Part I

Neuroscience, ethics, agency, and the self

Chapter 1

Moral decision-making and the brain

Patricia Smith Churchland

Neuroethics: The coming paradigm

As we understand more about the details of the regulatory systems in the brain and how decisions emerge in neural networks, it is increasingly evident that moral standards, practices, and policies reside in our neurobiology. As we learn more about neural development, the evolution of nervous systems, and how genes are regulated, it has become evident that our neurobiology is profoundly shaped by our evolutionary history. Our moral nature is what it is because our brains are as they are; so too, for our capacities to learn, reason, invent, and do science (Roskies 2002).

Although our moral capacities are prepared during embryological development, they are not wholly configured at birth. One's social and cultural world, with its various regulatory institutions, deeply shapes the exercise of moral capacities in adulthood. These regulatory institutions include the standards prevailing in one's particular family and clan, the prevailing criminal justice system, the organization and style of government, schools, guilds, religions, and professional societies (P.M. Churchland 2000).

Recognition of these various determinants means that the traditional field of ethics must itself undergo recalibration. Philosophers and others are now struggling to understand the significance of seeing morality not as a product of supernatural processes, 'pure reason' or so-called 'natural law', but of **brains**—how they are configured, how they change through experience, how cultural institutions can embody moral wisdom or lack of same, and how emotions, drives, and hormones play a role in decision-making. Some traditional assumptions concerning the roots of moral knowledge have been exposed as untenable. As these assumptions sustain reconfiguration, the beginnings of a new paradigm in ethics can be seen emerging. Owing to the natural and biological roots of morality, this new approach to ethics may be referred to as 'naturalized ethics', or more simply, 'as neuroethics' (P.S. Churchland 1991, 2002; Flanagan 1996; Campbell and Hunter 2000; Illes and Raffin 2002; Roskies 2002; Casebeer and Churchland 2003; Goodenough and Prehn 2004).

The new research on the nature of ethics is located at the interface of philosophy, jurisprudence, and many sciences—neuroscience, evolutionary biology, molecular biology, political science, anthropology, psychology, and ethology. These interdisciplinary inquiries will have profound, and rather unpredictable, social consequences, as people in general rethink their conventional ideas concerning the basis for moral standards and practices. In this context, it will also be important to consider the impact on, and interactions with, organized religion, although I shall not address that matter in this chapter. Here, I shall focus mainly on one central aspect motivating the changing view, namely what we are learning about the neurobiological nature of decisions.

Decisions and decision-making

Brains incessantly make decisions: some pertain to the immediate future, and others to the distant future; some are trivial, and others are momentous. Some decisions concern only oneself and one's own interest; others concern the interests of offspring, family, clan and distant neighbors; yet others may pertain to the welfare of animals, birds, and the land. From the perspective of the brain, these are all just decisions, involving emotions, memory, prediction, evaluation and temperament. From the perspective of human society, we find it useful to characterize a subset of these, namely the decisions pertaining to the interests of others, as moral decisions. Nevertheless, within the huge domain of practical decision-making, no sharp line demarcates the moral from the non-moral issues. Instead, decisions fall on a continuum, with prototypically non-moral decisions at one end (e.g. Should I have grapefruit or bananas for breakfast?), and prototypically moral decisions at the other end (e.g. How should we punish a juvenile capital offender?). In between, there are many cases where, if inclined, we may argue about whether the matter is best considered genuinely moral or merely pragmatic; whether it concerns justice, or only manners and politeness. All decisions involve some evaluation of the predicted consequences. Such evaluation is anchored by the brain's reward system, by the complex networks supporting emotions, appetites, and moods, and finally by one's background knowledge of the way the world works and one's own varying capacities for acting in it.

In childhood we normally acquire the skills to navigate the physical world of food, danger, and shelter. Acquired in tandem are the skills to navigate the social world; we learn how to please and avoid displeasure, how to find and give comfort, and how to cooperate, share, and benefit from compromise. Children discover the prototypes of exploitation, fairness, cheating, and altruism. They acquire the skills needed to survive and, if they are lucky, to flourish in both the family and the wider social world. Tacitly, they become sensitive to parameters of time to act, and the need to decide on the basis of imperfect and imprecise knowledge, of balancing risk and caution. Contemplation of theoretical knowledge may not require courage, good sense, resolve, or balance, but acting on knowledge surely does. As we all know, a theoretically clever person can, for all that, be a practical fool.

Aristotle (384–322BC) was really the first thoroughly to articulate the idea that the substance of morality is a matter of practical wisdom, rather than a matter of exceptionless rules received from supernatural or other occult sources. On the Aristotelian conception of morality, practical wisdom (praxis) requires development of appropriate habits and character, as well as a thoughtful understanding of what social institutions and judgments best serve human flourishing, all things and all interests considered. Just as theoretical knowledge of the physical world can evolve over generations and through the lifetime of a single person, practical understanding of the social world can evolve likewise (P.M. Churchland 1989, 2000). The evaluation of practices such as child labor, public education, trial by ordeal, trial by jury, income tax, slavery, cannibalism, human sacrifice, separation of church and state, military draft, and so forth occurs as people reflect on the benefits and costs of these institutions.

The role of the brain's reward system in social learning normally fosters respect, or even reverence, for whatever human social institutions happen to exist. Therefore change in those institutions may be neither fast nor linear, and may be vigorously resisted even by those who stand to benefit from the change; for example, women who opposed the vote for women and the poor who oppose taxation of the very rich. Despite the power of social inertia, modifica-

tions, and sometimes revolutions, do occur, and some of these changes can reasonably be reckoned as moral progress (P.M. Churchland 1989).

