16

The Neurobiology of Individual Decision Making, Dualism, and Legal Accountability

Paul W. Glimcher

Departments of Neural Science, Economics and Psychology, Center for Neuroeconomics, New York University, New York, NY 10003, U.S.A.

Abstract

Introduction

Over the course of the last decade there has been an increasing interest in neurobiological analyses of the causes of behavior among many practitioners of criminal law. In some institutional circles this has crystallized as an interest in providing a physical method for classifying the actions of human agents according to preexisting social–legal categories. The impetus driving this search for neurobiological classification tools stems from both the longstanding Western legal requirement that actors be held accountable only for those voluntary actions which are preceded by what is termed a culpable mental state (those for which *mens rea* can be established) and the longstanding legal difficulty in establishing culpable mental state at trial. Thus a pressing question for many legal practitioners is whether existing neurobiological techniques or data can be used to identify the socially defined categories that guide law and punishment.

If neurobiological measurements did suggest a division of the physical causes of behavior into biological categories closely aligned with social categories, then neuroscience might indeed be very useful for making this legal distinction. In contrast, if neurobiological data suggested that (at a physical rather than a social level) no fundamental division of behavior into these categories could be supported, then we would face a social dilemma of sorts. We would have to decide whether the social consensus that supports differentially

punishment of actors based on their psycho–legal mental state should persist even if there is compelling neurobiological data that the physical causes of real human actions cannot be divided into the category or categories this legal classification imposes.

In this chapter, I suggest that both the modern epistemological views of natural scientists and the available neurobiological evidence indicates that there is no meaningful sense in which the possible states of the brain can be reduced to a standard psycho–legal state. Indeed, our current level of understanding suggests that at an empirical neurobiological level the distinctions employed by the criminal justice system may be nearly meaningless. These data suggest that there is no compelling evidence for a neural partition that uniquely includes either "rational," voluntary, conscious, or for that matter even "unemotional," mental states.[1]

If in the near future these hints from the neural data become certain conclusions, society and its institutions will face an interesting problem. We will have to decide whether to continue to regard socially defined categories of behavior, like the legal notions of rational and irrational, as fundamental to our institutions.

The Law and Neuroscience

The Legal Classes of Behavior: Involuntary, Compelled, or Rational

Nearly all extant Western legal systems reflect a widely held social conviction that whether an actor should be punished for actions that can be attributed to the actor depends on more than whether the agreed upon events contravene legal statute. Indeed, even whether an agreed upon physical event (e.g., "Jill's arm put a knife into Jack,") can or cannot be considered an "action" in the legal sense depends on Jill's mental state when the knife entered Jack's body. Whether Jill is responsible for a crime depends upon a series of behavioral classification judgments which seek, ultimately, to separate actions into those for which an actor should be held responsible and those for which the actor should not be held responsible. In a perhaps overly simplified sense, the goal of these classifications is to establish whether the observed act was the conscious product of a rational mind, or whether it reflected involuntary, unconscious, or automatic processes as defined at a psycho–legal level.

To make these classifications clear (or at least clearer) to psychologists and neurobiologists, consider the example of an individual who enters knowingly an environment in which it is illegal to utter a profanity. In a first example, the actor suffers from a suddenly acquired and quite severe case of Tourette's

[1] There may be, however, significant evidence being developed to support the conscious/non-conscious distinction in a manner that separates it from the rational/irrational distinction (Dehaene et al. 2006)

syndrome and while in the environment swears loudly. A crime? No. The existing legal structure leads us to conclude that this action was "involuntary"; it was committed without the "intent" to swear. In the absence of a voluntary action, there is no crime. In a second example, the actor is approached by an individual with a knife and told to swear loudly on pain of death. The actor swears. A crime? In this case there is no doubt (according to standard legal definitions) that the actor swore voluntarily, but he did so under an external compulsion to which a normal person would have yielded. No crime. A third actor steps on an exposed nail and in response to the bolt of pain from his foot swears loudly. A crime? There are two ways to go here that yield the same result. We can conclude that the act was involuntary; the actor could not have done otherwise given the pain in "the foot," so no crime has been committed. Alternatively, we could conclude that the action was the irrational product of the pain. More precisely we conclude that a normal person experiencing this pain could not be expected to control his vocal behavior rationally. Again, no crime. A final agent enters the same environment and seeing someone he dislikes swears loudly at them. In legal terms this is a voluntary act, and one committed by a rational person. A crime.

To psychologists and neurobiologists, these categories may seem a bit arbitrary, but they provide the tools by which legal professionals categorize behavior into punishable and nonpunishable (or less punishable categories). In addition, these two main categories (punishable and nonpunishable) are ones to which nearly all humans are committed. To make that clear, consider two cases. Two men return home to find their wives in bed with other men. In the first case the husband, enraged by this act of infidelity and seeing a loaded gun on the table, shoots and kills his wife. In the second case the husband deliberates for two days, then goes to a gun store, purchases a gun, and returns home to kill his wife. In almost all Western legal systems, the first man's rage is at least partly exculpatory because that rage impaired his rationality; it caused him to behave irrationally at the time of the shooting. The absolutely key point then becomes: was the shooting a "rational" act? If, at the time of the shooting, the man was acting "irrationally" (or more precisely was incapable of acting rationally) or "involuntarily," then he is innocent.

So what are exactly these legal categories that determine culpability? Acting "rationally" here has a very specific legal meaning. It does not mean that the man was producing behavior that maximized anything in the economic sense, although the sense of calculation this evokes does suggest rationality of the legal kind. It does not require that he was engaged in explicitly symbolic psychological processing, although the sense of this class of processing is what the legal definition clearly invokes. It does not mean that regions of frontal cortex were at least potentially in control of his motor system, although again this is a related sense of the word. According to the Oxford English Dictionary "rational action" in this sense means: exercising (or able to exercise) one's reason in a proper manner; having sound judgment; sensible, sane.

The critical notion that emerges from these two criteria is that actors can be in one of two states (although there may be gray areas between these two states): rational with intent to commit a crime or everything else. If you are rational in this sense at the time of the action, you are guilty; otherwise you are innocent, or at least less guilty. This notion of rational intent is closely tied to terms from psychology such as voluntary, conscious, or (to follow a more recent psychological trend) simply System 2. It is criminal behavior that rests within this category which Western legal systems tend to punish most severely. When an actor commits a crime while in a rational mental state, then he receives maximal punishment. The man above who deliberated and purchased a gun before killing his wife provides an example of this kind of deliberative, rational action.

Behaviors for which actors are not criminally responsible (or at least less responsible) are those lawyers call involuntary or irrational, and this category is tied to psychological (but not legal) notions of involuntary behavior, nonconscious, automatic, (sometimes) emotional, or simply System 1 behavior. If it can be demonstrated that a specific criminal act engaged an emotional, automatic, or involuntary process that limits or eliminates the agent's ability to act "rationally," then the process itself becomes exculpatory. There are many examples of behavior of this type, and it cannot be overstated that the Western legal tradition places lighter sanctions on actors guilty of crimes that can be attributed to this class of mental state. If, for example, an actor can be shown to be enraged (under some circumstances), this can mitigate against punishment for a crime.

While each of these categories (rational and the super-category of all exculpatory states) constitutes, in a sense, a doctrinal universe of its own, each of them has recently become of interest to neurobiologically minded lawyers. Precisely because lawyers have become interested in the relationship between neurobiological measurements and rational versus exculpatory classes of behavior, these categories are the subject of this review. These categories are often presumed by legal professionals to be particularly tractable to neurobiological analysis, and it is with that in mind that we begin by examining the social, ethical, and physical basis of the rational/exculpatory distinction—why it is that we punish.

