of neuronal activity, rising to a maximum value at about 5 s after a short-duration stimulus. This maximum is known as the peak of the hemodynamic response. If the neuronal activity is extended in time, the peak may be similarly extended into a plateau, typically maintained at a slightly lower amplitude than the peak (see Figure 7.12B for a typical example).

After the neuronal activity has ceased, the BOLD signal decreases in amplitude to a below-baseline level and remains below baseline for an extended interval. This effect, known as the poststimulus undershoot, has been attributed to both biophysical and metabolic effects. Some biophysical models, such as the balloon model advanced by Buxton and colleagues, postulate that neuronal activation evokes an inflow of blood that is initially greater than its outflow. The result is an increase in blood volume so that the venous system expands like a balloon. Then, following cessation of neuronal activity, blood flow decreases more rapidly than blood volume. While volume remains elevated but flow has returned to baseline, a greater amount of deoxygenated hemoglobin will be present. This reduces the overall fMRI signal below baseline levels. As blood volume slowly returns to normal levels, the fMRI signal will similarly increase to baseline, ending the undershoot. However, other data indicate that changes in oxygen metabolism play a key role. For example, Lu, van Zijl, and their colleagues have hypothesized that a continuously elevated cerebral metabolism would lead to the poststimulus undershoot. Supporting this perspective, a study published in 2008 by Harshbarger and Song investigated whether the application of diffusion weighting, which should minimize large-vessel contributions, influenced the amplitude of the BOLD undershoot.

Targeting voxels in the visual cortex that were activated by a flashing visual stimulus, they found that diffusion weighting modulated the undershoot amplitude in some voxels, but not others. Still other voxels, primarily those in higher-order visual regions, exhibited no undershoot at all. These results, along with similar data from other groups, are difficult to reconcile with a purely vascular explanation, but instead suggest that sustained metabolic demand is an important contributor to the poststimulus undershoot.

To illustrate a sample hemodynamic response, Figure 7.13A provides data from a single voxel showing its change in MR signal over time, or its time course. In this experiment, the subject squeezed both hands whenever a brief flashing checkerboard was presented. There were long intervals between the stimuli so that there would be time for the hemodynamic response to return to baseline. Each line in Figure 7.13B presents the change in MR signal within a single voxel over a 21 s epoch, beginning 3 s before the stimulus through to 18 s after. Immediately apparent is the variability in the response over time. Even for very robust responses, the noise in the data has an amplitude similar to that of the hemodynamic response, making it difficult to identify the exact response evoked by each presentation of the stimulus (see Chapter 8 for an extended discussion of this problem). However, as data are combined from many evoked responses, a pattern similar to that shown in Figure 7.11A emerges.

The initial dip

Thus far, we have emphasized that neuronal activity leads to an increase in the BOLD signal. However, some research suggests that this positive change is preceded by a smaller decrease in the MR signal, which is often called the initial dip. Data from several studies using optical spectroscopy in animals peak The maximal amplitude of the hemodynamic response, occurring typically about 4 to 6 s following a short-duration event. undershoot The decrease in MR signal amplitude below baseline due to the combination of reduced blood flow and increased blood volume. balloon model A model of the interaction between changes in blood volume and changes in blood flow associated with neuronal activity. time course The change in MR signal over a series of fMRI images. epoch A time segment extracted from a larger series of images, usually corresponding to the period in time surrounding an event of interest.
Figure 7.13 Examples of the BOLD hemodynamic response. (A) A sample fMRI time course from a single voxel in the motor cortex during a task in which the subject squeezed her hand for 2 s every 16 to 18 s. Even though this is a very high signal-to-noise task with clear responses present following every stimulus event, there remains substantial variability in the amplitude and form of the hemodynamic response over events. (B) Data from the individual events that make up (A). Note the variability from event to event, although all have generally similar hemodynamic responses that peak about 5 to 6 s following the hand squeezing. Three randomly selected individual trials are highlighted in color for clarity.

found a rapid increase in deoxygenated hemoglobin that had good spatial correspondence with active neurons, at least initially. In 1995 Menon and colleagues attempted to identify a similar phenomenon within the fMRI hemodynamic response. They presented a visual pattern that flashed for 10 s while collecting data using high-field (4.0 T) and fast-rate (TR of 100 ms) echo-planar imaging, with a local surface coil to increase detection power in the visual cortex. They found that some activated voxels showed an initial reduction in signal that had less than half the amplitude of the positive hemodynamic response (Figure 7.14). Those voxels were found within gray matter along the calcarine sulcus, which corresponds to the primary visual cortex. However, the activated region associated with the later positive response was more spatially diffuse and extended into neighboring veins and white matter.
To demonstrate the spatial specificity of the initial dip, other investigators have used experimental designs that evoke highly specific activation patterns. In a study conducted in 2001, Duong and colleagues examined the BOLD response in the primary visual cortex of anesthetized cats using high-field (4.7 T and 9.4 T) fMRI. They targeted orientation columns, which are small, vertically organized collections of neurons that respond preferentially to stimuli of a particular orientation (e.g., lines). By collecting data at very high spatial resolution (about 150 μm²), the researchers could identify individual orientation columns, which they expected would be differentially activated by stimuli with perpendicular orientations (e.g., 45° and 135°). They found that the initial dip showed good spatial specificity, in that voxels with an initial dip in response to one orientation did not show an initial dip in response to the stimuli in the perpendicular orientation. In contrast, the later positive BOLD response was blurred over the columns, such that most voxels showed positive BOLD activation in response to both orientations. The authors noted that the spatial specificity of the BOLD response was greatest over the first 2 s and decreased over time. This result suggests that the initial dip may reflect decreased oxygenation in nearby capillaries, while the positive response reflects overcompensatory delivery of oxygenated blood within the surrounding venous drainage system.

Despite these intriguing results, the initial dip is a controversial topic for a simple reason: despite many examples of its existence, including several dozen well-controlled studies in anesthetized animals, it is not reported in the vast majority of fMRI studies. (See the 2004 review article by Anes for an extended discussion.) So what could explain its infrequent observation? One likely factor is the rarity of high-field MR scanners for functional studies. The amplitude of the initial dip seems to scale dramatically with field strength. When measured at 1.5 T, the initial dip was only 12% of the magnitude of the positive response, which is only one-third of the proportion measured at 4.0 T. This result is consistent with the idea that the initial dip has a microvascular origin, since signals recorded from small blood vessels should scale more dramatically with field strength than signals recorded from large blood vessels. Another possible factor is that averaging over a large spatial region or over an extended time period obscures the smaller-scale effects of the initial dip. Highly specialized image acquisition methods, like those reported in a 2008 article by Lindquist and colleagues, may be useful for isolating the rapid but small-scale negative BOLD changes.
While the evidence in support of the initial dip has strengthened in recent years, its underlying mechanisms still require further investigation. For example, even though the oxygen extraction hypothesis remains the best theoretical interpretation, a transient decrease in flow or increase in volume would also result in a decrease in the BOLD signal. Furthermore, the specific relationships between neuronal activity and metabolic changes also require further elucidation. As suggested in a 2007 experiment by Li and Freeman, the increase in deoxygenated hemoglobin that may underlie the initial dip may exhibit different patterns of non-linearity than those of the positive response. We emphasize that the physiological changes that underlie the initial dip are themselves well-supported by a wealth of evidence from multiple imaging and optical recording techniques. What remains (somewhat) controversial is whether those physiological changes affect the BOLD signal and, assuming they do, under what conditions an initial BOLD decrease would be observed.