That individuals are to be held responsible for their actions is a common human practice, and typically involves punishment in some manner when actions violate the established standards. By its very nature, punishment inflicts pain (or more generally, dis-utilities) on the punished. Consequently, the nature and justification of punishment, and its scope, mode, and limitations, have traditionally been the locus of much reflection and debate. As science has come to understand the physical basis for insanity and epilepsy, there have been changes in the criminal law to accommodate the scientific facts. Insanity, for example, is rarely considered now to be demonic possession, best treated by isolation to a dung heap. In this century, neuroscientific advances in understanding higher functions inspire renewed reflections on the fundamentals of responsibility and punishment. At the most basic level reside questions about the relation between free choice, punishment, and responsibility.

Brains, souls, and causes

The brain is a causal machine. Or, perhaps more accurately, given everything that is so far known in neuroscience, it is very probable that the brain is a causal machine. By calling it a causal machine, I mean that it goes from state to state as a function of antecedent conditions. If the antecedent conditions had been different, the result would have been different; if the antecedent conditions remained the same, the same result would obtain. Choices and evaluation of options are processes that occur in the physical brain, and they result in behavioral decisions. These processes, just like other processes in the brain, are very probably the causal result of a large array of antecedent conditions. Some of the antecedent conditions result from the effects of external stimuli; others arise from internally generated changes, such as changes in hormone levels, glucose levels, body temperature, and so forth.

Available evidence indicates that the brain is the thing that thinks, feels, chooses, remembers, and plans. That is, at this stage of science, it is exceedingly improbable that there exists a non-physical soul or mind that does the thinking, feeling, and perceiving, and that in some utterly occult manner connects with the physical brain. Broadly speaking, the evidence from evolutionary biology, molecular biology, physics, chemistry, and the various neurosciences strongly implies that there is *only* the physical brain and its body; there is no non-physical soul, spooky stuff, or ectoplasmic mind-whiffle. For example, there is no reason to believe that the law of conservation of mass/energy is violated in nervous systems, which it would have to be if the non-physical soul could make changes in the physical brain. The most plausible hypothesis on the table is that the brain, and the brain alone, makes choices and decides upon actions. Moreover, it is most likely that these events are the outcome of complex—extremely complex—causal processes (P.S. Churchland 2002).

If an event is caused by antecedent causal factors, does this mean that the event is predictable? Not necessarily. When a system is very complex, and when small changes at one time can be amplified over time to result in large differences in the end result, it is often very difficult to predict exactly the behavior of the system. This characterizes many dynamical systems. For example, it is impossible to predict with great precision exactly whether a tornado will emerge, and exactly when and where it will emerge. Exact predictability is elusive, not because tornadoes are uncaused, but because the weather is a complex dynamical system, there are very many variables to measure, and the values of the variables can change over very short time-scales. Consequently, in practice we cannot make all the measurements and perform all

the calculations in real time in order to make precise predictions. The logical point of importance here is that the proposition 'events in system S cannot be precisely predicted in real time' is entirely consistent with the proposition that all the events in system S are caused, i.e. no uncaused events occur in S.

Nevertheless, even when a dynamical system is too complex for exact moment-to-moment prediction, general or rough predictions are certainly possible and technological advances may enhance predictability. Thus satellite photographs and radar maps make it possible to predict roughly where a hurricane will make landfall, at least within several hundred miles, and when, at least within tens of hours. However, greater precision remains impossible with current technology.

Similarly, although it may not be possible to predict exactly what a person may decide on a question regarding which seat to take in a bus, general rough predictions are possible, especially when one knows another person well. In a classroom, students tend to sit in the same seat each time the class meets, and hence I can predict that it is quite likely that Bill will sit in the front row aisle seat, since he has sat there for the last three times the class has met. I can reliably, if imperfectly, predict that another person will be offended if I gratuitously insult him, or that a 6-year-old child will prefer a sweet thing to a bitter thing, that a student will say 'apple' but not 'pomegranate' when asked to name the first fruit he thinks of, or that a person thrown into icy water will want to scramble out quickly. Technology and knowledge allow for improvements in our predictions about people and their behavior. Imaging techniques (functional MRI) showing unusually low levels of activity in orbitofrontal cortex can help us predict that a person is depressed. If a person has a mutation in the gene that normally produces the enzyme monoamine oxidase A (MAOA), and if he has also had an abusive upbringing, we can reliably predict that he will display irrationally violent and self-destructive behavior. An electroencephalogram (EEG) in which electrodes are placed on the scalp can detect brain changes that predict the imminent onset of an epileptic seizure. It has also been shown that criminal recidivism can be roughly predicted by using EEG in a Go–No Go paradigm (Howard and Lumsden 1996).

To be free, must our choices be uncaused?

A common assumption states that when one's choice is genuinely free, then one created that choice—created it independently of whatever causal events might be occurring in the brain (Van Inwagen 1983; Allison 1990; Kane 2001; Korsgaard 2001; Pereboom 2001). It may be thought that such a choice springs into being as a result of the exercise of pure agency unfettered by any causal antecedents. According to this view, reasons may justify choosing one option rather than another, but reasons do not causally affect the will, for the will acts in a kind of causal vacuum. A free choice will have causal effects, but it has no causal antecedents. This view is known as libertarianism, or the thesis of contra-causal free will.