Consequentialist vs. Retributivist Justice

One of the most interesting features of judicial systems, and a feature that engages closely the categorization of behavior into rational and involuntary-irrational, is the underlying reason why institutions punish actors who commit illegal acts. Speaking very broadly, reasons for punishment can be classed as either retributivist or consequentialist. In these simple black and white terms, a consequentialist legal system is one in which punishments serve only to reduce or eliminate crimes. In designing a consequentialist legal code, for example,

one need only consider whether a proposed punishment will reduce crime or improve Society. In contrast, a retributivist system seeks to punish actors who commit crimes specifically because their actions deserve punishment in an ethical or moral sense.

One place where this distinction becomes particularly clear might be in a hypothetical discussion of capital punishment. Consider an imaginary Society in which one could chose between imprisoning particularly heinous criminals for life or executing them. Let us further assume that the monetary cost of executing one of these criminals was higher than the monetary cost of life imprisonment. If we observed that the Society executed criminals, what would this tell us? We might draw one of two conclusions. Working from Jeremy Bentham's consequentialist analysis, we might conclude that the rate of murders was probably being reduced by the visible execution of murders. However, what if we had unimpeachable evidence that executing these criminals had no effect on the rate at which these crimes were committed *and* that everyone was aware of this evidence? Under these conditions a perfect consequentialist system would never perform executions, for executions neither decrease crime nor reduce costs. If one observed executions under these conditions, one could conclude that one was studying a retributivist system of justice. Of course real legal systems do not admit such a simple dichotomous analysis. Executions may reduce murders, or at least members of the Society might believe that executions reduce murders. On the other hand, it does seem that there may be more than just consequentialist motives at work in many legal systems.

The reason I bring this up is because retributivist systems seem to rely particularly heavily on the voluntary-rational/involuntary-irrational distinction for their existence. To make this clear, consider a legal system that lacked this distinction. Actors commit crimes. Demonstrating that an individual killed someone, a simple finding of fact, would be a complete finding of guilt. Once we determine that fact-finding establishes guilt, then we must decide whether everyone receives the same punishment for the same crime. Without the voluntary-rational/involuntary-irrational distinction there can be only one reason for differential punishment of our different actors: The conclusion that minimizing the risk of future crimes required that different degrees of punishment be meted out to different individuals. If there is a single category of behavior then we punish all actors who commit that externally observable crime in the same way, or we adjust punishment individual-by-individual so as to minimize future crime.

However, and this may not be immediately obvious to natural scientists, this is not what we often observe in the Western legal tradition, particularly in the U.S. tradition. Some actors are punished more severely than others and not because those receiving harsher punishment require a higher level of deterrence. Instead, the force of punishment in these Western laws reflects both our convictions about the actor's culpability for his or her actions and considerations about the positive effects of punishment for Society at large. To make

that clear, let us return to the example of the two men who kill the wives. Why do we punish more severely the man who deliberates for two days than the one who kills impulsively while enraged? Certainly not because we want to discourage deliberation when enraged. We punish the deliberating man because he *deserves* more punishment, because his murder was rational and voluntary. If we were simply attempting to deter crime, we might even conclude that it was the enraged man who should be punished more severely. We might well hypothesize that only the threat of very severe punishment has a hope of deterring someone in that mental state. In other words, the structure of our punishment system is, if anything, anti-consequentialist. Behaviors that are more automatic would probably be punished more harshly in a consequentialist system. We observe the opposite in our system. People are largely being punished because they deserve it.[2]

Many natural scientists may object to the conclusion that retributivist approaches to justice reflect a social consensus. They may argue instead that retributivism in our culture reflects an antiquated feature of the Western legal tradition, and one that is on the wane. It is important to point out that this is simply not the case. Notions of fairness drive human behavior in a wide variety of situations, many of which have been well studied by social scientists. Consider, for example, the ultimatum game popular in behavioral economics (Blout 1995; Guth et al. 1982; McCabe et al. 2001). Two players in different cities, who have never met and who will never meet, sit at computer monitors. A scientist gives one of these players $10 and asks her to propose a division of that money between the two strangers. The second player can then decide whether to accept the proposed split. If she accepts, the money is divided and both players go home richer. If she rejects the offer, then the experimenter retains the $10, and the players gain nothing. What is interesting about this game is that when the proposer offers the second player $2.50 or less the second player rejects the offer. The result is that rather than going home with $2.50, the second player goes home with nothing. Why does she give up the $2.50?

We might derive a partial answer to this question from examining what happens when the first player is replaced by a computer program which, the second player is informed, plays exactly the same distribution of strategies that real human players employ. Under these conditions, if the computer offers $2.50 the second player almost invariably accepts it. Why?

The standard interpretation of this finding is that players refuse the $2.50 from a human opponent because they perceive the offer to be unfair. They want to punish the proposer by depriving her of $7.50, which the players have

[2] This point (i.e., that our legal systems are significantly retributivist in structure) is widely acknowledged in legal circles, although it may be unfamiliar to natural scientists. In developing this point in more detail, the American legal scholar, Owen Jones, has argued that we must be aware of the biological imperatives that drive some classes of crimes when we design criminal codes. This issue was addressed in his famous article, "The Law of Law's Leverage" (Jones 2001).

determined that she does not deserve. Players will happily spend $2.50 to achieve this goal. When playing a computer, whose actions they see as mechanical and involuntary, they happily accept the $2.50. I think that we have a clear pattern here. Under conditions ranging from splitting $10 to murder, nearly all people indicate a conviction that actors should be punished when they voluntarily act unjustly. One can speculate about the evolutionary roots of this taste for justice (e.g., Brosnan and deWaal 2003) or this drive for fairness, but there is no escaping the conclusion that it is an essential feature of human behavior today.

Wherefore Neuroscience in Law?

In the preceding sections, I hoped to make two points. The first was that the Western judicial tradition, and nearly all members of Western societies, possesses a preexisting consensus that there is a social distinction between voluntary-rational and involuntary-irrational behavior (among a set of several distinctions of this type). This is a deeply fixed element in our institutional designs and probably in the evolved biological fabric of our brains. The second point was that our judicial systems are at least partially retributivist in nature. Like the distinction between these two classes of behavior, this reflects a strong social consensus. It is important, however, to note that these two principles interact. We take retribution for actions socially defined as voluntary and rational. For acts socially defined as involuntary, irrational, or compelled we mete out limited punishment—punishment that is often more consequentialist in nature.

The point that we have to keep in mind here is that, in a very real sense, this works. We have a working social consensus, so what possible role could neuroscience play in any of this from the point of view of a practicing lawyer or judge? The answer stems from the institutional need to segregate (in particular) voluntary and rational behavior from these other classes of behavior which are deemed exculpatory[3]. This is a fundamental problem with the system as it currently exists, and there is much hope in some legal circles that neuroscientific evidence can be used to segregate voluntary, conscious, and rational acts from involuntary, nonconscious, irrational acts. The reason this is necessary should be immediately obvious: Because we punish voluntary-rational acts more severely, it is in the interest of all defendants to establish that their crimes were committed in an irrational mental state. This means that juries and judges are often in the position of trying to decide whether an act was rational. The question in legal circles, then, is: Can neurobiological data, for example from a

[3] Currently, there is also great interest in the possibility of using neuroscientific methods to establish issues of liability: "Did she do it?" The rising interest in using brain scanners to identify lying is an example of this interest. Here I am discussing only the issue of responsibility which bears on the voluntary/involuntary distinction. In considering issues of liability, the reader is referred to Wolpe et al. (2005) and Garland and Glimcher (2006).

brain scanner, be used to identify voluntary-rational behavior as the preexisting legal system defines it?

At least logically, the answer to this seems like it should be yes. Imagine that we used our social consensus to label each of one million particularly unambiguous crimes as voluntary-rational or excusable: five hundred thousand in each category. Then imagine that we subjected all one million of these individuals to brain scans. Intuitively, it seems obvious that such an endeavor would yield a portrait of the distinction between voluntary-rational and involuntary-irrational crime at the neural level. But would this really have to work?

