THE ROYAL SOCIETY

Published online 26 November 2004

For the law, neuroscience changes nothing and everything

Joshua Greene* **and Jonathan Cohen**

Department of Psychology, Center for the Study of Brain, Mind, and Behavior, Princeton University, Princeton, NJ 08544, USA

The rapidly growing field of cognitive neuroscience holds the promise of explaining the operations of the mind in terms of the physical operations of the brain. Some suggest that our emerging understanding of the physical causes of human (mis)behaviour will have a transformative effect on the law. Others argue that new neuroscience will provide only new details and that existing legal doctrine can accommodate whatever new information neuroscience will provide. We argue that neuroscience will probably have a transformative effect on the law, despite the fact that existing legal doctrine can, in principle, accommodate whatever neuroscience will tell us. New neuroscience will change the law, not by undermining its current assumptions, but by transforming people's moral intuitions about free will and responsibility. This change in moral outlook will result not from the discovery of crucial new facts or clever new arguments, but from a new appreciation of old arguments, bolstered by vivid new illustrations provided by cognitive neuroscience. We foresee, and recommend, a shift away from punishment aimed at retribution in favour of a more progressive, consequentialist approach to the criminal law.

Keywords: law; brain; morality; free will; punishment; retributivism

1. INTRODUCTION

The law takes a long-standing interest in the mind. In most criminal cases, a successful conviction requires the prosecution to establish not only that the defendant engaged in proscribed behaviour, but also that the misdeed in question was the product of *mens rea*, a 'guilty mind'. Narrowly interpreted, *mens rea* refers to the intention to commit a criminal act, but the term has a looser interpretation by which it refers to all mental states consistent with moral and/or legal blame. (A killing motivated by insane delusional beliefs may meet the requirements for *mens rea* in the first sense, but not the second.) (Goldstein *et al.* 2003) Thus, for centuries, many legal issues have turned on the question: 'what was he thinking?'.

To answer this question, the law has often turned to science. Today, the newest kid on this particular scientific block is cognitive neuroscience, the study of the mind through the brain, which has gained prominence in part as a result of the advent of functional neuroimaging as a widely used tool for psychological research. Given the law's aforementioned concern for mental states, along with its preference for 'hard' evidence, it is no surprise that interest in the potential legal implications of cognitive neuroscience abounds. But does our emerging understanding of the mind as brain really have any deep implications for the law? This theme issue is a testament to the thought that it might. Some have argued, however, that new neuroscience contributes nothing more than new details and that existing legal principles can handle anything that neuroscience will throw our way in the foreseeable future (Morse 2004).

In our view, both of these positions are, in their respective ways, correct. Existing legal principles make virtually no assumptions about the neural bases of criminal behaviour, and as a result they can comfortably assimilate new neuroscience without much in the way of conceptual upheaval: new details, new sources of evidence, but nothing for which the law is fundamentally unprepared. We maintain, however, that our operative legal principles exist because they more or less adequately capture an intuitive sense of justice. In our view, neuroscience will challenge and ultimately reshape our intuitive sense(s) of justice. New neuroscience will affect the way we view the law, not by furnishing us with new ideas or arguments about the nature of human action, but by breathing new life into old ones. Cognitive neuroscience, by identifying the specific mechanisms responsible for behaviour, will vividly illustrate what until now could only be appreciated through esoteric theorizing: that there is something fishy about our ordinary conceptions of human action and responsibility, and that, as a result, the legal principles we have devised to reflect these conceptions may be flawed.

Our argument runs as follows: first, we draw a familiar distinction between the consequentialist justification for state punishment, according to which punishment is merely an instrument for promoting future social welfare, and the retributivist justification for punishment, according to which the principal aim of punishment is to give people what they deserve based on their past actions. We observe that the common-sense approach to moral and legal responsibility has consequentialist elements, but is largely retributivist. Unlike the consequentialist justification for

* Author for correspondence (jdgreene@princeton.edu).

One contribution of 16 to a Theme Issue 'Law and the brain'.

doi:10.1098/rstb.2004.1546

punishment, the retributivist justification relies, either explicitly or implicitly, on a demanding—and some say overly demanding—conception of free will. We therefore consider the standard responses to the philosophical problem of free will (Watson 1982). 'Libertarians' (no relation to the political philosophy) and 'hard determinists' agree on 'incompatibilism', the thesis that free will and determinism are incompatible, but they disagree about whether determinism is true, or near enough true to preclude free will. Libertarians believe that we have free will because determinism is false, and hard determinists believe that we lack free will because determinism is (approximately) true. 'Compatibilists', in contrast to libertarians and hard determinists, argue that free will and determinism are perfectly compatible.