Looked at in the context of the preceding discussion, the idea of contra-causal free will has very low figures of merit. If choices are brain events, and if brain events have causal antecedents, then choices have causal antecedents. In addition, we can understand quite well why it may seem, introspectively, as though one's choice was uncaused. There are basically two reasons for this. First, our brains are not conscious of all relevant neural events antecedent to choice. From the inside—introspectively, as it were—a person will have no apprehension of non-conscious antecedent causes. Hence one may be inclined to consider the choice as springing from nothing—nothing but his free will. But the fact is, we have no introspective access to many events that happen in our brains. I cannot, for example, introspect the events occurring in the

retina, or in the spinal cord. One just feels sexual appetite, without conscious access to the biochemical changes underlying those feelings, and one may imagine that they emerge uncaused. I cannot introspect the processes that typically precede normal unrehearsed speech, and may fancy that speech emerges from the freely acting will, independently of any causal antecedents. But by using imaging techniques such as functional MRI (see also Chapters 11 and 12), the causes can be seen to occur before the conscious awareness of intent.

A second reason helps to explain the introspective sense that choices are innocent of antecedent causes. When a decision is made to move—whether to move the eyes, the whole body, or the attention to a new task—a copy of the command goes back to other regions of the brain. This signal is called efference copy. As Helmholtz observed, one visible demonstration of efference copy consists in observing the difference between moving your eyes to the right and pressing your eyeball to the right. In the latter case, the world seems to move; in the former case, the brain knows, via efference copy, that the movement is my movement, not the world's. Consequently, the conscious visual effect is completely different. The wiring for efference copy means, for example, that it feels different when I lift my arm in deep water and when my arm rises in the water. One might, mistakenly, attribute that difference in feeling to something else: in one case, the buoyancy of the water causes my arm to rise; in the other case, uncaused free will exerts itself so that I raise my arm. However, the actual difference in feeling depends on efference copy or the lack thereof.

Reflecting on libertarianism in the eighteenth century, the Scottish philosopher David Hume realized clearly that the idea of contra-causal free will was muddled (Hume 1739). Hume pointed out, quite correctly, that choices are caused by desires, beliefs, hopes, fears, drives, intentions, and motives. He realized that our preferences are affected by our character, temperament, hormones, and childhood experiences; they are affected by how sleepy or alert we are, by how sick or well we are, by our habits and history. Moreover, and this was the crucial logical point, Hume realized that if a choice could be made independently of all these factors—if, *per impossibile*, it were independent of character, habit, inclination, belief, desire, and so on—the choice would be so bizarre, so 'out of the blue' as to raise the question of whether it was a real choice at all. Those choices we consider free choices, argued Hume, are those that result from our character, needs, habits, and beliefs about duty, among other things. Moreover, these are the choices for which we hold people responsible. If I suddenly choose to throw my phone out of the window, but had no desire, intention, or inclination to do so, if I had no belief that it was my duty to do so, no need to do so, no belief that it was in my self-interest to do so, then, as Hume saw, this would be a rather absurd action. This action would certainly *not* be considered the paradigm of a freely chosen action.

In contrast, consider as a paradigm of free choice President Truman's decision to drop an atomic bomb on Hiroshima in 1945. He believed that the war would continue for a long time if fought in the traditional manner and that American causalities would continue to be extremely high. He believed that Emperor Hirohito would not surrender unless it were starkly clear to him that his civilian losses would be catastrophic. Truman deliberated, reflected, considered the options, weighed the intelligence reports, and finally chose to drop the atomic bomb. His decision appears to have been the outcome of his desires, motives, beliefs, fears, and predictions. It is a decision for which he is held responsible. Some historians praise him for the decision and others blame him, but no one doubts that he was responsible. Was he caused to make the decision? Not by coercion or panic or reflex was he caused. Nevertheless, the various states of Truman's brain—states that we call motives, beliefs, and so on—did jointly cause a particular decision. Some of those causes, such as certain beliefs, may also be called reasons.

This does not distinguish them from causes, but is a conventional way of marking certain causes, such as beliefs and goals, as distinct from certain other causes, such as panic and obsession.

Those who find the idea of contra-causal freedom appealing have sometimes looked to quantum physics as providing the theoretical framework for contra-causal free choice. Roughly, the idea is that the indeterminacy at the quantum level—the collapse of the wave function—might provide the physical platform for contra-causal agency. Notice that Hume's argument, although some 200 years old, is sufficiently powerful to sideline this strategy.

If Hume's argument is correct, then free choice is not uncaused; rather, it is caused by certain appropriate conditions that are different from the set of conditions that result in involuntary behavior. Consequently, the appeal to quantum physics is completely irrelevant (P.M. Churchland 2002). As Hume might have said: Why would you consider an act free if it came about through pure randomness, as opposed to coming about as a result of desires, temperament, motives, and goals? Sometimes of course we are flippant about a choice, when there is little to choose between options or when we are indifferent to the consequences of the options, especially if they are trivial. It really does not matter whether I have a blue or a red toothbrush, and so in shopping for a new brush I just take one without deliberation or evaluation of the consequences. However, these sorts of cases are not the paradigm cases of reflective or deliberative or thoughtful choice. They are the minor players, at least in terms of momentous issues of responsibility and punishment.

The further point is that classical physics appears to be entirely adequate to account for the behavior of neurons and their interactions with one another. If uncaused events do actually exist in the neurons, perhaps in tiny structures such as microfilaments, such nanostructural events are unrelated to the level of neural network events that result in mental states, including deliberation and choice. Just as perception and memory storage involve information distributed across many neurons in a network, so a decision to act will be made not by a single neuron, but by many neurons interacting in a network. These interactions at the network level can safely be assumed to overwhelm any non-determined quantum level events that might happen to take place in the inner structures of a single neuron.