To answer that question, consider more mathematically what we are trying to accomplish. During the brain scans, we measure the average activity of each cubic millimeter of brain, or voxel. Imagine that we scanned only three of the cubic millimeters in each individual. Then we could represent the activity of each brain as a trio of measurements, the activity in each of the three voxels, a point on a three-dimensional graph. Now imagine that we do this on 500,000 people guilty of voluntary crimes. The result is a cloud of points, we can imagine them colored red, in this three-dimensional space. Then we do the same thing for the people who committed crimes involuntarily. They produce a second cloud of points, imagine them colored green. Of course in a real brain scan we would observe the activity of about 60,000 voxels simultaneously so these clouds of points would be distributed in a much more complicated (60,000-dimensional) space, but we have good mathematical tools for moving back and forth between these kinds of graphs so the higher dimensional space presents no conceptual barrier that we cannot overcome. So having generated a graph of these one million points, here is the critical question that we want to answer: Can we draw a circle[4] around the voluntary-rational points that includes none (or at least very few) of the exculpatory-state points? That is the critical question.

If the answer is yes, then whenever we want to establish if an actor committed a crime in a culpable mental state, we place her in a brain scanner. If her point falls within the circle, then she is guilty. What this analysis reveals is that for this approach to work, the clouds of points must land in separate places in the 60,000-dimensional space; if the points are all intermingled, then the approach will fail. Now if, for example, conscious rational-voluntary acts (as defined legally) are the specific product of a single brain area; if brain area X, made up of 1000 voxels, was more active when a crime was committed in a culpable mental than when it was committed in an exculpatory state this would cause the two clouds of points to separate, then the method will work. If, on the other hand, there is no coherent logical mapping between the states of these brain voxels and the legal notion of culpable mental state, then this approach is doomed to failure.

[4] More formally, we would be searching for a hyperplane in the 60,000-dimensional space after computing and removing the covariance matrix.

Although I have explained this logic for a contemporary brain scanner, it is important to note that what I have explained is true, in principle, for all of the neural (or more generally all physical) measurements we could imagine making. Consider measuring the brain levels of a single neurotransmitter. If culpable crimes are uniquely associated with low levels of this neurotransmitter, then the method will work. If there is a completely overlapping distribution of neurotransmitter levels in the culpable and exculpatory groups, then the method will fail. The same is true for some bigger and better future brain scanner. If there is any feature of the anatomy or physiology of the human brain that can support a partition of behavior into these categories, then neuroscience will be relevant to this problem. If no feature of the natural structure of the brain can support this categorization of action into two domains, then neuroscience will not be of use, and it may even call into question the wisdom of this categorization depending upon your convictions about the relationship between physical and social phenomena.

To proceed, what we have to do next is to understand how likely it is that neuroscientific evidence can support the division of behavior into these two categories. It is certainly true that both philosophy and physiology seemed to support the existence of these two categories up until about a hundred years ago, but that may be changing. So next we turn to both the epistemology and physiology of voluntary-rational action. What we will find is compelling evidence to discard, at the neurobiological level of analysis, the philosophical notion of Free Will. Secondary to that conclusion, we will also find that the available empirical evidence leans against the notion that the socially defined categories of voluntary-rational and involuntary-irrational can be identified neurobiologically. While this second point is admittedly a preliminary empirical conclusion, it will raise some potentially troubling issues.

Neuroscience and the Law

To understand how neuroscientists see the issues of voluntary-rational and involuntary-irrational behavior, it is critical to recognize what it is that neuroscientists are trying to understand when they study the brain. Consider the hotly debated topic of face perception. The human ability to identify familiar faces is astonishing. Show a human a stack of 24 pictures of different peoples' faces for a few seconds each. Wait 15 minutes. Then show the person a stack of 48 faces: the 24 old faces and 24 new faces. The average human can correctly pull out 90% of the faces that they saw (Bruce 1982). The average human cannot do this with pictures of sheep, pictures of cars, or pictures of houses (cf. Kanwisher and Yovel 2006; Gobini and Haxby 2007). Now how does this ability arise at the mechanistic level?

Prior to the research that yielded the neuroscientific evidence on this issue, there were two theories. Theory 1 argued that humans had a highly specialized

(and probably anatomically localized) brain system for recognizing human faces: a machine, in our heads, optimized for the recognition of faces and nothing else. In essence, this theory argued that there was a face recognition system and an "everything else" recognition system. It was a two-system view of face recognition. Theory 2 argued that there was only one system for general recognition, but this theory went on to argue that since humans spend so much time recognizing human faces in normal everyday life, this general-purpose system happens to be particularly good at recognizing faces. This theory, in essence, argued that while it is true that we are better at recognizing human faces than houses, this ability arises from the actions of a single system.

Brain imaging weighed in on this issue when it was discovered that if you show a human subject pictures of faces for two minutes and then pictures of houses for two minutes, you see a very differential pattern of brain activation (Kanwisher at al. 1997). When humans view faces, a small region in the temporal lobe of the brain, now called the fusiform face area, becomes highly active. This phenomenon is highly reproducible. In fact, the level of activation in this area can be used to determine whether a human subject is looking at a picture of a face or a house even if she refuses to tell you verbally. So what does this mean? First, and unarguably, this means you can tell what a human subject is looking at in this example, but does this also mean that neuroscientists have concluded that the brain can be usefully described as being made of two recognition systems: one for faces and one for other things? The answer to this is much more complicated and far from settled. To resolve that question, one group of neurobiologists trained a small group of humans to become experts at recognizing a class of non-face geometric objects called "greebles" and then asked whether "greeble experts" used the "fusiform face area" when they recognized greebles. Their evidence suggests that this was the case, and so they concluded that either (a) there is a brain area responsible for "expert" recognition and a more general system for "nonexpert" recognition or (b) there is no expert system, but the fusiform face area is active for any recognition problem which is difficult and for which subjects are well trained (Gauthier et al. 1999).

These experiments were followed by a host of other experiments: on car recognition by car experts, studies of patients with damage to the fusiform face area (who cannot recognize faces but can perform normally at many other recognition tasks), recognition of inverted faces (Haxby et al. 2001). In total, literally millions of dollars and hundreds of first-class scientists struggled to resolve this problem. The problem that they wanted to resolve was whether the architecture of the brain could be described as having a face recognition system and an "other-stuff" recognition system in a deep and natural sense. They were asking: Is there an intrinsic distinction in the neural architecture between face and non-face? There may not yet be a consensus on this issue, but at the moment it is probably safe to say that more neuroscientists believe that face recognition is accomplished by a discrete system than feel otherwise. Still, this

is an incredibly controversial issue, and the case is far from closed. It does, however, look like faces may be a natural neural category.

Is there a similar neural distinction between voluntary-rational and involuntary-irrational? Can we determine whether there is a meaningful neural category for voluntary-rational behavior? Or at least the possibility of segregating voluntary-rational states from all other brain states? To answer those questions, neurobiologists take two approaches: an epistemological approach that asks whether, in principle, there are reasons to expect such a category and an empirical approach, such as the one described above, for face perception. First, let us turn to the philosophical roots of the voluntary-rational versus involuntary-irrational problem.

Epistemic Beliefs about the Voluntary vs. Involuntary Distinction

Free Will

Prior to the enlightenment, Aristotelian and Platonic notions of the mind dominated scholarly debates about the sources of human action. Aristotle had concluded, largely in *De Anima*, that all living beings possessed souls and that the complexity of these nonphysical objects increased as the mental complexity of living beings increased. Aristotle saw the nonmaterial souls of humans as causal agents uniquely responsible for observed behavior. Although questions were raised during this period about whether souls really occurred in nonhuman organisms, the notion that a nonmaterial soul was the unique causal agent responsible for human behavior became a widely held idea. During the Reformation, this notion was challenged by emerging Protestant theologians (e.g., Luther and Calvin) who, working from the ideas of Aquinas and Augustine, developed the doctrine of predestination. This doctrine argued against the classical notion of a causal soul on at least two grounds. First, the doctrine argued that only God could act as a causal agent. Second, and more importantly, the doctrine recognized that making the human soul a causal agent independent of *God's will* or *God's omniscient knowledge* meant that at least some future events would be unknown to God. This set up a conflict between the omniscience of God and the causal independence of the human soul. They resolved this conflict by concluding that although there was a nonmaterial human soul, it was not a causal agent. (This is a theology typically referred to as "double predestination" and is particularly associated with the works of Calvin). Thus according to the early Protestants, human action was the mechanical product of a deterministic soul that reflected a preordained divine program. For these theologians, and a major segment of the European intelligentsia during the Reformation, human actors were, at least at a philosophical level, viewed as entirely deterministic devices.