Spatial Resolution

The spatial resolution of an fMRI study, or its ability to distinguish differences between nearby spatial locations, depends on several factors. One straightforward influence is the voxel size. Voxels are three-dimensional rectangular prisms whose dimensions are specified by three scanner parameters: field of view, matrix size, and slice thickness. The field of view describes the extent of the imaging volume within a slice and is generally expressed in centimeters. The matrix size determines how many voxels are acquired in each dimension, within that slice. Most studies use matrix sizes that are symmetric powers of two, such as $64 \times 64$ or $256 \times 256$, to facilitate the use of the fast Fourier transform for image reconstruction. So, for a field of view of 24 by 24 cm and a matrix size of $64 \times 64$, the resulting within-slice (in-plane) voxel size would be 3.75 mm by 3.75 mm. Slice thickness provides the third dimension (through-plane) and is generally the same or larger than the in-plane voxel size (e.g., 5 mm). When the slice thickness is equal to the in-plane resolution, the voxels are cubic and the spatial resolution is said to be isotropic.

The size of the voxel used in a study may depend on the research question. Studies that examine the entire brain will use relatively large voxels, often around 4 to 5 mm on each side. In contrast, studies that examine a single brain region, such as the visual cortex, may use smaller voxels of 1 to 2 mm. For comparison, the letters making up this sentence are each about 2 to 3 mm in size. It is common to acquire anatomical images with smaller voxel dimensions (e.g., 1 mm by 1 mm in-plane), and functional data are often displayed on these high-resolution anatomical images. Note that the actual spatial resolution of the fMRI data is not affected by the resolution of the images on which it is displayed.

It seems obvious that increased spatial resolution would carry advantages for fMRI studies: more voxels within a brain region should improve the ability to distinguish boundaries between neighboring functional areas. So if fMRI data can be collected at any voxel size, why do researchers not always use the smallest possible size? There are two primary challenges for using small voxels in fMRI: reduced signal compared with noise and increased acquisition time.

First, variation in the BOLD signal depends on the change in the total amount of deoxygenated hemoglobin within a voxel. So, if we reduce the voxel volume by a factor of two, the BOLD signal changes will be half as large, resulting in a smaller signal-to-noise ratio. In some brain regions, such as the primary motor or visual cortex, a finger flexion or a visual flash may evoke a very

---

**spatial resolution** The ability to distinguish changes in an image (or map) across different spatial locations.

**isotropic** Having similar properties in all directions.
large brain response, and thus the reduced signal amplitude generated from small voxels may not be a problem. But in other brain regions, like the frontal lobe, a complex cognitive task may evoke much smaller changes in neuronal activity, and larger voxels may improve the ability to detect those changes in the presence of noise. We will discuss the concepts of signal and noise in more detail in the next chapter.

Second, as voxel size decreases, the time needed to acquire a given volume of the brain increases. While slice acquisition rates will vary between scanners and pulse sequences, increasing the within-image resolution can double or even quadruple acquisition time. Increasing the through-plane resolution also increases acquisition time, in this case exactly in proportion with the resolution change. If one halves the slice thickness, then twice as many slices will be needed to cover the same volume, and twice the acquisition time will be required. For example, a typical scanner might acquire 16 slices per second at 4 × 4 × 4 mm resolution, but only 1 slice per second at 1 × 1 × 1 mm resolution. The vertical extent of the cerebrum is about 110 mm, so 27 4-mm slices that cover most of the brain could be acquired in less than 2 s. At 1 mm resolution, that same volume would take nearly 110 slices and two minutes to acquire. Though extreme, this example points out the difficulty in using high spatial resolution for full-brain functional imaging. And, even if possible in theory, the long data acquisition periods needed for high-resolution images can cause $T_2^*$ blurring. Remember from Chapter 4 that data acquisition (e.g., filling k-space) takes time; this can be only a few tens of milliseconds for a typical 64 × 64 image but is much longer for very-high-resolution images. During this acquisition period, the spins are continuously undergoing $T_2^*$ decay, so if the acquisition period is very long compared with the $T_2^*$ value of the tissues being imaged, there will be virtually no signal for k-space locations collected toward the end of the acquisition period, which results in blurred BOLD images.

Conversely, using voxels that are too large can also reduce detection power. All fMRI studies, especially those using relatively large voxels, suffer from partial volume effects. Even the smallest voxel may contain multiple tissue types, each contributing differently to the total MR signal from that voxel. Figure 7.15

**Figure 7.15** Partial volume effects. A single voxel may contain many different types of tissue, including gray matter, white matter, cerebrospinal fluid, or blood vessels. The MR signal recorded from that voxel is the sum of signals recorded from all of these different tissue types. So, if a voxel on a $T_1$ image contains 25% cerebrospinal fluid (with low signal), 50% gray matter (with medium signal) and 25% white matter (with high signal), the MR signal recorded from the voxel will contain contributions from all three, potentially taking an intermediate value.
shows the possible contents of a single 4 × 4 × 5 mm voxel. A typical voxel within the brain consists mostly of neurons (and their processes, like axons) and glia, with only about 3% of the volume made up of blood vessels. That voxel could include a few million neurons and some tens of billions of synapses, all of which contribute to the combined metabolic demand and thus the total BOLD signal from the voxel. See the 2008 article by Logothetis in the Suggested Readings for further calculations and discussions. In addition to the active neurons of interest and the local capillary bed, there may be other brain tissue that does not contribute to the measured activation. For example, voxels on the edge of the brain can contain gray matter, white matter, and cerebrospinal fluid. Only the gray matter will contribute to the BOLD signal, but protons in the other tissue types may contribute to the noise.

**Spatial specificity in the vascular system**

The functional resolution of fMRI depends on more than voxel size; it also depends on the concordance of neuronal activity and vascular responses to that activity (see Box 7.3). In an important study conducted in 2001, Logothetis and colleagues investigated the correspondence between the BOLD signal and electrophysiological measures, using simultaneous recordings in the primary visual cortex of monkeys. They found good spatial correspondence between BOLD changes and local field potentials that reflect summed EPSPs and IPSPs. Similarly, a 2000 study by Disbrow and colleagues used both fMRI and separate microelectrode recordings to map hand and face representations in the somatosensory cortex of the monkey. The maps of activation produced by the two methods largely overlapped, although the areas of activation were somewhat larger for fMRI compared with the more precise electrode recordings (Figure 7.16). This discrepancy was attributed to the filtering effects of the vascular system. These and related studies demonstrate the limitations in the spatial resolution of BOLD fMRI: in the best case, it can provide information about the capillaries closest to active neurons, but in the worst cases, its signal can come from larger and more distant blood vessels.

How might different parts of the vascular system contribute to the BOLD effect? Recall that the BOLD signal depends on the amount of deoxygenated hemoglobin, which is present in capillaries and veins but absent in fully oxygenated arterial blood. Because deoxygenated hemoglobin molecules are paramagnetic, they create magnetic field gradients within the vessel that extend into the surrounding tissue. The primary mechanism for the BOLD signal is the dephasing of spins within water molecules as they diffuse through these gradient fields. The spins located within the vessel itself give rise to the intravascular component of the BOLD signal, while the spins located in the surrounding tissue (i.e., parenchyma) give rise to the extravascular component of the BOLD signal. In a typical fMRI experiment using gradient-echo sequences, the BOLD signal reflects both intravascular and extravascular signal sources.

Both of these signal sources can arise from capillaries that are adjacent to and perfuse the active neurons. Capillaries are very small, typically less than a millimeter in length, and are generally separated from each other by tens to hundreds of microns. This suggests that the absolute lower limit for functional resolution of any hemodynamic measure of brain activity is a few hundred microns. Even if the brain were sampled at a finer spatial resolution, the functional resolution would not improve. Nevertheless, given the much coarser resolution of most current fMRI research, improving the spatial specificity of BOLD contrast would be of considerable value.
Yet, the BOLD signal is not restricted to nearby capillaries. As described earlier in the chapter, as the flow of oxygen-rich blood increases dramatically in response to neuronal activity, a smaller proportion of oxygen molecules are extracted as fuel. The remaining hemoglobin-rich blood enters the venous system, displacing deoxygenated hemoglobin and increasing the BOLD signal downstream from the active neurons. This effect led to the question “Brain or vein?” that was posed by Frahm and colleagues in 1994, and is a vitally important issue for fMRI studies. Many reported areas of activation may be a consequence of venous drainage, not local neuronal activation, as any researcher who has found significant activation in the superior sagittal sinus will attest. Large-vessel effects associated with large draining veins can compromise studies that require high functional resolution.