Hume's argument against the contra-causal view of rational choice seems to me essentially correct, and, as far as I can determine, no one has ever successfully refuted this argument, or even come close to doing so (see also Dennett 2003). Hume went on to develop a further important point that is especially powerful in the context of the developments in neuroscience that have occurred since about 1980. Assuming the cogency of his argument showing that decisions and choices are caused, Hume went on to propose that the really important question to answer is this: What are the differences between the causes of voluntary behavior and involuntary behavior? What are the differences in causal profile between decisions for which someone is held responsible, and decisions for which someone is excused or granted diminished responsibility? (See also Chapter 3).

From the point of view of civil and criminal law, and of society's interests more generally, these are indeed the important questions. Moreover, at bottom these questions are empirical, for they inquire into the nature of causes of responsible behavior, i.e. for behavior distinguishable on the social criteria that it either does or does not make sense to hold the agent responsible. Hume can be construed as proposing a hypothesis: there are discoverable empirical differences between the causes of accountable behavior and excusable behavior. In the next section, I shall develop Hume's hypothesis using data from the neurosciences. First, however, we need briefly to recall how the law regards choice and responsibility.

Criteria for accountability and diminished responsibility

In *The Nichomachean Ethics*, Aristotle raises the following question: When should someone be held responsible for his actions? Very insightfully, Aristotle noted that the best strategy is to make responsibility the default condition, i.e. a person is assumed to be responsible for his action unless it can be shown that certain conditions in the person or his environment reduce or excuse responsibility. The main idea was that if punishment in a certain type of case would neither deter nor improve the future behavior of persons, including the defendant—if the punishment in these circumstances fails provide a reason to avoid the action in future—then full responsibility does not apply. To be excused from responsibility, Aristotle reasoned, unusual circumstances must obtain. For example, the person might be insane, and hence completely fail to understand the nature of what he is doing. Or the person might be sleep walking, and injure someone while imagining that he is rowing a boat. If a defendant injured someone while involuntarily intoxicated, for example, then he may be excused, but if he was voluntarily intoxicated, he must be held responsible, since he is expected to have known the dangers of drinking. Aristotle also recognized that responsibility can come in grades and degrees. Depending on conditions, someone may be granted diminished responsibility and therewith, reduced punishment, rather than be excused entirely from responsibility

In many Western countries, standards for determination of responsibility are rooted in the Aristotelian foundation. The standards have been elaborated and revised with the advent of new knowledge concerning defects of decision-making, and the developments in moral standards consequent upon appreciating the significance of the new knowledge. For example, in the USA, Canada, and England, for a person to be convicted of a crime, (a) he have performed the action, (b) the action must be a violation of a criminal statute, and (c) he must have a criminal state of mind. The latter phrase means that the act was willful, knowing, and involved reckless indifference or gross negligence. The state of mind provision is known as *mens rea*. The law assumes as the default condition that adults and mature minors are moral agents, i.e. they are held responsible for the action unless exculpating conditions obtain.

The *mens rea* defense involves trying to prove that the defendant was not, in the conditions under which the action was performed, a moral agent in the full sense. Thus a defendant who can be shown to have acted under duress, who was insane when the act was performed, or who was involuntarily intoxicated will be excused. When the behavior was an automatism, such as sleep-walking or otherwise acting without consciousness, the defendant may be excused. Note also that the excuse from punishment does not entail that the defendant can walk free. Institutions for the criminally insane, for example, will house defendants who are found not guilty by reason of insanity, but who are considered a danger to society nonetheless.

Other factors may show diminished responsibility. These factors do not abolish responsibility. Rather, they play a role in the mitigation of the sentence. Diminished responsibility provisions are complicated, but roughly mean that the person was not in full control of his actions. For example, diminished responsibility may be acknowledged if the person has very low intelligence, or is mentally impaired by disease or injury, or if the action occurred in response to provocation sufficient to make him lose self-control, i.e. where the provocation was such that a reasonable person might well lose self-control. In England, this defense can be used only in the charge of murder, and the defense must show that the defendant suffered from an abnormality of mind such as to reduce his ability to have mental responsibility for the murder. If the defense is successful, the charge is reduced to manslaughter. The young are treated differently from adults on the grounds that the capacity for control develops as the child

matures, and that we cannot expect the same degree of control in adults and immature minors (Brink 2004).

The insanity defense requires proving first that the defendant is in fact insane or impaired. This requires the testimony of medical experts. In addition, the defendant must prove that the cognitive impairment rendered him unable to appreciate the criminal nature of the act. This requirement is generally known as the McNaghten Rules. In some states, another option is also available, which is to prove that the defendant, although cognitively able to appreciate the criminal nature of the act, was unable to conform his behavior to the requirements of the law. This option concerns volition rather than cognition.

Philosophers such as Van Inwagen (1983) and Allison (1990) argue that no one can justly be held responsible unless contra-causal free will exists. The sustaining assumption is that it would be unjust and immoral to hold someone responsible if his decision were caused by antecedent factors. In contrast, Hume, as we have seen, believed that it would be unjust to hold someone responsible unless the decision were caused—caused by his beliefs, hopes, desires, plans, motives, and so forth. Contemporary philosophers such as Van Inwagen tend to believe that because we do hold people responsible, agents must in fact enjoy contra-causal free will. The trouble with this view is that the facts about the universe, including the facts about the neural basis of choice, cannot be made to conform to our philosophical assumptions about our social institutions and conventions. The neurobiological facts are the neurobiological facts, whatever our social conventions might wish the facts to be.