Within the Catholic Church, this philosophy prompted a heated debate during the 16th and 17th centuries. At that time, a group of scholars centered around the Flemish theologian, Cornelius Jansen, argued that this Protestant notion was logical and compatible with Catholic doctrine, while the Jesuits and others close to the Papacy believed that the human soul must be viewed as a causal agent for the notions of moral judgment, especially salvation and damnation, to be meaningful. In the end, the Jesuits triumphed in this debate, and much of Jansen's writings were declared heretical by the middle of the 17th century. The result was that the Catholic Church adopted initially a unitary classical position on the causes of human behavior: The soul of man makes choices that are causally responsible for all actions taken by a person. The choices that the soul makes are unconstrained; man is free to produce any action. God judges all of those actions as just or unjust, and on this basis saves or damns each individual.

For our purposes, I want to draw attention to the fact that neither of these approaches is really compatible, at a deep philosophical level, with the Western retributivist tradition, which rests on two classes of behavior. If all human action is equally predestined, then how can one support different levels of culpability based on mental state? The same question, in reverse, can be asked of the Catholic tradition: If all human action is the product of a unitary voluntary-rational mechanism, then how can we single out some actions as exempt from punishment? The existence of these philosophical traditions, however, had very little impact on the ongoing Western legal tradition. When the city of Geneva became Protestant, it did not abandon Roman law. The law and its psycho–legal consensus were effectively partitioned from these metaphysical conclusions. I mention this because it is a point to which we will have to return later.

In any case, it was the reconciliation of these two models, in the 17th century, that led to the widely held modern metaphysical and empirical conviction that there are at least two independent sources of human action. I turn next to the reconciliation that occurred principally within the Catholic Church.

The Catholic Dualization of Human Action

The history of this dualization began with the surge in anatomical studies conducted during the 1500s, which revealed that the human physical corpus was surprisingly material and much more mechanistic than had been previously supposed. Great European anatomists, like Vesalius, Sylvius, and Fallopius, catalogued the intricacies of human anatomy meticulously (cf. Vesalius 1543/1998). Concurrently, the progenitors of physiology sought to provide clear mechanistic explanations of the functions of these human anatomical components. William Harvey (1628/1995), to take the most prominent example, went so far as to explain that the human heart, which Aristotle (in *De Anima*) had identified as the physical location of the soul, was nothing more

than a pump. These scientists, modeling their physiological investigations on the emerging field of mechanics, described simple cause and effect relationships that accounted for the actions of the human body. Basing their understanding of physiology on the clockwork deterministic mechanisms that were being invented every day during this period, they began to perceive the human body as a predictable machine.

Descartes' avowed goal was to provide a philosophical basis for understanding the material interactions that governed the physical world while preserving the Catholic notion that the actions of the human soul were a distinct category of events that could not be deterministically tied to the material events and processes of the corporeal realm. He accomplished this goal by dividing all of human behavior into two exclusive categories. He presumed the first category to be the deterministic product of the physical body—a class of behaviors we now often call *reflexes*. The second category included those actions that could be attributed to the causal force of the human soul. These *voluntary* actions were characterized empirically by being unpredictable, and it was Descartes' conclusion that the causes of these actions lay outside the material world and thus could not form the subject of physiological inquiry. Descartes thus opened a segment of human behavior for physiological study while reserving to voluntary behavior an extra-physical notion of agency. However, from a legal and ethical point of view, Descartes did something else: for the simple behaviors that he classed as *deterministic reflexes*, there could be no (or at the very least a diminished) moral culpability; for the more complicated and unpredictable behaviors, there was complete moral culpability. The distinction between voluntary and involuntary was reified at both the metaphysical and empirical levels, thus removing a preexisting tension between the law and philosophers (natural and otherwise).

The critical feature of the scientific study of voluntary and involuntary behavior that I am trying to bring out here is a debate about whether the causes of behavior can be usefully described as intrinsically breaking apart along the voluntary-rational and involuntary-irrational boundary. Descartes argued that this would be the case: that involuntary behaviors, reflexes, would be shown to have a straightforward physical implementation and that if one could trace a deterministic path through the nervous system that accounted for the behavior then it could be labeled involuntary. The voluntary behaviors, he argued, could not be deterministically accounted for by simple mechanical and deterministic components; these behaviors were the product of what philosophers now term "contra-causal metaphysical libertarian freedom."

One really nice feature of this approach is that it suggests that the neurobiological approach will identify two categories of mechanisms that map directly to the social–legal notion of voluntary-rational/involuntary-irrational. In other words, if neurobiologists persist long enough, then we will have simple and direct neurobiological markers for the voluntary-rational/involuntary-irrational

distinction. What is most amazing to many neurobiologists is that by the early 20th century, the courts had begun to settle on this approach.

A particularly clear example of this is a U.S. Supreme Court decision from the early 1900s: Hotema vs. the U.S. The case concerns a plaintiff, Solomon Hotema, who was convicted of killing Vina Coleman on April 14, 1899. Hotema plead not guilty by reason of insanity, although this defense failed. In evaluating his insanity plea the court noted:

> A safe and reasonable test is that whenever it shall appear from all the evidence that at the time of committing the act the defendant was sane, and this conclusion is proved to the satisfaction of the jury, taking into consideration all the evidence in the case, beyond a reasonable doubt, he will be held amenable to the law. Whether the insanity be general or partial, whether continuous or periodical, the degree of it must have been sufficiently great to have controlled the will of the accused at the time of the commission of the act. Where reason ceases to have dominion over the mind proved to be diseased, the person reaches a degree of insanity where criminal responsibility ceases and accountability to the law for the purpose of punishment no longer exists.

The decision continues in this vein, citing the lower court's decision:

> The real test, as I understand it, of liability or nonliability rests upon the proposition whether at the time the homicide was committed Hotema had a diseased [186 U.S. 413, 417] brain, and it was not partially diseased or to some extent diseased, but diseased to the extent that he was incapable of forming a criminal intent, and that the disease had so taken charge of his brain and had so impelled it that for the time being his will power, judgment, reflection, and control of his mental faculties were impaired so that the act done was an irresistible and uncontrollable impulse with him at the time he committed the act. If his brain was in this condition, he cannot be punished by law. But if his brain was not in this condition, he can be punished by law, remembering that the burden is upon the government to establish that he was of sound mind, and by that term is not meant that he was of perfectly sound mind, but that he had sufficient mind to know right from wrong, and knowing that the act he was committing at the time he was performing it was a wrongful act in violation of human law, and he could be punished therefore, and that he did not perform the act because he was controlled by irresistible and uncontrollable impulse. In that state of case the defendant could not be excused upon the ground of insanity, and it would be your duty to convict him. But if you find from the evidence, or have a reasonable doubt in regard thereto, that his brain at the time he committed the act was impaired by disease, and the homicide was the product of such disease, and that he was incapable of forming a criminal intent, and that he had no control of his mental faculties and the will power to control his actions, but simply slew Vina Coleman because he was laboring under a delusion which absolutely controlled him, and that his act was one of irresistible impulse, and not of judgment, in that event he would be entitled to an acquittal.