Several characteristics are indicative of large-vessel effects. The simplest is magnitude of signal change. Veins have much greater volume, and thus larger potential BOLD changes, than capillaries. Systematic changes in the phase of the MR signal may also indicate large-vessel effects, since large vessels have specific orientations within a voxel, unlike capillaries, which are randomly oriented (see the 2002 article by Menon for an approach to overcoming this prob-

large-vessel effects. Signal changes in veins that drain a functionally active region but are distant from the neuronal activity of interest.
spin-echo (SE) imaging. One of the two primary types of pulse sequences used in MRI, it uses a second 180° electromagnetic pulse to generate the MR signal changes that are measured at data acquisition.

**diffusion weighting.** The application of magnetic gradients to cause changes in the MR signal that are dependent upon the amplitude and/or direction of diffusion.

Voxels containing draining veins will have reduced functional specificity, since they may be downstream from multiple functionally distinct populations of neurons. Finally, the initial dip that is sometimes observed at high field is thought to represent oxygen extraction in the capillaries, so it will not be seen within voxels containing large vessels that are distant from the active neurons.

Advanced acquisition techniques can minimize the BOLD signal originating from large vessels. These techniques take advantage of the different magnetic properties of large- and small-caliber vessels and the different diffusion properties of extravascular and intravascular spins. The magnetic field generated by the deoxygenated hemoglobin in large vessels changes slowly over space as it extends into the surrounding tissue and fluid (Figure 7.17A). Thus, nearby extravascular spins within diffusing water molecules experience relatively small magnetic field changes as they travel. Indeed, within the few tens of milliseconds that are typical for a BOLD fMRI acquisition period, the deoxygenated hemoglobin within large vessels can be approximated as having a constant pattern of field inhomogeneity. Spin-echo sequences (see Chapter 5) can reverse the loss of phase coherence, and therefore can eliminate the BOLD signal that arises from the extravascular compartments of large vessels. The situation is quite different for small vessels (Figure 7.17B). They generate steeper magnetic field gradients, compared with the diffusion distance of nearby water molecules, in the surrounding parenchyma. The lost phase coherence caused by these magnetic field inhomogeneities cannot be completely refocused by a 180° pulse, and thus spin-echo sequences retain their sensitivity to the small-vessel extravascular component of the BOLD signal.

Spins within diffusing intravascular water molecules also experience dynamic magnetic field inhomogeneities, because their diffusion distance is large compared with the local magnetic field gradient, for both large and small vessels. For this reason, spin-echo sequences are still sensitive to the intravascular BOLD signal, and another approach must be employed to eliminate those effects. Because these intravascular spins are in flowing blood, they have a higher mobility (especially within large vessels), and thus motion-weighted image acquisition techniques like diffusion-weighted imaging (see Chapter 5).
can selectively suppress the intravascular component of large vessels. The combined use of spin-echo and diffusion-weighted sequences can eliminate the signals from large vessels while preserving the small-vessel signals most critical for functional resolution. Yet, this approach is hardly a panacea: it leads to greatly reduced BOLD sensitivity. For this reason, the introduction of refocusing pulses to eliminate large-vessel effects may be practical only at very high fields, where the increased overall signal can overcome the lost sensitivity to functional changes.

**What spatial resolution is needed?**

The right spatial resolution for an experiment depends on the question being asked. If you are a neurologist examining the effects of frontal lobe damage on intelligence tests, you may examine lesions that span 5 centimeters or more. If you are an electrophysiologist recording the firing of layer 4 neurons within the parietal lobe, you may need to localize the tip of your microelectrode within a few hundred microns of the area of interest. As shown in Table 7.1 (see also Figure 1.8), the properties of the human brain span about seven orders of magnitude from large-scale anatomy to small-scale microbiology. The spatial resolution of fMRI is intermediate between these extremes, and fMRI is most suited for examining a spatial range from millimeters to centimeters. Many aspects of brain function vary over this spatial range. Brain regions identified by cytoarchitectonic features, such as those identified by Brodmann in 1909, generally are several centimeters in size. Individual functional regions within the visual cortex extend from a few millimeters to a centimeter or more, although the entire visual pathways span several centimeters. Subcortical nuclei such as the caudate, putamen, and thalamus are all sufficiently large to encompass multiple fMRI voxels. Nevertheless, many brain structures, including the horizontal cortical layers and the vertical cortical columns (see the ocular dominance columns illustrated in Figure 7.18), exist on a much smaller scale and are difficult (but not impossible) to elucidate functionally using fMRI.

| TABLE 7.1 Different Spatial Scales in the Human Brain |
|----------------|----------------|
| Structure      | Scale          |
| Brain          | 100 mm         |
| Gyri           | 10 mm          |
| Dominance column | 1 mm        |
| Neuron         | 0.01 mm        |
| Synapse        | 0.001 mm       |
| Ion channel    | 0.00001 mm     |

**Figure 7.18 Ocular dominance columns in the visual cortex.** Neurons in the primary visual cortex are organized into very small (~1 mm) columns that are sensitive to information coming from one eye. The same subject participated in two fMRI sessions, shown in panels A and C, that mapped the relative sensitivity of each visual cortex voxel to stimulation from each eye. Note that the outlines of the areas of ocular dominance from the first session (B) correspond well to the results from the second session (D). (From Cheng et al., 2001.)
spatial smoothing  The blurring of fMRI data across adjacent voxels to improve the validity of statistical testing and maximize functional SNR, at a cost of spatial resolution.

normalization  The transformation of MRI data from an individual subject to match the spatial properties of a standardized image, such as an averaged brain derived from a sample of many individuals.

region-of-interest (ROI) analyses  Evaluations of hypotheses about the functional properties of brain regions (i.e., aggregated over a pre-determined set of voxels), often chosen to reflect a priori anatomical distinctions within the brain.

temporal resolution  The ability to distinguish changes in a signal (or map) across time.

While this discussion has focused on the effects of data acquisition methods on spatial resolution, choices made in experimental analysis are also important. A common preprocessing step explicitly reduces spatial resolution by smoothing fMRI data using a three-dimensional Gaussian filter of several voxels in width (see Chapter 8). Typical smoothing parameters can increase the effective voxel size to $6 \times 6 \times 6$ mm or greater. Note that such a voxel contains more than three times the volume of a voxel $4$ mm on each side, and 27 times the volume of a voxel that is $2$ mm per side. While smoothing can reduce spatial resolution, it can improve the validity of statistical tests and comparisons between subjects. Other analysis steps also reduce spatial resolution, through more subtle effects. Algorithms for transforming subjects to a common stereotaxic space, a process known as normalization, further reduce spatial resolution due to the difficulty in matching a person's individual anatomy to a stereotaxic template. Moreover, combining data from multiple subjects introduces spatial blurring associated with functional differences between the subjects. Using anatomically based region-of-interest analyses changes the basic spatial unit from a single voxel to a region containing many voxels, greatly reducing spatial resolution. However, if the chosen regions accurately map onto functional divisions within the brain, the functional resolution of the data may be greatly increased by averaging many similar voxels. For example, the putamen is a relatively small structure within the basal ganglia that is associated with motor preparation, interval timing, learning, and some cognitive processes. Since there are clear anatomical divisions between the putamen and the surrounding white matter, it is simple to create a region of interest that includes the entire structure. This will prevent the identification of functionally distinct subregions (e.g., medial vs. lateral), but may increase the ability to detect changes in the putamen as a whole. As a general rule, many analysis steps sacrifice spatial resolution in order to increase functional resolution.