My hypothesis is that holding people responsible and punishing the guilty is rooted not in some abstract relationship between a Platonic conception of justice and contra-causal willing, but in the fundamental social need for civil behavior. (As someone once whimsically said, there is no *Justice*, there is *just us*.) Humans are biologically disposed to be social animals, and, in general, one's chances of surviving and thriving in a social group are better than if one lives a solitary existence. In this respect, we are more like wolves and less like cougars. In social groups, cooperation, dividing labor, sharing, and reciprocity are essential for well-functioning—the civility of the group. Maintaining civility requires that children come to acquire the social practices of cooperation and so forth, and this acquisition typically involves both inflicting pain and bestowing reward. The young learn to value sharing, for example, and to suffer the consequences of failures to reciprocate.

The criminal justice system is a formal institution that aims to enforce laws regarded as necessary to maintain and protect civil society. To a first approximation, it functions to protect by removing violent offenders from society, to deter people who might otherwise violate the laws, and to provide a formal structure for revenge. The last function is not always emphasized, but its importance derives from recognition that humans who have suffered an injury are apt to retaliate, and that clan feuds can undermine the well-being of the group. Formal mechanisms for revenge are, on the whole, more consistent with civil stability.

Toward a neurobiology of 'in control' versus 'not in control'

Until the second half of the twentieth century, it was not really possible to explore systematically the neurobiological differences between a voluntary and an involuntary action. However, developments in neuroscience in the last 50 years have made it possible to begin to probe the neurobiological basis for decision-making and impulse control. Emerging understanding of the role of prefrontal structures in planning, evaluation, and choice, and of the relationship between limbic structures and prefrontal cortex, suggests that eventually we will be able

understand, at least in general terms, the neurobiological profile of a brain that is in control, and how it differs from a brain that is not in control (Fig. 1.1). More correctly, we may be able to understand the neurobiological profile for all the degrees and shades of dysfunctional control. For the time being, it seems most useful to frame the hypothesis in terms of 'being in control', where this is commonly assumed to involve the capacity to inhibit inappropriate impulses, to maintain goals, to balance long- and short-term values, to consider and evaluate the consequences of a planned action, and to resist being 'carried away by emotion'. Although these descriptions are not very precise, most people roughly agree that they describe what is typical for being in control, and agree on what behaviors are paradigmatic examples of compromised control.

The criminal law recognizes that control is lacking in automatisms, such as sleep-walking or movements made during and shortly following an epileptic seizure. Insanity is also understood to involve disorders of control, for example when the person does not appreciate that his thoughts or actions are in fact his. Other kinds of syndromes implicating compromised control include obsessive–compulsive disorder, where a patient has impaired inability to resist endlessly repeating some self-costly action such as hand-washing, severe Tourette's syndrome,

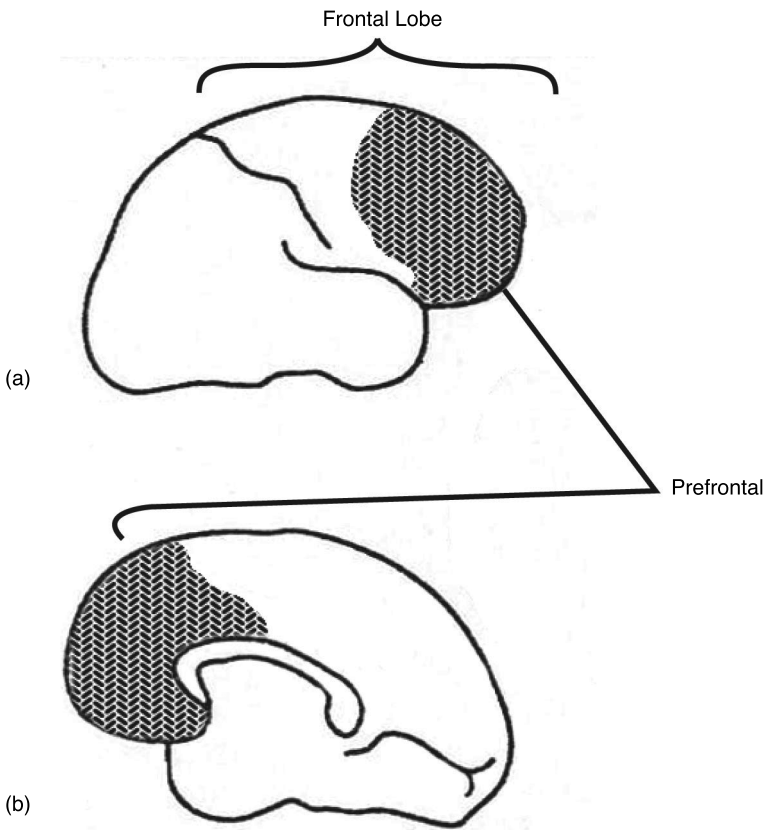

Fig. 1.1 Schematic drawing showing the regions considered to be the prefrontal cortex: (a) lateral view; (b) medial view.

where the person finds it almost impossible to inhibit particular ticking movements, or leptin disorder, where abnormally low levels of the protein leptin in the hypothalamus cause the person constantly to feel extreme hunger, no matter how much he eats. Obesity is the typical outcome of leptin disorder. Lesions to the prefrontal cortex have long been known to alter the capacity for planning, appropriate emotional response, and impulse control (Stuss and Benson 1984; Damasio 1994). The aim is to advance from a set of cases illustrating prototypical examples of control dysfunction to addressing the neurobiological basis of control and disorders thereof. Consequently, we need to determine whether there is a neurobiological pattern, or a set of common neurobiological themes, underlying the behavioral-level variations in compromised control.