What is clear here is that the court is separating behavior into two possible categories: those which are rational and those which are irrational or involuntary—a perfectly normal, if in this case somewhat ambiguous, legal thing to do. Rational behaviors are subject to legal sanction; irrational, involuntary, or compelled actions are not. A second feature of the decision is the court's effort to tie the irrational or compelled behavior to the properties of Hotema's brain. Punishing Solomon Hotema reflects a conviction that it was his voluntary-rational self, and not simply an irrational or involuntary action (in this ruling an action linked to his brain), that is responsible for this action.

To be clear, let me stress that this was not the only way that the court could have gone in interpreting the issue of insanity. In other settings the same court stressed the idea that what makes an act insane is the mental (and here I specifically use mental not neural) state of the person at the time of the criminal act. When that occurs, a behavioral criterion is used to identify behavior that lies beyond legal sanction in a very traditional way. What makes this case so interesting are two things. First, the court extended a tradition hundreds of years old when it argued that behavior should be categorized into what are basically voluntary-rational and exculpatory divisions. Second, it did something fairly novel in trying to develop a biological marker for an exculpatory class of behavior.

Beyond Descartes and Hotema: Metaphysical Issues

To understand how these philosophical notions of voluntary and involuntary behavior influence the thoughts of neurobiologists *today*, we have to look beyond Descartes and the Hotema decision at two critical advances in Western scientific thought. The first of these advances is the rise of materialism during the last century and a growing conviction among both scholars and much of the lay public that all phenomena, even all of human behavior, are the exclusive product of purely material events. The second is a recent (and perhaps unexpected) challenge to determinism. This challenge reflects a growing conviction among some scientists that although all of the events that we can observe are the product of the material world, not all of these events are necessarily predictable in principle. Some events reflect irreducible randomness in the physical world.

One of the central, if not *the* central, products of the Enlightenment was the philosophical stance of materialism. The philosophers of the Enlightenment argued against the notion that magical powers (like those implicitly evoked by traditional notions of Free Will: an unmeasurable force that cannot be studied with physical methods but which shapes the universe around us) played a causal role in the physical world. Instead they sought to explain, at a material level, everything about the world. By the late 20th century, this notion had been extended even to the study of human behavior. It is now commonplace to state that the mind is produced entirely by the brain. Essentially all Western

scientists now accept the notion that materialism extends unequivocally to the human mind.

If we accept that even human behavior is a material product of the brain, then it seems likely that we have to accept the notion that all of human behavior is deterministic in character. And, if we accept that this is true, then humans are no more causal agents than billiard balls interacting on a pool table. Indeed, this extreme stance argues that humans cannot be seen by biologists as causal agents in any meaningful way, irrespective of whether jurists choose to retain a social consensus that labels some behaviors voluntary or rational. Many modern thinkers, however, find this notion implicitly distasteful and are troubled by the fact that much of human behavior does not seem deterministic. On these grounds, many scholars (including many biologists) dismiss materialist notions that all of behavior obeys simple physical laws. This deterministic conclusion, however, is not necessarily true for reasons that are not always obvious, and I want to take a few paragraphs to explain why. Understanding this last issue is important before we try to understand the current tension between consequentialist and ethical approaches to human behavior, because we have to understand why materialism and determinism are not identical before we try to understand the sources of human behavior.

Prior to the 20th century, it was assumed that all physical objects would obey the laws of physics which were then believed to be fully deterministic. In the early 20th century, however, quantum theory challenged that notion. What quantum theory demonstrated was that under some conditions, physical events are fundamentally unpredictable because they are random[5]. Let me be very clear about this though. This does not mean that some events in the physical world are unpredictable because we do not yet understand them, or because they show a pattern of behavior that is so complicated that they simply appear random to human observers (a mathematical property called "complex nonlinear dynamics" or more popularly "chaos theory"). This means that they are unpredictable in a fundamental way, and that this unpredictability follows clear physical laws. There is nothing magical here, just a recognition that fundamentally random processes (e.g., the Brownian motion of atoms studied by Einstein) are one of the processes within the physical domain. The point is that these processes are both fully lawful (in the physical sense) and unpredictable. For a physical scientist there is no conflict or magical thinking implicit or implied by these notions.

[5] In this regard, quantum theory differed in a philosophical way from the work of earlier scientists such as Quetelet, who studied human behavior at the statistical level, or Maxwell and Boltzmann, who studied the statistical behavior of small particles. Although their work rested on a strongly probabalistic foundation, it did not challenge the scientific assumption that physical events were deterministic. In fact, Quetelet was often attacked for suggesting that the statistical regularities in human behavior implied some kind of determinism—a point from which he often tried to distance himself.

So what does the existence of fundamentally random processes in the physical world mean for a philosophical understanding of the causal role of human agents in the physical world? First, it means that an agent could, at least in principle, be unpredictable while being fully material. This distinction is important because it removes a central barrier (for many people) to accepting a material stance with regard to human action. Human action does not appear deterministic, and there is no material reason why it should be deterministic. Second, it means that the events which follow from the actions of a stochastic actor cannot be predicted completely from the state of the environment or from the state of actor. Unpredictable events are set in motion, *caused,* by the stochastic actions of the actor. It does not mean that any special property that lies outside the physical world, such as the magical force of Free Will, is required. The critical points are:

1. The apparent contradiction between materialism and the notion that humans are fundamentally unpredictable actors is not a contradiction at all. It would only be a contradiction if it were the case that all of human action is deterministic.
2. If we allow humans to incorporate stochasticity (true randomness) into their behavior, then we recover the notion that humans can cause unpredictable events *de novo*, but *not* the notion that they can exercise Free Will in any meaningful way, a kind of *will-free agency* (or agency without Free Will).
3. Only if we place the source of human action outside the material domain, beyond the laws of both determinate and indeterminate physics, can we recover *free-willed Agency* for human actors.

The problem, of course, is that the last of these possibilities contradicts the philosophical stance that Western science has taken as axiomatically true for the last century or two. (It is also undeniable that the past two centuries, i.e., the period since the adoption of that axiom, have been very productive for the physical and natural sciences.)

Beyond Descartes and Hotema: Empirical Issues

The Empirical Search for a Voluntary-rational Boundary

The discussion above reflects a view held by many neurobiologists; namely, that humans do not have Free Will in any deep sense. Still, even in the absence of such a belief, we may well be able to divide behavior into the socially defined categories of voluntary-rational and involuntary at the neurobiological level. To see how these philosophical insights have influenced the empirical data we use to parse the natural categories of neural function, consider a recent experiment on monkey decision making which was conducted under two

sets of conditions: one yielded behavior that could be socially classified as involuntary whereas the second yielded behavior that could be classified as voluntary and rational (Glimcher 2005; Dorris and Glimcher 2005 – **include in ref lis**). About four years ago Michael Dorris and I trained both humans and monkeys to perform two tasks. The first was very simple and designed to elicit an involuntary behavior. A central spot of light appeared on a computer monitor in front of the subject. The subject's task was to fixate that light. After a delay, two other lights illuminated (one green and one red) on either side of the fixation spot. After a pause, the central spot then turned red or green and the subject was rewarded for simply looking at the color-matched target as quickly as possible. To make this orienting eye movement as reflexive as possible, we overtrained our subjects on this task. They performed this response literally tens of thousands of times. They did this task until it was as automatic as possible. Then, we traced some of the neural pathways active during this behavior in the monkeys and found that much of the nervous system governing eye movements behaved deterministically under these conditions and accurately foretold the deterministic actions of the monkeys. In particular, we focused on a group of cells in the posterior parietal cortex that behaved deterministically and could, at least in principle, account for the behavior of the monkeys during this automatic and presumably involuntary task.