Temporal Resolution of fMRI

For many experimental questions, it is important to measure the timing of brain activity with accuracy, or with high temporal resolution. Neuroscience techniques differ dramatically in their ability to assess the relative timing of events (see also Figure 1.8). Recordings from microelectrodes within the brain can identify the firing of a single neuron as it occurs, resolving activity in time to the millisecond level, but these can only be made in nonhuman animals or (rarely) in humans who are undergoing special tests associated with neurosurgical procedures. Lesion studies, drug manipulations, and even some imaging techniques like PET provide almost no information about the timing of brain activity. Functional MRI has an intermediate level of temporal resolution, because it can discriminate events that are separated by a few seconds. Just as the basic sampling unit for spatial resolution is the voxel, the basic sampling unit for temporal resolution is repetition time, or TR. Depending on the experiment, the TR may vary from very short (e.g., 500 ms) to very long (e.g., 3000 ms), with even more extreme values used in specialized experiments. While the duration of the TR contributes to the temporal resolution of an experiment, it is not the only factor.

The fMRI BOLD hemodynamic response rises and falls over a period of more than 10 seconds, even if the duration of neuronal activity is very short (e.g., less than a second). So, when we collect fMRI data, we do not take snapshots of neuronal activity, but estimate that activity based on slower changes
in the vascular system. Decreasing the TR to better sample the fMRI hemodynamic response improves our estimate of these vascular changes, which in turn improves the inferences we can make about neuronal activity.

We emphasize this framework because it suggests that there may be a preferred temporal resolution for a given experimental question. Consider the very simple event-related design in which a subject squeezes her hand whenever she sees a visual stimulus. To determine whether an area of the brain becomes active due to hand motion (i.e., detection of the active region), a relatively slow sampling rate will suffice. At a 3-s TR, the hemodynamic response may be easily identified when compared with the prestimulus baseline, but its exact shape may be difficult to estimate (Figure 7.19A). Halving the TR to 1500 ms improves our estimates of the shape and timing of the hemodynamic response but does not substantially change the measured amplitude (Figure 7.19B). Something very interesting becomes evident if we shorten the TR, to 750 or even 375 ms (Figure 7.19C and D): the measured hemodynamic response, though sampled much more often, does not appreciably change.

Figure 7.19 Effects of sampling rate (TR) on the measured hemodynamic response. In each figure, an idealized hemodynamic response is sampled at a different rate.
Thought Question

How does the temporal resolution that is needed to detect significant activation change for long-interval blocked designs? Is the required TR larger or smaller than for event-related designs?

It seems counterintuitive that dramatic increases in sampling rate would have little effect on temporal resolution. However, think about how additional samples might change our estimate of the hemodynamic response. In the case of a 3-s TR, what do we know about the time points between samples (i.e., 1.5 s, 4.5 s, etc.)? Although they are not measured directly, a reasonable assumption might be that the hemodynamic amplitude at these points would be intermediate between those of the recorded samples. For now, we can consider simple linear interpolation, such that the midpoint would be given by the average of the two adjacent samples. At TRs greater than about 2 s, linear interpolation does not provide a good estimate of the values that would have been recorded from the intervening points. But as we shorten the TR to 1.5 s or less, even simple linear interpolation will accurately reproduce intermediate values, because the hemodynamic response has reproducible structure. The changes in blood flow and oxygen extraction that form the basis of BOLD contrast occur as a result of slow physiological processes. Only if these processes varied wildly within short intervals, say 100 ms, would increasing the sampling rate be critical. Note that this example shows the results of an event-related design, so that the hemodynamic response evolves over about 10 to 15 s. If a long-interval blocked design were used, then the changes in the hemodynamic response would be much slower and an even longer TR would be adequate. In summary, the temporal resolution of fMRI is determined both by the repetition time, TR, and by the limitations of the vascular system. For many experimental questions, TRs of about 1 to 2 s are sufficient.

Moreover, using very short TRs introduces some problems. In Chapter 3 we learned that one parameter of an MR pulse sequence is the flip angle, which reflects how far the net longitudinal magnetization is tipped toward the transverse plane by an excitation pulse. Since the amount of MR signal is proportional to the projection of the magnetization vector on the transverse plane, large flip angles provide greater MR signals. For typical gradient-echo sequences with long TRs (i.e., greater than about 2 s), a flip angle of 90° can be used to recover maximal MR signal. But at shorter TRs, a smaller flip angle is required so that the magnetization will reach a steady state over repeated excitations. As a result, the amplitude of the transverse magnetization following excitation will be reduced, and less MR signal will be measured. Short repetition times also reduce spatial coverage. If a scanner can acquire 14 slices per second with a given pulse sequence, then only 7 slices could be acquired with a 500-ms TR, while 28 could be acquired with a 2000-ms TR.

Temporal resolution can be improved by using an interleaved stimulus presentation, in which the experimental stimuli are presented at different points within a TR during different trials. Note that this should not be confused with interleaved slice acquisition, which refers to the order of slice excitation within a TR. Figure 7.20 illustrates the basic approach of interleaved stimulus presentation. In a typical experiment with a 3-s TR, the experimental stimuli might be presented at TR onset, so the hemodynamic response is sampled at 3 s, 6 s, 9 s, etc., following stimulus presentation. In an interleaved design with three presentation times, the stimuli could be presented either at TR onset, one sec-
ond into the TR, or two seconds into the TR. Thus, one-third of the trials would be sampled normally, one-third would be sampled at 1 s, 4 s, 7 s, etc., and one-third would be sampled at 2 s, 5 s, 8 s, etc. By combining data from all three sets of trials, the hemodynamic response can be estimated with a temporal resolution of 1 s. Interleaved presentation can therefore provide improved temporal resolution without limiting spatial coverage or reducing signal amplitude, and is thus an attractive option for many studies. Its primary disadvantage, however, lies in the reduction in the number of trials conducted for each delay condition, which reduces the precision of the estimated hemodynamic response. Researchers must always balance improvements in temporal resolution against possibly diminished spatial coverage, spatial resolution, or experimental power.

What temporal resolution is needed?

To understand the use of fMRI in studying the timing of mental processes, it is necessary to appreciate the different time scales over which such events occur. Imagine that you are driving a car and must quickly swerve to avoid an obstacle in the road (Figure 7.21). Within a millisecond or so after the image of the obstacle hits your eye, photoreceptors in the retina begin to release neurotransmitters. Over the next few milliseconds, those neurotransmitters influence the activity in adjacent bipolar neurons, which in turn evoke action potentials in retinal ganglion cells that project to the lateral geniculate nucleus of the thalamus (as well as to a few other targets). Transmission of visual information through the thalamus to the primary visual cortex requires a few tens of milliseconds, and significant changes in neuronal activity can be detected in secondary visual areas after about 100 ms. Yet, the reaction time before you move the steering wheel has a lower limit of about 200 ms. (For comparison, in Olympic track events, runners who leave the starting blocks within 100 ms of the starter’s pistol are penalized, because there has not been enough time for them to react to the stimulus.) By about 500 ms, you become aware of the obstacle and begin reflecting on the near accident. More complex processes, like retrieving the memory of a similar event, may take several seconds, and changes in your emotional or physiological state may last for minutes or hours.

Figure 7.21. Timing of mental events. When a person makes a simple motor response, like turning a steering wheel to avoid an obstacle (A), a large number of different brain regions will become increasingly active. However, across these regions, the timing of activation may vary considerably (B). Neuronal activation may be present within about 100 ms of stimulus presentation, in regions that support basic sensory processing (e.g., the lateral geniculate nucleus, LGN; primary visual cortex, V1; motion-sensitive cortex, V5; parietal cortex, PC; motor regions, SMA/M1). But in regions supporting more complex cognitive functions (i.e., the prefrontal cortex), activation may persist for tens of seconds.
Finally, there may be long-term effects, such as learned changes in your driving patterns, that persist for days, months, or even indefinitely.