What regions or systems in the brain are particularly important in maintaining control? Cortically, the structures that appear to have a pre-eminent role are anterior and medial: the orbitofrontal, ventromedial frontal, dorsolateral frontal, and cingulate cortices. Subcortically, all those regions that have a role in emotions, drives, motivations, and evaluations are important. These include the hypothalamus, amygdala, and ventral tegmental area (VTA) in the midbrain, and the nucleus accumbens (part of the basal ganglia, and a crucial part of the reward system) (Fig. 1.2). These are part of what is usually referred to as the limbic system, which also includes the cingulate cortex. Circumscribed lesions in these areas have helped reveal the interdependence of cognition, the emotions, and reward, both positive and negative (Stuss and Benson 1984; Damasio 1994; Schore 1994; Panksepp 1998).

Anatomical investigations of the patterns of connectivity among these regions, as well as between these regions and more posterior regions such as the parietal cortex, have begun to fill out the general picture of the dependency relations and the connection to action systems (Schore 1994; Fuster 1995; Panksepp 1998). Imaging studies have produced data that are generally consistent with the lesion data (MacDonald *et al.* 2000; Gray *et al.* 2003). For example, when subjects encounter conflict and must withhold or inhibit the typical response to a stimulus, activity in specific prefrontal and limbic regions increases. For other examples, chronic depression characteristically shows up as reduced activity in orbitofrontal cortex, and

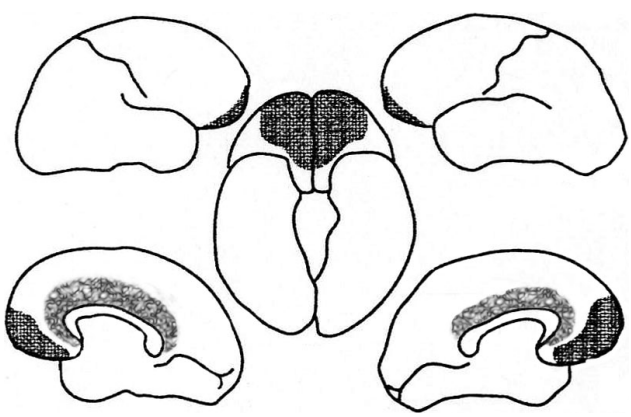

Fig. 1.2 Schematic drawing showing the location of the orbitofrontal cortex (hatched) and the cingulate cortex (shaded): upper, lateral view of the right and left hemispheres; lower, medial view; center, underside (ventral view).

disruption of connectivity between amygdala and prefrontal cortex results in an absence of normal activity in prefrontal areas in response to fearful or horrifying stimuli.

A common thread linking these anatomical regions is the so-called 'non-specific' neurotransmitter projection systems, each originating in its particular set of brainstem nuclei. There are six such systems, identified via the neurotransmitter secreted at the axon terminals: serotonin, dopamine, norepinephrine, epinephrine, histamine and acetylcholine (Fig. 1.3). These are referred to as the 'non-specific systems' because of their generality of effect and their role in modulating the activity of neurons. Abnormalities in these systems are implicated in mood disorders, schizophrenia, Tourette's syndrome, obsessive–compulsive disorder, social cognition dysfunctions, and disorders of affect. The pattern of axonal projection for each system is unique, and all exhibit extensive axonal branching that yields a very broad distribution of connections. The prefrontal and limbic areas of the brain are universal targets, with proprietary projections to specific subcortical structures such as the substantia nigra (dopamine), the thalamus (acetylcholine, serotonin), and thehypothalamus (norepinephrine, serotonin, acetylcholine).

On the surface, the aforementioned data from lesion studies, anatomy, neuropharmacology and so forth do not seem to coalesce into an orderly hypothesis concerning the neural basis for control. Looking a little more deeply, however, it may be possible to sketch a framework for such a hypothesis. Perhaps we can identify various parameters of the normal profile of being in control, which would include specific connectivity patterns between amygdala, orbitofrontal cortex, and insula, between anterior cingulate gyrus and prefrontal cortex, and so forth. Other parameters would identify, for each of the six non-specific systems, the normal distribution of axon terminals and the normal pattern of neurotransmitter release, uptake, and co-localization with other neurotransmitters such as glutamate. Levels of various hormones would specify another set of parameters. Yet other parameters contrast the immature with the adult pattern of synaptic density and axon myelinization (Sowell *et al.* 2003). At the current stage of neuroscience, we can identify the normal range for these parameters only roughly, not precisely.