Next, we trained the same subjects to play a strategic game developed by economists known as "Work or Shirk" or "The Inspection Game" (Fudenberg and Tirole 1991). In this situation, two agents face each other and play a repeated game of strategy in which they have to outwit their opponents in order to maximize their winnings. Importantly, we taught both humans and monkeys to play this game. When we asked the humans if their behavior during the game should be classified as voluntary, they all responded by saying yes. When we compared the behavior of the humans and the monkeys during the game, their behavior was identical. One could not tell the behavior of an individual monkey from the behavior of an individual human. If monkeys are capable of voluntary-rational behavior, then we reasoned that this must be it.

Now the interesting part is what happened when we studied the posterior parietal cortices of these monkeys. Once again we found that the same brain area was active, and that this same single brain area continued to predict the behavior of the monkeys. Only now, these same neurons were behaving stochastically. In other words, we found no evidence for the voluntary/involuntary distinction at a neural level. Instead, we found a single neural machine that could produce (and from which we could predict) both the deterministic and stochastic behavior of our monkeys. From the point of view of these neurons, there was no distinction that we could find between voluntary and involuntary action.

While this is admittedly only one experiment (and I have only provided the very briefest description of that experiment), it does seem to suggest that the voluntary/involuntary distinction may not be a natural category at the neurobiological level. Since that time a number of other experiments in humans have

seemed to bear out this conclusions (see the many chapters in this volume for more examples).

For me, this comes as no surprise. The entire Western scientific tradition rests on the axiom that all the phenomena that we observe are the product of events governed by physical law and thus that all phenomena in the universe can, at least in principle, be the subject of scientific inquiry. While I recognize that there are some scholars who would disagree with this axiom, I take it as a starting point. Second, given that starting point, the classical metaphysical notion of Free Will—a causal process not constrained by either the deterministic or the stochastic laws of physics—is untenable at this time and can be rejected. Third, even the notion that the socially defined construct of voluntary behavior will have a clear and meaningful neurobiological substrate seems unlikely. In fact, much of the data presented in this volume seems to suggest that this is now a consensus view. While the brain involves many systems for generating actions, there seems no compelling evidence that the two-system description is of any utility in describing the brain.

Do Neuroscience and Law Collide?

What all of this suggests is that we need to be very careful how we use neurobiological evidence in addressing the question of responsibility. Neurobiological evidence, from its metaphysical stance to the available empirical data, seems to argue against the existence of a voluntary/involuntary distinction at the physiological level. Our legal systems, our (presumably evolved) sense of fairness, our willingness to place justice in the hands of government all rest on this distinction between voluntary-rational and involuntary-irrational behavior at the psycho–legal level. That inconsistency makes reductionist approaches to the law, which seek to map explanations explicitly at these two levels to one another, perilous at best.

As a society we have a consensus that children are less responsible for their acts than adults. At a scientific level, it is easy to see how we can categorize actors into children and adults. We can even use neurobiological evidence to support this categorization in a fairly clear way (although for reasons that will be come clear below, I think that even this is a slippery slope that should be avoided). As a society we have a consensus that people experiencing strong emotional states such as rage are acting less rationally than those in emotionally neutral states and so should be held less responsible for their actions than others. However, where exactly do we draw the line between the two categories of voluntary-rational and exculpatory under these emotional conditions? In the American legal system, we draw the line on this issue by asking if the stimulus that produced the emotional state would have enraged an "ordinary and reasonable person" making it difficult or impossible for him to act rationally. What the many contributions in this volume and the work of neurobiological

scholars who are experts on emotion tell us is that this distinction does not correspond to a natural category of neural function.

So what emerges from this discussion, at least for me, is a profound skepticism about the notion that behavior can be meaningfully divided into two useful legal categories by any of the materialistic methods we encounter typically. I am arguing that the common belief that we can divide behavior into voluntary-rational and involuntary-irrational categories on neurobiological grounds and can conclude that involuntary-irrational behaviors lie beyond legal sanction is an artifact of how Descartes chose to engage the issues of Free Will and ethical responsibility in a material world. Conscious or nonconscious (as it is typically used), voluntary or involuntary, mind or brain (the worst of the three when used in this way) are all notions that I believe were originally rooted in ideas about Free Will as a contra-causal mystical force.

Are these ruminations simply the puzzled thoughts of a neurobiologist, or are any of them directly relevant to contemporary legal issues? To answer that question let me turn to studies of serotonin (a neural chemical also called 5-hydroxytriptamine or 5-HT), depression (a psychologically defined state), and violence. This is relevant because a number of legal cases involving violent acts have begun to involve measurements of brain serotonin, and I believe that these cases involve the critical error of trying to map a neurobiological phenomenon onto the voluntary/involuntary distinction—a mapping which I have argued is probably impossible, and certainly premature.

The data we have that motivates this legal use goes like this: We begin with a psychological-level definition of clinical depression and note that individuals meeting these psychological criteria are more prone to violence than the average human. Second, we treat most forms of depression today with drugs that increase brain levels of serotonin. Drugs of this type include the widely known Eli Lilly drug Prozac. Increasing brain levels of serotonin decreases the risk of violence and controls the psychological state of depression in many individuals[6]. Thus we have a clear set of isomorphic relationships at the neurobiological, psychological, and behavioral levels. Given these facts, the goal of some criminal defense lawyers has been to argue, from neurobiological measurements, toward the definition of the defendant's action as involuntary or irrational. This means that the defense argues in violent crimes, typically homicides, that either (a) low brain levels of serotonin are evidence that the defendant could not commit a willfully criminal (or intentional) act because of a diminished capacity for rationality or (b) punishment would be unjust because the low serotonin levels indicate a diminished voluntary component to the crime[7].

---

6 Indeed, we even know that decreases in brain serotonin levels increase the risk of both depression and violent acts.

7 I think that these two arguments are different for reasons that I hope will be clear in the following pages.

The first of these issues, the issue of criminal intent and its relation to the existence of a predisposing brain state, arises principally in defenses during second-degree murder trials. One of the requirements in such a trial is for the prosecution to show that the defendant formed a clear criminal intent that motivated his act. That it was his intention—and here I use a psychological word strongly but not uniquely associated with the idea of the conscious self—to commit homicide. What we find is that recently a number of defendants have argued that the existence of their low brain serotonin levels means that they could not form a true criminal intent, essentially because their behavior can be related to a measurable state of their brain (for an excellent review of this literature and its often circular reasoning, see Farahany and Coleman 2006). In my reading of these cases, the defense in question basically seeks (a) to tie the behavior to the involuntary-irrational class *on the grounds that a brain chemical is involved* and (b) to place the involuntary-irrational behavior (involuntary because it involved a brain chemical) beyond the bounds of legal sanction on ethical grounds. For me this is essentially the Hotema case—antique notions of the relationship between brain and psychology (and Free Will) in a modern legal defense. Although I am sure these revisitations of the Hotema case deserve a more complete treatment, I am going to dismiss these classes of defense as silly now that we know that "caused by the brain" cannot possibly mean exactly the same thing as involuntary (or irrational).

The second of these issues, the more challenging one, involves the use of brain levels of serotonin to mitigate punishment in a more subtle way. In Hill vs. Ozmint, an important case carefully reviewed in Farahany and Coleman (2006), David Clayton Hill was convicted of shooting a South Carolina police officer at a car wash. His lawyers made several related arguments during the sentencing phase of his trial, described here in the record of his appeal:

> In his IAC claim, Hill maintains that his defense lawyers were ineffective in calling Dr. Edward [**39] Burt to testify during the trial's sentencing phase. At sentencing, Hill's lawyers sought to show that Hill suffered from a genetic condition that caused neurochemical imbalances in his brain. Specifically, they contended that Hill suffered from a genetically based serotonin [*202] deficiency, which resulted in aggressive impulses. After his arrest and incarceration, Hill had been prescribed medication that they believed had successfully curbed these impulses. Thus, according to Hill's lawyers, the death penalty was not warranted because Hill's aggressive behavior was genetic (i.e., beyond his control) and treatable. To this end, Hill's lawyers presented the testimony of Dr. Emil Coccaro, who explained the role of serotonin in brain chemistry, as well as how genetics affects serotonin levels. Next, the defense called Dr. Bernard Albiniak, a forensic psychologist, who had performed a series of spinal taps on Hill to monitor his serotonin levels. Dr. Albiniak opined that Hill suffered from a chronic serotonin deficiency.
>
> Finally, the defense called Hill's psychiatrist, Dr. Edward Burt. Dr. Burt was expected to testify that he had prescribed Prozac to treat Hill's serotonin

> deficiency, and [**40] that Hill had responded favorably to the medication. Dr. Burt's testimony sought to establish that Hill's serotonin deficiency caused his aggressive behavior, and that a long history of violence and suicide in his family indicated that his aggressive impulses resulted from a genetic condition. Dr. Burt, however, apparently suffered a breakdown while on the witness stand. Thus, while testifying during the trial's sentencing phase, Dr. Burt had difficulty responding to questions, particularly on cross-examination.