A single stimulus can thus evoke changes in the brain that span more than eight orders of magnitude in time, from milliseconds to days. While this range is very large, most psychological experiments manipulate cognitive processes that occur over periods of a few seconds, well within the range of fMRI. For example, studies examining executive processing, long-term memory, or decision making typically present stimuli every few seconds, while experiments of working memory may require the subject to maintain a stimulus in memory for about 10 s. It is more difficult to use fMRI to study topics involving very small or very large time scales. A rough estimate of the lower limit of fMRI’s temporal resolution for most conditions is a few hundred milliseconds. For example, there are well-known reentrant circuits running from higher visual processing regions back to the primary visual cortex. Attending to a visual stimulus has been shown to increase fMRI activation in the primary visual cortex, but that activation might be attributed to the initial neuronal activity or to feedback from other brain regions that occurs hundreds of milliseconds later.

**Thought Question**

How could experiments be designed to investigate feedback circuits in the brain using fMRI?

While there is no limit, in principle, on the length of long-term fMRI studies, practical factors make very long studies unrealistic. One common manipulation is to change experimental instructions between runs, so that the participant does one task for several minutes and then another during the next run. This approach is necessary for many types of experimental questions, such as memory studies that separate stimulus learning from memory retrieval. When using such an approach, the researcher must monitor for problems like head motion and scanner drift that can cause systematic signal changes during and between runs. If the time scale needed is longer than an hour or two, such as when examining the encoding and retrieval of long-term memories, then multiple scanner sessions will be necessary. Comparisons of data from different sessions raises many additional problems, since there may be differences in head position, fatigue level, and practice with the task. Nevertheless, it is important to emphasize that while the lower and upper limits discussed here apply to most fMRI BOLD studies, it is possible to overcome them with careful selection of imaging parameters and clever experimental design.

The relationship between timing and hemodynamic amplitude is critical for models of fMRI activation. Because of the constraints of the vascular system, one might expect a larger response to reach its peak more slowly than a smaller response. Alternatively, a large-amplitude response might reflect greater neuronal activity, and thus greater metabolic demand, which could cause an earlier supply of oxygen and an earlier peak. While both arguments are plausible, neither is correct. For a given stimulus and brain region, the amplitude of the fMRI hemodynamic response appears to be independent of the timing of the response in both latency to onset (i.e., time before an initial rise above baseline) and latency to peak (Figure 7.22). This means that if you measure the fMRI BOLD response in the auditory cortex to the presentation of a 2-s music video clip, there will be no correlation across subjects between its amplitude and timing. However, if you measure the activation in that region in response to several different stimulus durations, the response at longer durations will be of
both greater amplitude and increased latency to peak. Likewise, if you compare activation in the auditory cortex with activation in another brain region (e.g., the visual cortex), the two brain regions may differ in both amplitude and latency. These differences could reflect differences in neuronal processing as a function of stimulus duration or brain region, or differences in the hemodynamic response itself. In summary, changes in the amplitude of the fMRI hemodynamic response do not appear to cause changes in the timing of the response, or vice versa, although external factors may affect both measures similarly.

**Effects of stimulus duration and timing**

Over intervals from a few seconds to a few tens of seconds, the duration of the fMRI hemodynamic response provides a reasonable estimate of the duration of neuronal activity. A good example of this was reported in 1997 by Richter, Kim, and their colleagues, who investigated the role of the superior parietal lobe in the mental rotation of complex objects. Decades of cognitive psychology studies have shown that the time needed to judge whether or not two similar objects are identical depends on the rotation angle between them. Thus, response time is twice as long for pairs that differ by 90° compared with pairs that differ by 45°. Richter and colleagues hypothesized that since the superior parietal lobe is critical for spatial processing, including mental rotation, its activation should extend in time for the entire rotation process. To test this idea, they measured the duration of the BOLD response in this region during each of 16 different trials that required objects to be rotated through different angles. They found that for each trial, the duration of the BOLD response matched well with the subject’s response time (Figure 7.23). The onset of the BOLD response did not differ across trials, suggesting that neuronal activity began at the start of the mental rotation process and stopped when mental rotation was complete.

Recognize, however, that the duration of the stimulus or behavior does not necessarily correspond with the duration of the neuronal activity. Imagine seeing a photograph of a familiar face that is presented for 5 s. Neurons in the primary visual cortex will have their largest response over the first 100 ms or so, then become much less active. Neurons within face-sensitive regions in the inferior temporal lobe will be active at stimulus onset, as well as throughout the 5 s period, due to feedback from other brain regions. Some neurons within
the frontal and parietal lobes may become active after the face disappears, as you reflect on what you just saw. For these reasons, researchers construct their fMRI analysis models based on estimated neuronal activity, not necessarily stimulus duration.

Even though it is difficult to determine the absolute timing of neuronal activity based on the fMRI BOLD signal, the relative timing of activation can be identified with great precision. In 1998 this was demonstrated in an elegant study by Menon and colleagues, who investigated relative timing in two fMRI experiments at high field (4.0 T) and high temporal resolution (TR of 100 ms). Their first experiment used a visual checkerboard stimulus that was split into left and right hemifields, each presented for 2 s. The left hemifield was presented either before the right with a delay interval ranging from 125 to 1000 ms, or at the same time. When they measured the difference in hemodynamic latency between the two hemispheres in the visual cortex, they found a near perfect correspondence with the presentation delay.

**Thought Question**

How is it possible that researchers can identify timing differences of a few hundred milliseconds using fMRI if the hemodynamic response is delayed by a few seconds following the neuronal activity?

The second experiment extended this result by introducing a motor reaction-time task. Each trial began with the subjects watching a fixation cross at the center of a screen for 6 s. When the screen flashed bright yellow, the sub-
ject used a joystick to move a cursor to a target as quickly as possible. One second after the cursor was successfully moved, the screen returned to gray and the subjects returned the joystick back to its center. The authors investigated the timing of activation in three cortical regions: the primary visual cortex (V1), the primary motor cortex (M1), and the supplementary motor area (SMA). Activation in V1 was observed, on average, about 200 ms before activation in SMA, which in turn was activated about 30 ms before M1. Especially remarkable were the effects of reaction time on these differences. Reaction times varied between subjects in a range from about 200 to 300 ms. The time differences between activation in V1 and SMA were proportional to the differences in reaction times, suggesting that the pathway between them may be associated with preparatory processes that differ between subjects (Figure 7.24A). However, the delay between SMA and M1 activation was constant across reaction times, suggesting that the link between these regions supports more basic response-execution processes. The timing differences found between these regions using fMRI were larger than those calculated from electrophysiological studies, where a latency of around 100 ms between V1 and M1 is more typical. This discrepancy illustrates the basic problem with comparing the timing of neuronal activity between brain regions based on fMRI data: differences in vascular properties also contribute to the timing of activity. This caveat, however, does not call into question the results of Menon and colleagues, because they demonstrated that the timing differences between regions depended on reaction time. To cognitive neuroscientists, these sorts of results suggest that variability in behavior, such as reaction time, the most commonly used measure in psychological experiments, can provide important information about the functions and connectivity between brain regions. We return to this theme in Chapter 11.

Within a single region, however, any vascular properties will be similar across timing conditions, allowing for accurate estimations of small timing differences. In 2000, Miezin and colleagues investigated the issue of relative timing using a clever visuomotor task. They presented checkerboard stimuli in a rapid event-related design. Each stimulus was presented for 1.5 s, and the time

Figure 7.24 Using fMRI to identify the relative timing of activation across brain regions. (A) After the presentation of a visual cue, subjects moved a target square across the display with a joystick. The relative latency of the fMRI signal in different brain regions as a function of the subject's reaction time (RT) on the task was measured. The latency between BOLD activation in the primary visual cortex (V1) and in the supplementary motor area (SMA) increased roughly linearly with RT, while the latency between SMA and the primary motor cortex (M1) was unchanged with RT. This result suggests that processes influencing the speed of a response occur between V1 and SMA but not between SMA and M1. (B) In an experiment with a similar approach, subjects pressed a button with one hand when a stimulus of 1.5 s duration appeared and another button with the other hand with that stimulus disappeared. The order of activation between motor cortex regions depended on the order of responses, while activation in the visual cortex was independent of response order. (A after Menon et al., 1998; B after Miezin et al., 2000.)
between stimuli was randomly varied, in a technique called jittering, with a mean of about 5 s between stimuli. The subjects' task was to press a button with one hand when the stimulus appeared and press another button with the other hand when the stimulus disappeared. In half of the trials, the right hand was used first, and in the other half, the left hand was used first. They found that the BOLD response in the motor cortex corresponding to the hand used at the onset of the stimulus was shifted earlier in time by about 0.75 to 1.0 s, compared with the response in the motor cortex corresponding to stimulus offset (Figure 7.24B). The fact that this shift was slightly less than the stimulus duration likely reflected preparation processes, since the subjects knew when the stimulus would disappear and could prepare for that button press. A subsequent power analysis suggested that timing differences as small as 100 ms could be reliably detected using this type of experimental design.