Once a set of N parameters is identified, each can be represented as a dimension in an n-dimensional parameter space. Visually, one can depict at most a three-dimensional parameter space, and the extrapolation to n dimensions is conceived via generalization without

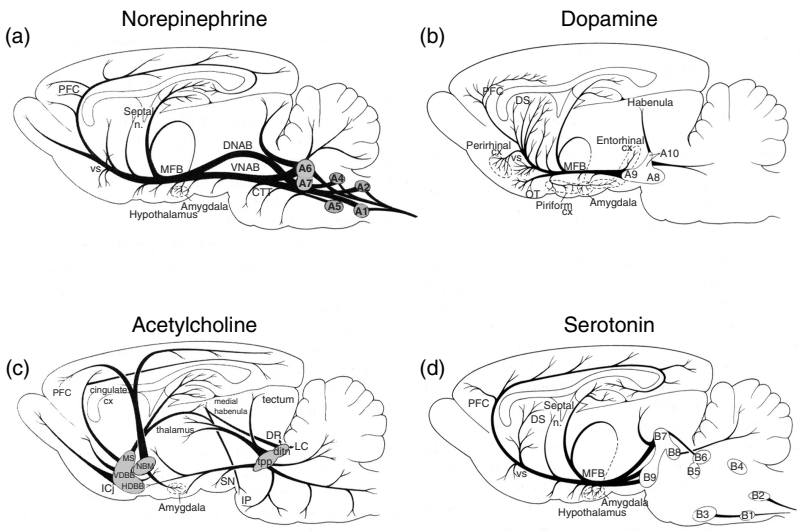

Fig. 1.3 Projection patterns for four non-specific systems originating in the brainstem.

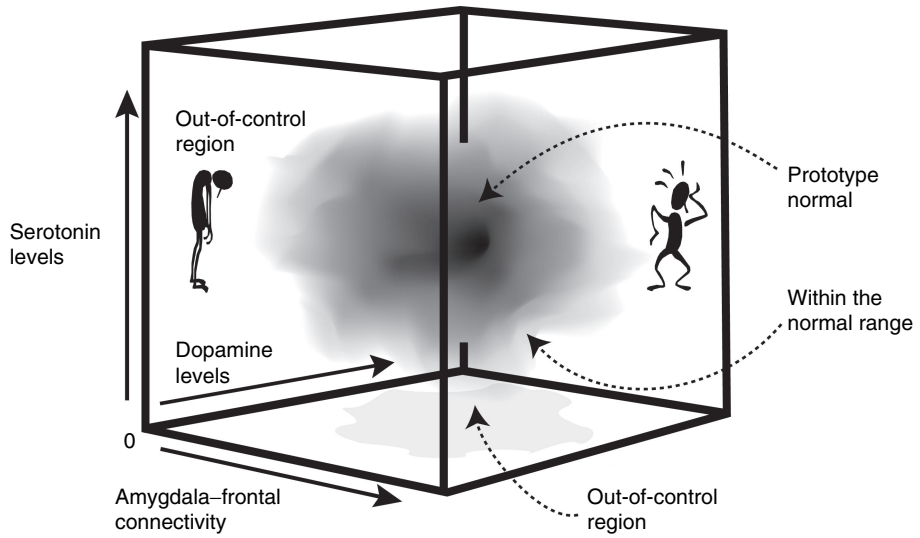

Fig. 1.4 Cartoon of parameter space showing an inner volume where the values of the parameters mean that the person is in control, and regions outside that volume where the parameter values mean that the person is not in control. Only three parameters are depicted, but in fact the parameter space is multidimensional. Notice that the boundaries are fuzzy rather than sharp. Dynamical properties are omitted.

visualization. Figure 1.4 depicts such a three-dimensional parameter space; for clarity, many dimensions are omitted.

The hypothesis on offer is that within the described *n*-dimensional parameter space, there is a volume such that when a brain's values for those parameters are within that volume, the brain is 'in control', in the sense in which I am using that term, i.e. the person's behavior exhibits those features implying that the person is in control. I suspect that the in-control volume of the control-space is rather large relative to the not-in-control space, suggesting that different brains may be in control by virtue of somewhat different values of the parameters. To put it simply, there may be many ways of being in control. Equally, there are also many very different ways of not being in control, i.e. of being outside the in-control volume. I also assume that the boundaries of the 'in control' volume are fuzzy, not sharp, suggesting that a brain may be 'borderline' in control; it may drift out of the volume, and perhaps drift back in again, as a function of changes in a parameter such as hormone levels.

From the perspective of evolutionary biology, it makes sense that brains normally develop so that they are in the preferred volume (in control). Even simple animals need to be wired so that they flee from a predator, despite being hungry, or that they seek warmth if they are cold, even though they are keen to find a mate. Species with large prefrontal cortex, such as primates, are able to defer gratification, to control impulses, to generate rather abstract goals, and to plan for future satisfactions. Individuals lacking control in these dimensions are at a disadvantage in the struggle for survival (Panksepp 1998; Gisolfi and Mora 2000; Dennett 2003).

Figure 1.4 is a cartoon. It is meant to be a conceptual tool for thinking about 'in control versus not in control' in terms of a parameter space and a preferred volume within that space. So as not to be misled, one must be clear about its limitations. For example, although some of

the parameters may interact with one another, this is not reflected in the diagram. Additionally, it is impossible to depict dynamical properties in a static diagram, although dynamical properties, such as changes during early development, adolescence, and senescence, certainly exist. Perhaps there are also diurnal changes or seasonal changes, and there is probably some context-dependency. Despite the many limitations of the diagram, the general concept of a control parameter space lends manageability to the hypothesis that there is a neurobiological basis in terms of which we can understand what is it for a brain to be in control. One can envisage much refinement to the basic conceptual point as neuroscience continues to discover more about prefrontal and limbic functions, and their role in planning, decision-making, self-representation, and evaluation.

Conclusions

The important core of the idea of free will consists not in the notion of uncaused choice, whatever that might be, but in choices that are made deliberately, knowingly, and intentionally; where the agent is in control. This aspect of our idea of free will is the justifying platform for reward and punishment, both in the informal institutions of childrearing and social commerce, and in the formal institutions of the criminal justice system. Because of developments in neuroscience and cognitive science, it is now possible to formulate a rough hypothesis concerning the neurobiology of 'in-control' brains, and the respects in which it differs from that of 'not-in-control' brains.