While these claims were ultimately rejected, the legal argument raises interesting questions. If the defense had been able to establish that (a) Hill's behavior was causally related to his serotonergic brain chemistry and that (b) this brain state (the level of serotonin) was unusual, would this mitigate his punishment[8]?

My own feeling is that this is a difficult question even for a consequentialist and that it has to be approached carefully. In thinking about it, first we have to be clear that saying Hill had lower than average serotonin levels and that he was more violent than average may be saying the same thing at two different levels of reduction. (At least this is the argument that his lawyers were trying to make.) If this is true, can the argument that he had low serotonin levels (equivalent to saying that he was a violent kind of person—remember that his lawyers are arguing that these are the same thing) be used to mitigate the punishment? To answer that question I think that we have to get a bit more quantitative both in how we define "different" and in how we approach issues of punishment efficacy. Consider the graphs in Figure 16.1. Let me stress that *both are imaginary*; I use them only to illustrate a point. The first plots brain serotonin levels in a population of individuals. Note that serotonin levels can be high or low with a Gaussian distribution centered around a mean level, which I have arbitrarily called 100. Beneath that graph I have included a hypothetical plot of the likelihood that a person at any given serotonin level will commit a violent act. These (*imaginary*) numbers report how much of an increase in the chance of violent behavior is associated with any given level of whole-brain serotonin. People to the left of the graph are, on average, more violent. People on the right are, on average, less violent. How then should we use this data?

One approach, and the one for which Hill's lawyers in effect argued, is that we should set some lower bound on this graph and not punish people below that bound. Where, however, should we set this lower bound? If we believe that there are two categories of violent behavior and that these two categories map to serotonin levels[9], then where shall we find the boundaries between these two categories? At two standard deviations below the mean? At

---

8 Let me also put aside, for the purposes of this discussion, the consequentialist notion that Hill should be put away to protect society specifically because his behavior is refractory to punishment. Of course, this is almost certainly a point juries and judges consider, but not a point made in this case.

9 Something that I have argued is unlikely, but which is in the end an empirical question.

3? At 5? Where to set that lower limit seems problematic because there does not seem to be anything qualitatively different about a specific group of these people. There do not seem to be two *categories* here in any meaningful physical sense. For reasons like these, it seems almost impossible to argue up from the raw neurobiological data to any conclusion about the natural categories of social responsibility.

What we have here is a typical physically continuous variable and that suggests, if anything, that the deterrence effect of punishment may influence the behavior of all of these people in different ways. Imagine we knew (and again I recognize that this is not the case) that different degrees of punishment reduced the likelihood that any given individual would commit a violent act. Of course people on the right of the distribution shown here do not need much incentive to act nonviolently while people on the left do. A rational choice theorist *working from this neurobiological data to a theory of law* might argue that the likelihoods of violence should be matched by a strength-of-punishment function. Indeed, working from the neural data, the logical conclusion would be that punishment should be a continuous function of the convicted criminal's serotonin level. If convicted with a high serotonin level, Hill should get a lighter sentence; that is, his sentence should be inversely proportional to his serotonin level. I think that for almost everyone reading this article, the idea that if Hill was a nonviolent (or equivalently "high serotonin") individual then his crime should go largely unpunished is completely unacceptable.

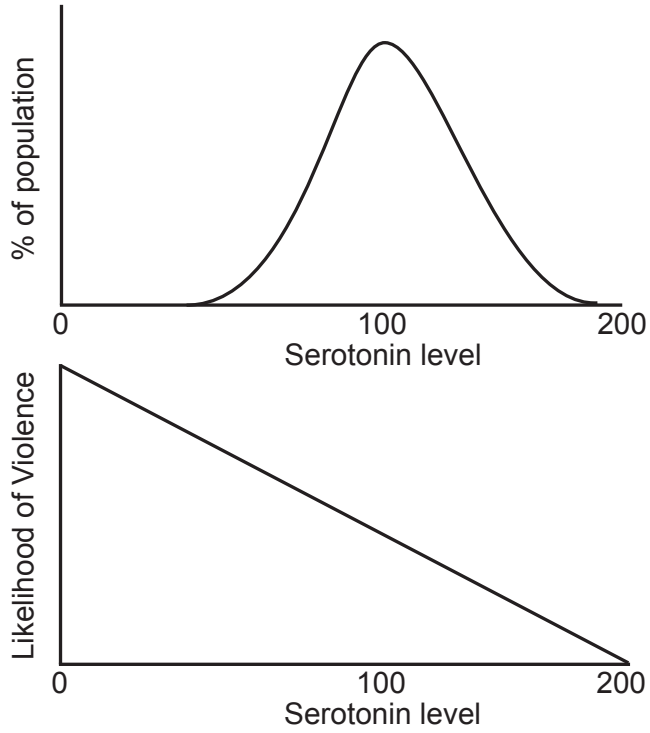

Figure 16.1 Imaginary graphs of (a) brain serotonin levels in a population of individuals and (b) a hypothetical plot of the likelihood that a person at any given serotonin level will commit a violent act.

Let me finish, though, by saying that it is entirely possible (at least in principle) to reason in the other direction. If we found (though I think this highly unlikely) that killers we define on social grounds as not responsible for their actions had particular serotonin levels, then we could begin to use those serotonin levels (be they high or low) to identify these socially categorized individuals. Whether this works is an empirical, rather than a philosophical issue. My own suspicion is that this approach will also fail. Social categories of this type will not yield robustly to neural measurements because the social categories we have are too different from the physical structure of the brain. However, this is a hypothesis that will continue to be tested in the years to come.

Reductionism, Law, and Neuroscience Together?

Over the course of the last decade or two, there have been tremendous steps made towards a reductive synthesis that relates neurobiology, psychology and economics[10]. Each of these disciplines can be seen as a description of the causes of human behavior at differing levels of reduction. Neurobiology seeks to describe the physical processes that generate behavior. Psychology seeks to describe the mental processes that generate this same behavior, and, at a more global level of analysis, economics seeks highly parsimonious models that predict behavior from initial conditions without explicit regard to underlying mechanism.

Beginning in earnest with the introduction of modern brain scanners less than two decades ago, there have been strong linkages created between the neurobiological and psychological levels of analysis. The premise that guides the formation of these links is that psychological theories compatible with the underlying neural architecture are more likely to be robust and extensible than those psychological theories that are incompatible with the underlying physiology. At about the same time economics and psychology began to interact in a similar way, and more recently, a similar set of linkages has begun to emerge between neuroscience and economics. The result of this interdisciplinary activity has been a growing alignment of the models and theories that guide these three disciplines.