In summary, fMRI can measure small latency differences between brain regions, under appropriate task conditions (Figure 7.25). These differences may reflect the timing of neuronal activity and the form of the hemodynamic response. To draw inferences that are specific to neuronal activity, researchers

Figure 7.25 Maps of BOLD latency to peak in visual cortical regions. A 500 ms visual stimulus was presented and the latency to peak in voxels with significant BOLD activation was measured. Data are shown for two subjects [one subject shown in panels (A) and (B), and the other shown in (C) and (D)]. The colormap indicates the latency to peak of each voxel, with blue colors indicating more rapid hemodynamic responses, red colors indicating intermediate responses, and yellow colors indicating the slowest responses. Peak latency was shortest in voxels near the calcarine cortex (A, C), and was longest in the fusiform gyrus (B, D). (From Huettel et al., 2001.)

(A)  
(B)  
(C) Calcarine cortex  
(D) Fusiform gyrus
must selectively manipulate or measure one mental process while holding another constant. For example, decision processes may be proportional to response time, while low-level perceptual processes may be independent of response time, as in the study by Menon and colleagues. Comparisons within the same region do not suffer from the problem of neuronal vs hemodynamic activity, and thus very small within-region timing differences can be identified.

**Linearity of the Hemodynamic Response**

So far we have considered the properties of the hemodynamic response to a single isolated stimulus. What happens when multiple stimuli are presented in succession, as is necessary for most experiments? One possibility is that the same hemodynamic response is evoked for every stimulus, independently of the other stimuli presented, as shown in Figure 7.26A. If so, stimuli delivered sufficiently close together that their hemodynamic responses overlap would generate a total change in MR signal equivalent to the sum of the individual responses (Figure 7.26B). This is known as a linear system, and its properties are discussed in the following section. Another possibility is that the hemodynamic response to a given stimulus depends on what other stimuli are presented (Figure 7.26C and D). If two stimuli are presented very close together, the combined response might be less than the sum of the two individual responses. Reductions in hemodynamic amplitude as a function of interstimulus interval are known as BOLD refractory effects. If refractory effects are present, then a linear model will overestimate the hemodynamic response to closely spaced stimuli, potentially reducing the effectiveness of experimental analy-

**Figure 7.26** Linear and nonlinear addition of hemodynamic responses. If the same fMRI hemodynamic response were evoked by every stimulus (A), then the combined hemodynamic response to two stimuli would be a linear combination of two identical responses (B). But if the hemodynamic response were attenuated when stimuli were presented in rapid succession (C), then the combined response would be reduced in amplitude (D).
impulse A single input to a system. Impulses are assumed to be of infinitely short duration.

scaling A principle of linear systems that states that the magnitude of the system output must be proportional to the system input.

superposition A principle of linear systems that states that the total response to a set of inputs is equivalent to the summation of the independent responses to the inputs.

Properties of a linear system

A basic framework for measurement of signal in fMRI is shown in Figure 7.27. Under this framework, when a stimulus is presented, it induces neuronal activity within a particular region of the brain. Since the neuronal activity requires oxygen delivered via the vascular system, there are resulting flow and volume changes in that brain region, as discussed earlier in this chapter. The reduction in magnetic susceptibility associated with increased blood oxygenation then becomes measurable as BOLD contrast using fMRI. To reformulate this framework in terms of a linear system, we can consider the neuronal activity as a short-duration input, or impulse, to the hemodynamic system, whose output is measured as an MR signal. For a given impulse, the hemodynamic system is assumed to always respond in the same manner. From this assumption follow two basic principles of a linear system: scaling and superposition.

The principle of scaling states that the output of a linear system is proportional to the magnitude of its input (Figure 7.28A). If the input is doubled, then the output is likewise doubled; if the input is halved, then so is the output. For fMRI data, this principle would predict that changes in the relative amplitude of neuronal activity should lead to similar changes in the amplitude of the hemodynamic response. Why is this principle important? Remember that the goal of fMRI is to determine changes in neuronal activity, which must be inferred from the amplitude changes in the hemodynamic response. So, a study may involve a test condition and a control condition, find that activation in the brain region of interest was twice as large in the test condition, and then infer that neuronal activity was also twice as large in that condition. If the hemodynamic amplitude were independent of the amplitude of neuronal activity, then no such inferences would be possible.

Whereas scaling refers to the amplitude of activation, the principle of superposition refers to the timing of activation (Figure 7.28B). Stated simply, superposition means that the total response to two or more events is the summation of the individual responses. If a single event generates a hemodynamic response, then two events presented in succession will generate a combined response equal to two individual responses added together. To understand the importance of superposition, consider a very simple experiment. You are interested in studying the brain activation associated with short-term memory for verbal stimuli, so you devise an experiment in which a word is presented, followed by a 2 s delay and then another word. The subject’s task is to indicate whether or not the two words are the same. Since the two stimuli are both words, they will likely activate many of the same brain regions, and since they are presented so close together in time, their hemodynamic responses will overlap. How can you determine whether an area is active in response to the first word, the second word, or both? By assuming superposition, you can create
models for what combined hemodynamic response would be expected in each condition. If an area is active in response to the first word but not the second, it will show an early rise to a peak at about 5 s, whereas if it is active only in response to the second word, the rise will occur 2 s later. If it is active in response to both words, there will be an intermediate but larger peak. Without assuming superposition, or using approaches to correct for deviations from superposition, analysis of complex designs would be very challenging.

**Evidence for rough linearity**

The first study of the linearity of the fMRI BOLD response was reported by Boynton and colleagues in 1996. They investigated the effects of duration of visual stimulation, from 3 to 24 s, on activation in the primary visual cortex, using an EPI gradient-echo sequence at 1.5 T. The visual stimuli were checkerboard patterns that moved from right to left while flickering at a high rate (8 Hz). The contrast between dark and light squares on the checkerboard was manipulated over trials. The authors hypothesized that the duration and contrast of the stimuli should have independent and additive effects on the BOLD signal if the hemodynamic response behaved as a linear system. They found that, for all stimulus durations, the amplitude of activation was greater at higher visual contrast levels, but the hemodynamic response had the same basic shape. This was consistent with the scaling principle. To test whether the superposition principle also held for the BOLD data, they investigated whether the response to a longer stimulus could be predicted by the sum of multiple shorter stimuli. Their results indicated linear superposition for most stimulus durations, so the response to a 24-s stimulus could be approximated by the addition of two sequential responses to the 12-s stimuli or four sequential responses to the 6-s stimuli (Figure 7.29). The one exception they noted was that using the response to a 3-s stimulus led to an overestimation of the responses to longer stimuli, and this discrepancy was attributed to neuronal
adaptation effects (i.e., the neurons themselves decrease in activation over the first few seconds of a stimulus). In total, these results provided strong evidence for the linearity of the hemodynamic response over stimulus durations ranging from a few seconds to a few tens of seconds.