My proposal is that we frame this hypothesis in terms of a parameter space, the dimensions of which are specified in terms of neurobiological properties, especially of the prefrontal cortex, the limbic system, and the brainstem. As a consequence, 'in control' can be characterized neurobiologically as a volume within that parameter space. This provides a framework for further research on planning, decision-making, evaluation, and choice in nervous systems.

These developments in the biological sciences give rise to difficult but important issues concerning the possible revision and improvement of particular legal practices, especially in the criminal law (Goodenough and Prehn 2004). A wide range of potential policy changes need careful consideration; options need to be thoroughly articulated, explored, and debated in order for us as a nation to find our way toward wise policy decisions.

No single professional or social group is adequately equipped to solve these problems; no single group or person can claim moral authority to the answers. We will need to count on the thoughtful opinions and solid common sense of people everywhere—in industry, academia, the military, government, the press, religion, and business. Neither dogmatism nor intolerance nor self-righteousness will be an aid progress. Basically, we have to reason together to try to determine how best to proceed.

Aristotle believed in moral progress. In his view, as we search and reason about social life and its perils, as we experience life and reflect on its complexities and surprises, we come to a finer appreciation of what is decent and fair, and of the conditions conducive to human flourishing. We learn from each other, and from those whose lives exemplify the human virtues. We learn from the past—our own, and those in the history of our species. Aristotle's view is not a flashy theory of the Archetypal Good, nor it is the sort of theory to whip up moral zeal. Nevertheless, it is a reasonable and sensible approach to achieving some measure of human good, succumbing neither to invocations of the supernatural, nor to self-destructive skepticism. It is a pragmatic approach anchored by grace, dignity, empathy, and courage (see also A. Roskies, to be published).

References

Allison H (1990). *Kant's Theory of Freedom*. Cambridge, UK: Cambridge University Press.

Brink DO (2004). Immaturity, normative competence and juvenile transfer: how (not) to punish minors for major crimes. *Texas Law Review* **82**, 1555–85.

Campbell R, Hunter B (eds) (2000). *Moral Epistemology Naturalized*. Calgary: University of Calgary Press.

Casebeer, WD (2004). *Natural Ethical Facts*. Cambridge, MA: MIT Press.

Casebeer WD, Churchland PS (2003). The neural mechanisms of moral cognition: a multiple-aspect approach to moral judgment and decision-making. *Biology and Philosophy* **18**, 169–94.

Churchland PM (1989). *A Neurocomputational Perspective*. Cambridge, MA: MIT Press.

Churchland PM (2000). Rules, know-how, and the future of moral cognition. In: Campbell R, Hunter B (eds) *Moral Epistemology Naturalized*. Calgary: University of Calgary Press.

Churchland PS (1991). Our brains, our selves: reflections of neuroethical questions. In: Roy DJ, Wynne BE, Old RW (eds) *Bioscience and Society*. New York: Wiley, 77–96.

Churchland PS (2002). *Brain-Wise: Studies in Neurophilosophy*. Cambridge, MA: MIT Press.

Damasio AR (1994). *Descartes' Error*. New York: Grosset/Putnam.

Dennett D (2003). *Freedom Evolves*. New York: Viking.

Flanagan O (1996). *Self-Expressions: Mind, Morals and the Meaning of Life*. Oxford: Oxford University Press.

Fuster JM (1995). *Memory in the Cerebral Cortex*. Cambridge, MA: MIT Press.

Gisolfi CV, Mora F (2000). *The Hot Brain: Survival, Temperature, and the Human Body*. Cambridge, MA: MIT Press.

Goodenough OR, Prehn K (2004). A neuroscientific approach to normative judgment in law and justice. *Philosophical Transactions of the Royal Society of London, Series B* **359**, 1709–26.

Gray JR, Chabris CF, Braver TS (2003). *Nature Neuroscience* **6**, 316–22.

Howard R, Lumsden J (1996). A neurophysiological predictor of reoffending in special hospital patients. *Criminal Behaviour and Mental Health* **6**, 147–56.

Hume D (1739). *A Treatise of Human Nature* (ed. Selby-Bigge LA). Oxford: Clarendon Press, 1967.

Illes J, Raffin T (2002). Neuroethics: a new discipline is emerging in the study of brain and cognition. *Brain and Cognition* **50**, 341–4.

Kane R (ed) (2001). *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Korsgaard C (2001). *Self-Constitution: Action, Identity and Integrity*. The John Locke Lectures, Oxford University 2001–2002. Copies of the lecture handouts are available online at http://www.philosophy.ox.ac.uk/misc/johnlocke/index.shtml.

MacDonald AW, Cohen JD, Stenger VA, Carter CS (2000). *Science* 288, 1835–8.

Panksepp J (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York: Oxford University Press.

Pereboom D (2001). *Living Without Free Will*. Cambridge, UK: Cambridge University Press.

Roskies A (2002). Neuroethics for the new millennium. *Neuron* **35**, 21–3.

Schore AN (1994). *Affect Regulation and the Origin of Self*. Hillsdale, NJ: Lawrence Erlbaum.

Sowell ER, Peterson BS, Thompson PM, Welcome SE, Henkenius AL, Toga AW (2003). Mapping cortical change across the brain. *Nature Neuroscience* **6**, 309–15.

Stuss DT, Benson DF (1986). *The Frontal Lobes*. New York: Raven Press.

Van Inwagen P (1983). *An Essay on Free Will*. Oxford; Clarendon Press.