One example of that growing alignment seems to be an emerging consensus that the boundary between voluntary-rational and involuntary-irrational behavior (as either a lawyer or a psychologist might define them) is not nearly as clear at the economic, psychological, or neurobiological level as had been previously supposed. This is a fundamental theme that has emerged throughout this Ernst Strüngmann Forum. From the four working groups, the first three

[10] Note, however, that there is no evidence that economic and psychological levels are about to be reduced completely to neurobiology. Indeed, the history of chemistry and physics suggests that strong reductive relationships will continue to emerge while full reduction will probably remain elusive.

challenge repeatedly the distinction between conscious/voluntary and nonconscious/involuntary. The corresponding chapters convey an emerging empirical consensus at the neurobiological, psychological, and economic levels that one cannot successfully describe behavior as the product of two independent systems (or groups of systems).

The fourth working group (see Lubell et al., this volume) and its corresponding chapters reveal a different story. At the level of institutions, the notion of a voluntary-rational/involuntary-irrational distinction is more than just alive and well. It seems clear that this distinction and the notion of justice to which it is closely related are necessary components of our institutions. Institutional actors, like courts, serve as proxies for those they govern. People accept the rule of law and grant governments the power to imprison when the actions of those governments align with their own goals.

Summary

We have overwhelming evidence that the Western legal system reflects a widely held conviction among its citizens both that the voluntary/involuntary distinction is meaningful and that punishment for crimes must reflect this distinction. My own opinion is that this distinction has its roots in the classical division of behavior into reflex and voluntary/free-willed categories, a division that was stated so clearly by Descartes and which now permeates so many aspects of Western culture. Regardless of the cultural source of this division of behavior into two categories, it is clear that people have a "taste" for fairness and this taste is rooted irrefutably in distinctions between voluntary and involuntary or irrational behavior. Criminal law, and perhaps institutions in general, implement this division with a complex patchwork of tools. Some behaviors are labeled involuntary, others are labeled compelled, some are described as rational, others as the product of emotional states. What cuts across all of these categories, for me, is that we observe two super-categories: actions for which an actor will be held responsible and actions for which he will not be held responsible. There may be gray areas between these two super-categories, but it is these super-categorical boundaries that do all the work of deciding who will be punished and who will not. Legal systems are, in essence, evolved social systems that both function effectively and rest on this distinction.

A century ago, neurobiological and psychological analyses of behavior reflected a widely held conviction among scholars that the physical and mental roots of action supported the empirical division of behavior into two super-categories. In neurobiology these categories were called reflex and cognition. In psychology these categories have had many names, ranging from involuntary and voluntary to conscious and unconscious. In fairness, neoclassical economics never supported such a distinction—behavior was behavior—although more recently even economics has begun to search for such a distinction. Over

the last few decades, however, neurobiologists and psychologists (and now increasingly economists) have become more convinced that this dipartite approach to behavior is critically flawed. Most have become convinced that neither the physical brain nor any really workable psychological descriptions of mind can be made to rest on this dualist approach. If those conclusions are correct, then we simply have to make one of two choices. Either we abandon the hope for a reductive synthesis between institutional design and the neurobiological and psychological roots of behavior, or we abandon the super-categorical boundaries that underlie much of our institutional design.

Over the next several decades it seems unavoidable that descriptions of human behavior at the neural, psychological, and economic levels will become increasingly compatible. It also seems clear that those descriptions will not include a clear voluntary/involuntary distinction at the level of causal agency. Still, one has to be clear that this is a line of reasoning that, at least today, cannot propagate upwards to the institutional level. It is my conviction that neurobiology cannot guide law, because these two disciplines rest on differing, and to my mind irreconcilable, foundations. Law is based on social, not scientific, principles, and scientists must make their peace with that fact.

The implications for such a conclusion in law are significant. If these two systems for describing behavior rest on irreconcilable premises, then we simply cannot use neurobiological data to shape deep structural features of institutions. We can, for example, continue to search for neural measurements to identify culpable mental states (despite my personal skepticism about this endeavor) even if those measurements are not reducible to any theory of the brain. We can continue to use psychological theories to inform us about consequentialist issues in the design of punishments. What we cannot do, however, is to argue upwards to notions of responsibility and culpability from neurobiological data. We must continue to be cautious in our aspirations. Brains are exceedingly complicated devices, and it is not at all clear what constitute the natural categories, or even the system-level descriptions of these devices. Imposing social constructs on our interpretations of these categories is not guaranteed to yield legal clarity. Instead it may only yield injustice. Discussions like those in the Hotema case make that clear.

Acknowledgments

The author is grateful to Stephen Morse, Christoph Engel, and participants at this Ernst Strüngmann Forum for their comments and assistance.

References

Aristotle. 1987. *De Anima*. Oxford: Penguin Classics

Blount, S. 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. *Behav. Hum. Dec. Proc.* **63**:131–144.

Brosnan, S. F., and F. B. M. deWaal. 2003. Monkeys reject unequal pay. *Nature* **425**:297–299.

Bruce, V. 1982. Changing faces: Visual and non-visual coding processes in face recognition. *Brit. J. Psychol.* **73**:105–117.

Calvin, J. 1863/1964. *Ioannis Calvini Opera Quae Supersunt Omnia*, ed. J. H. Guilielmus, E. C. Baum, and E. Reuss. Braunschweig.

Dehaene, S., J. P. Changeux, L. Naccache, J. Sackur, and C. Sergent. 2006. Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends Cogn. Sci.* **10**:204–211.

Descartes, R. 1649/1989. *Passions De L'Ame* [The Passions of the Soul], trans. S. Voss. Indianapolis: Hackett Publ. Co.

Descartes, R. 1664/1972. *L'Homme*, trans. T. S. Hall. Harvard Univ. Press.

Dorris and Glimcher. 2005 – INCLUDE ref. info.

Farahany, N. A., and J. E. Coleman. 2006. Genetics and responsibility: To know the criminal from the crime. *Law Contemp. Probs.* **69**:115.

Fudenberg, D., and J. Tirole. 1991. Game Theory. Cambridge, MA: MIT Press.

Garland, B., and P. W. Glimcher. 2006. Cognitive neuroscience and the law. *Curr. Opin. Neurobiol.* **16**:130–134.

Gauthier, I., M. J. Tarr, A. W. Anderson, P. Skudlarski, and J. C. Gore. 1999. Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neurosci.* **2**:568–573.

Glimcher, P. W. 2005. Indeterminacy in brain and behavior. *Ann. Rev. Psychol.* **56**:25–56.

Gobbini, M.I., and J. V. Haxby. 2007. Neural systems for recognition of familiar faces. *Neuropsychologia* **45**:32–41.

Guth, W., R. Schmittberger, and B. Schwarz. 1982. An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Org.* **3**:367–388.

Harvey, W. C. 1628/1995. *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus* [An Anatomical Disquisition on the Motion of the Heart and Blood in Animals]. New York: Dover Publ.

Haxby, J. V., M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten et al. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**:2425–2430.

Jones, O. 2001. Time-shifted rationality and the law of law's leverage: Behavioral economics meets behavioral biology. *Northwestern Univ. Law Rev.* **95**:1141–1205.

Kanwisher, N., J. McDermott, and M. Chun. 1997. The fusiform face area: A module in human extrastriate cortex specialized for the perception of faces. *J. Neurosci.* **17**:4302–4311.

Kanwisher, N., and G. Yovel. 2006. The fusiform face area: A cortical region specialized for the perception of faces.. *Phil. Trans. Roy. Soc. Lond. B.* **361**:2109–2128.

McCabe, K., D. Houser, L. Ryan, V. Smith, and T. Trouard. 2001. A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl. Acad. Sci.* **98**:11,832–11,835.

Morse, S. 2006. The non-problem of free will in forensic psychiatry and psychology. *Behav. Sci. Law* **24**:1–17.

Vesalius, A. 1543/1998–1999. *De Humani Corpus Fabrica* [On the Fabric of the Human Body], trans. W. F. Richardson and J. B. Carman. Novato, CA: Norman Publ.

Wolpe, P. R., K. R. Foster, and D. D. Langleben. 2005. Emerging neurotechnologies for lie detection: Promises and perils. *Am. J. Bioethics* 5:39–49.