Building on the results of Boynton and colleagues, in 1997 Dale and Buckner investigated the linearity of the hemodynamic response to individual stimulus events, rather than extended stimulus blocks. They reasoned that neuronal adaptation effects could be eliminated by separating stimuli in time by an interval sufficiently long to allow complete neuronal recovery. To test whether complete linearity would be observed under such conditions, they presented clusters of one, two, or three stimuli at interstimulus intervals of either 2 or 5 s. The hemodynamic response to a second stimulus in a cluster was determined by subtracting the single-stimulus response from the response to a pair of stimuli. Likewise, the response to the third stimulus was determined by subtracting the two-stimulus response from the three-stimulus response. If the principle of superposition holds for fMRI BOLD data, then all responses should have the same amplitude and shape regardless of where they occur in the cluster. Dale and Buckner's evidence supported this interpretation, in that the responses to the second and third trials in each set were generally similar to that of the first trial, (Figure 7.30) especially for pairs of trials at a 5-s interval. They concluded that the fMRI BOLD response increases with additional stimuli in a "roughly linear" fashion at intervals typical of experimental testing (e.g., a few seconds).

This finding, that individual trials evoke recognizable hemodynamic responses even when separated by relatively short intervals, has become
extremely influential in the design and analysis of fMRI studies. At first interpretation, it suggests that individual trials can be closely spaced to increase the number of trials presented and thus the experimental power, without loss of response amplitude. But, as the authors themselves note, there might be significant limitations in such fast-rate studies. The most important of these are the deviations from linearity at short intervals. Close examination of the data from 2-s interstimulus intervals (see Figure 7.30B) reveals that the responses to the second and third stimuli in a cluster are reduced in amplitude and increased in latency compared with the response to a single stimulus. In the best case, such nonlinearities would reduce the power of experimental analyses but would have little additional effect. But in the worst case, large nonlinearities could preclude the use of very short intervals (e.g., 1 s) between experimental stimuli. As shown in a 2005 article by Wager and colleagues, accounting for nonlinearities within fMRI analysis models can dramatically improve their sensitivity.

**Challenges to linearity**

Subsequent work on nonlinearities in the fMRI hemodynamic response investigated whether there is a refractory period following stimulus presentation during which subsequent stimuli evoke smaller hemodynamic responses. A preliminary suggestion based on the results described above is that the fMRI hemodynamic response may be nonlinear at intervals of less than 6 s but linear at longer intervals, since superposition was found at durations of 6 s but not 3 s, and better scaling was observed at intervals of 5 s than 2 s. Tests of refractory periods have been conducted using both blocked and event-related designs. Blocked designs (e.g., the 1996 Boynton and colleagues study) usually manipulated stimulus duration, and event-related designs (e.g., the 1997 Dale and Buckner study) usually manipulated interstimulus interval. The issues investigated by the researchers included the timing of refractory effects, whether such effects differ across subject groups, and whether different brain regions have different refractory properties.

Blocked-design studies have verified that substantial refractory effects are present after short stimulus durations. In 1998 Robson and colleagues examined whether the response to a long-duration auditory stimulus could be predicted by the addition of multiple short-duration stimuli, testing durations from nonlinearity. The property whereby the combined response to two or more events is not equivalent to the summation of the responses to the individual events in isolation.

**Refractory period** A time period following the presentation of a stimulus during which subsequent stimuli evoke a reduced response. For BOLD fMRI, the refractory periods for many types of stimuli last approximately 6 s.
Figure 7.31 Nonlinearity of the hemodynamic response at short interstimulus intervals. Each trial consisted of a single visual stimulus or a pair of visual stimuli separated by a short interval. Single stimuli evoked a robust hemodynamic response (black line). At short interstimulus intervals (e.g., 1 to 2 s), the hemodynamic response was reduced in amplitude and increased in latency. These changes reflect the presence of refractory effects in the hemodynamic response. (After Huettel and McCarthy, 2000.)

100 ms to 25.5 s. As for the previous study by Boynton and colleagues, superposition held for stimuli of 6 s or more in duration, but not for shorter-duration stimuli. Furthermore, as the stimuli became shorter, the deviations from superposition became greater. Similar results were reported by Vazquez and Noll in 1998 for visual stimuli, with significant nonlinearities found for stimuli shorter than 4 s. Event-related studies also support the idea of a refractory period. Huettel and McCarthy presented short-duration visual checkerboard stimuli either singly or in pairs separated by a 1-s to 6-s interstimulus interval (Figure 7.31). Within the primary visual cortex, the response to a second stimulus in a pair separated by 1 s was reduced by more than 40% compared with the single-stimulus response and was delayed in time by nearly 1 s. By 6 s, both the amplitude and latency of the hemodynamic response to the second stimulus in a pair had returned to near normal values. The finding of both amplitude and latency changes is consistent with the earlier work of Dale and Buckner, as well as with computational simulations of the hemodynamic response.

The existence of refractory effects in the BOLD signal is now well-supported by many studies. As a result, the focus of research has shifted toward examining differences in refractory effects between individuals or brain regions. One of the first subject group comparisons was between young adults and healthy elderly adults. Given the many structural changes in the brain that accompany aging, including possible impairments in blood flow, it was predicted that refractory effects would be greater in elderly adults. This prediction, though sensible, turns out to be wrong. The magnitude of the refractory effect is similar for older and younger adults, as is the amplitude of the hemodynamic response itself. Another area of research is whether, in some patient groups, nonlinearities might be diagnostic of neurological or psychiatric deficits. Patients with schizophrenia, for example, exhibit abnormal sensory habituation, as indicated by electrophysiological measures. While neurologically normal adults show a significant decrease in electrophysiological responses when a stimulus is presented twice in rapid succession, a phenomenon known as sensory gating, schizophrenic adults show a much smaller decrease. While there are now many fMRI studies of sensory gating, whether these effects have consequences for the BOLD response remains unknown. For example, in a 2003 study by Barch and colleagues, individuals with schizo-
phrenia showed normal BOLD linearity with an interstimulus interval of 5 s. Future research that combines electrophysiological and fMRI data will provide important insights into these and other individual differences.

A number of studies have shown that refractory effects can also differ in different brain regions. Primary sensory or motor regions appear to have smaller refractory effects (i.e., are more linear) than other regions of the cortex. In 2001, Birn and colleagues measured activity in the motor regions M1 and SMA while subjects tapped their fingers for different durations. The hemodynamic response in M1 increased in amplitude with increasing duration, although there were deviations from superposition for stimuli shorter than about 4 s. But in the SMA, which is involved with the planning of motor behavior, activation amplitude was similar across all stimulus durations. Huettel and colleagues showed in 2004 that a similar relationship holds for visual stimuli: activation in the primary visual cortex scaled with duration of a visual stimulus, but activation in the motion-sensitive region V5 was independent of duration. However, comparisons of refractory effects between brain regions may be confounded by differences in neuronal activity. In both of the studies described above, the secondary regions examined (SMA and V5) may be transiently active at the beginning of the stimulus, while the primary cortex may show sustained activation throughout stimulus presentation. Thus, it is worth reemphasizing that refractory effects measured using fMRI may result from neuronal adaptation or from changes in vascular responsiveness, either of which could differ in different brain regions.

fMRI-adaptation

Although refractory effects are often seen as impediments to fMRI research, they can be used to understand brain function in cleverly designed experiments. The basic idea comes from physiological studies of adaptation. If one repeatedly presents the same stimulus to a subject, such as repeatedly exposing a rat to a loud noise or showing a bright red balloon to a human infant, the response to that stimulus will diminish over time. After a number of presentations, the rat will stop flinching at the noise and the infant will become bored with the balloon. When the response to a stimulus decreases with repeated presentation, the subject is said to have adapted to that stimulus. By changing the stimulus slightly, we can discover to which aspects of the stimulus the subject has adapted. If we replace the red balloon with a red fire truck and the infant is still bored, then we can infer that she adapted to the color red. But if she is now very interested in the fire truck, then she must have adapted to some other property of the balloon, such as its shape. Many neurons in the visual system will preferentially fire for one type of visual stimulus, due to its particular shape, color, or way of movement. But with repeated presentations of that stimulus, the neuron's firing rate will decrease dramatically.

Given the presence of fMRI refractory effects, we can apply similar principles to experimental designs in an approach known as fMRI-adaptation (Figure 7.32). First, we present a stimulus that evokes activation in a particular brain region. After a short time period, we then present another stimulus that differs in some fashion. If the region being investigated contains neurons that respond differently to the new stimulus, it will show an increase in fMRI activation. But if the neurons do not distinguish between the old and new stimuli, fMRI activation will be at a low level because of refractory effects. For example, refractory effects in the secondary (but not primary) visual cortex depend on the orientation of visual line gratings; refractory effects in face-sen-

adaptation A change in the response to a stimulus following its repeated presentation.

fMRI-adaptation Reduction in the BOLD response to the repeated presentation of a set of stimuli that differ in some attribute, indicating that the brain region being studied is insensitive (as measured by fMRI) to the stimulus attribute being varied.
Figure 7.32 A schematic overview of fMRI-adaptation. (A) In one test of fMRI-adaptation, here illustrated for a visual perception task, the fMRI subject first views a series of identical objects. This leads to adaptation of neurons throughout the visual pathway, from those supporting low-level properties to those supporting high-level object processing. (B) Then, other sets of objects are presented that differ from the adaptation set in some property, to evaluate whether there is release from adaptation. Shown here are examples of stimuli that differ in very simple physical properties (i.e., color and size) or complex categorical properties (i.e., car model and vehicle type). By comparing which stimulus sets lead to the greatest fMRI activation in a given brain region, the researcher can identify properties that the neurons in that region are sensitive to. Shown here are hypothetical BOLD responses in (C) a region whose neurons are sensitive to any physical change in a stimulus other than color and (D) a region whose neurons code complex object categories, like cars, but do not generalize to all vehicles.

Sensitive cortical regions are greater for pairs of identical faces than for pairs of different faces; and refractory effects in the motion-sensitive area V5 depend on the direction of motion.

While these and other findings provided insight into the functional specificity of BOLD effects in these regions, other studies have explored fMRI-adaptation in more complex forms of processing. In an elegant series of experiments published in 2001, Grill-Spector and Malach investigated whether neurons in the lateral occipital complex, which contains the lateral occipital lobe and the posterior fusiform gyrus, are sensitive to changes in higher-order properties of objects. For one of their experiments, they presented the same face repeatedly within a 10-s block. They compared the activation in this “identical face” condition with the activation measured in conditions where some property of the face was manipulated, such as position on the screen, viewpoint from left to right, illumination, size, or identity (i.e., different faces presented). Only one condition was varied within each block. In these brain regions, roughly similar activation would be observed in response to any of the faces presented in isolation, so differential activation in the blocks would indicate refractory effects.
associated with some aspect of the stimulus presentation. The researchers found that activation in the fusiform gyrus, for example, was greatly reduced in response to repeated presentations of identical, size-varying, or position-varying faces, compared with repeated presentations of entirely different faces. In contrast, variation in direction of illumination or viewpoint caused a recovery from adaptation. From these results, the researchers concluded that the fusiform gyrus recognizes facial identity despite size or position manipulations, but does treat illumination or viewpoint manipulations as new stimuli.

The power of the fMRI-adaptation approach comes from its flexibility. In principle, any stimulus property that can evoke activation within a region can be manipulated to evaluate adaptation effects. For example, two independent but simultaneously published 2007 studies by Piazza and colleagues and by Cohen Kadosh and colleagues used fMRI-adaptation to study number processing in the posterior parietal cortex. With a number of manipulations, both groups found clear evidence that voxels in this region coded for numerosity regardless of whether the stimulus was presented as a numeral, word, or dot pattern. That is, if the numeral “4” was followed by the word “four,” the response to that word was attenuated. (Note that there was also some evidence that at least some parts of the parietal cortex may code separately for numeral and word representations of a number.) Moreover, the amount of recovery from adaptation depended on the magnitude difference between successive stimuli: a larger response was observed when the new number was very different from the adaptation number (e.g., from 17 to 39) compared with when the numbers were only slightly different (e.g., from 17 to 20). Other domains to which fMRI-adaptation has been applied include commonalities between visual and motor processing, properties of word processing, and many other aspects of sensation and perception. Yet, there are limitations to these types of studies: effects attributed to adaptation within a region could also arise from adaptation within neurons elsewhere that project to the target region. For a broad overview of the strengths and limitations of fMRI-adaptation, see the 2006 review article by Krekelberg and colleagues in the Chapter References.

Summary

Nearly all fMRI studies rely on an endogenous measure known as blood-oxygenation-level dependent (BOLD) contrast. The idea of using blood properties as indices of brain functions is more than a century old. Early research found that deoxygenated hemoglobin is paramagnetic and altered the T₂* properties of nearby tissue. Important studies demonstrating the feasibility of BOLD contrast were conducted by Ogawa and colleagues using both test tube and animal models. These results, in conjunction with the increased prevalence of MRI scanners and the development of high-speed pulse sequences, set the stage for the growth of fMRI in the early 1990s. The BOLD response to brief neuronal activity (i.e., the hemodynamic response) consists of a short onset delay, a rise to a peak after a few seconds, a return to baseline, and a prolonged undershoot. Some researchers report the presence of an initial dip in the BOLD signal due to oxygen extraction before the large increase in blood flow, but this effect is not always seen. The spatial resolution of an fMRI study determines our ability to separate adjacent brain regions with different functional properties. It can be improved by collecting data using smaller voxels, at a cost of reduced BOLD signal. However, the organization and response properties of the vascular system introduce spatial constraints. The temporal resolution of fMRI refers to the ability to esti-
mate the timing of neuronal activity from the measured hemodynamic changes. While using a short repetition time can improve temporal resolution, the sluggishness of the hemodynamic response limits researchers’ ability to make precise temporal measurements. Nonlinearities in the fMRI hemodynamic response reflect the temporal dynamics of activation in a single spatial location. If the same brain region becomes active twice in rapid succession, within an interval of about 6 s, the hemodynamic response to the second event is reduced in amplitude compared with that evoked by a single event. While refractory effects present an analysis challenge for most studies, some researchers have taken advantage of these effects to study functional adaptation within a brain region.

**Suggested Readings**


*Indicates a reference that is a suggested reading in the field and is also cited in this chapter.
Chapter References


the cerebral oxygen/glucose uptake ratio by brain activation: evidence obtained
Magistretti, P. J., and Pellerin, L. (1999). Cellular mechanisms of brain energy metab-
olism and their relevance to functional brain imaging. Philos. Trans. Royal Soc. Lon-
Malonek, D., and Grinvald, A. (1996). Interactions between electrical activity and cor-
tical microcirculation revealed by imaging spectroscopy: Implications for func-
Mandeville J. B., Marota J. J., Kosofsky B. E., Keltner J. R., Weissleder R., Rosen B. R.,
and Weisskoff R. M. (1999). MRI measurement of the temporal evolution of rela-
Menon, R. S. (2002). Postacquisition suppression of large-vessel BOLD signals in
Menon, R. S., and Kim, S. G. (1999). Spatial and temporal limits in cognitive neu-
based functional MRI at 4 Tesla includes a capillary bed contribution: Echo-planar
imaging correlates with previous optical imaging using intrinsic signals. Magn.
Characterizing the hemodynamic response: Effects of presentation rate, sampling
procedure, and the possibility of ordering brain activity based on relative timing.
Moseley, M. E., and Glover, G. H. Functional MR imaging: capabilities and limita-
Coupling between neuronal firing, field potentials, and fMRI in human auditory
Ogawa S., Lee T. M., Nayak A. S., and Glynn P. (1990b). Oxygenation-sensitive con-
trast in magnetic resonance image of rodent brain at high magnetic fields. Magn.
89(13): 5951–5955.
mon to numerosities and number symbols in human intraparietal cortex. Neuron,
53: 293–305.
fMRI response of the human auditory cortex to trains of tones. NeuroImage, 7:
185–198.