We argue that current legal doctrine, although officially compatibilist, is ultimately grounded in intuitions that are incompatibilist and, more specifically, libertarian. In other words, the law *says* that it presupposes nothing more than a metaphysically modest notion of free will that is perfectly compatible with determinism. However, we argue that the law's intuitive support is ultimately grounded in a metaphysically overambitious, libertarian notion of free will that is threatened by determinism and, more pointedly, by forthcoming cognitive neuroscience. At present, the gap between what the law officially cares about and what people really care about is only revealed occasionally when vivid scientific information about the causes of criminal behaviour leads people to doubt certain individuals' capacity for moral and legal responsibility, despite the fact that this information is irrelevant according to the law's stated principles. We argue that new neuroscience will continue to highlight and widen this gap. That is, new neuroscience will undermine people's common sense, libertarian conception of free will and the retributivist thinking that depends on it, both of which have heretofore been shielded by the inaccessibility of sophisticated thinking about the mind and its neural basis.

The net effect of this influx of scientific information will be a rejection of free will as it is ordinarily conceived, with important ramifications for the law. As noted above, our criminal justice system is largely retributivist. We argue that retributivism, despite its unstable marriage to compatibilist philosophy in the letter of the law, ultimately depends on an intuitive, libertarian notion of free will that is undermined by science. Therefore, with the rejection of common-sense conceptions of free will comes the rejection of retributivism and an ensuing shift towards a consequentialist approach to punishment, i.e. one aimed at promoting future welfare rather than meting out just deserts. Because consequentialist approaches to punishment remain viable in the absence of common-sense free will, we need not give up on moral and legal responsibility. We argue further that the philosophical problem of free will arises out of a conflict between two cognitive subsystems that speak different 'languages': the 'folk psychology' system and the 'folk physics' system. Because we are inherently of two minds when it comes to the problem of free will, this problem will never find an intuitively satisfying solution. We can, however, recognize that free will, as conceptualized by the folk psychology system, is an illusion and structure our society accordingly by rejecting retributivist legal principles that derive their intuitive force from this illusion.

2. TWO THEORIES OF PUNISHMENT: CONSEQUENTIALISM AND RETRIBUTIVISIM

There are two standard justifications for legal punishment (Lacey 1988). According to the forward-looking, consequentialist theory, which emerges from the classical utilitarian tradition (Bentham 1982), punishment is justified by its future beneficial effects. Chief among them are the prevention of future crime through the deterrent effect of the law and the containment of dangerous individuals. Few would deny that the deterrence of future crime and the protection of the public are legitimate justifications for punishment. The controversy surrounding consequentialist theories concerns their serviceability as *complete* normative theories of punishment. Most theorists find them inadequate in this regard (e.g. Hart 1968), and many argue that consequentialism fundamentally mischaracterizes the primary justification for punishment, which, these critics argue, is retribution (Kant 2002). As a result, they claim, consequentialist theories justify intuitively unfair forms of punishment, if not in practice then in principle. One problem is that of Draconian penalties. It is possible, for example, that imposing the death penalty for parking violations would maximize aggregate welfare by reducing parking violations to near zero. But, retributivists claim, whether or not this is a good idea does not depend on the balance of costs and benefits. It is simply wrong to kill someone for double parking. A related problem is that of punishing the innocent. It is possible that, under certain circumstances, falsely convicting an innocent person would have a salutary deterrent effect, enough to justify that person's suffering, etc. Critics also note that, so far as deterrence is concerned, it is the *threat* of punishment that is justified and not the punishment itself. Thus, consequentialism might justify letting murderers and rapists off the hook so long as their punishment could be convincingly faked.

The standard consequentialist response to these charges is that such concerns have no place in the real world. They say, for example, that the idea of imposing the death penalty for parking violations to make society an overall happier place is absurd. People everywhere would live in mortal fear of bureaucratic errors, and so on. Likewise, a legal system that deliberately convicted innocent people and/or secretly refrained from punishing guilty ones would require a kind of systematic deception that would lead inevitably to corruption and that could never survive in a free society. At this point critics retort that consequentialist theories, at best, get the right answers for the wrong reasons. It is wrong to punish innocent people, etc. because it is fundamentally unfair, not because it leads to bad consequences in practice. Such critics are certainly correct to point out that consequentialist theories fail to capture something central to common-sense intuitions about legitimate punishment.

The backward-looking, retributivist account does a better job of capturing these intuitions. Its fundamental principle is simple: in the absence of mitigating circumstances, people who engage in criminal behaviour *deserve* to be punished, and that is why we punish them. Some would

explicate this theory in terms of criminals' forfeiting rights, others in terms of the rights of the victimized, whereas others would appeal to the violation of a hypothetical social contract, and so on. Retributivist theories come in many flavours, but these distinctions need not concern us here. What is important for our purposes is that retributivism captures the intuitive idea that we legitimately punish to give people what they deserve based on their past actions—in proportion to their 'internal wickedness', to use Kant's (2002) phrase—and not, primarily, to promote social welfare in the future.

The retributivist perspective is widespread, both in the explicit views of legal theorists and implicitly in common sense. There are two primary motivations for questioning retributivist theory. The first, which will not concern us here, comes from a prior commitment to a broader consequentialist moral theory. The second comes from scepticism regarding the notion of desert, grounded in a broader scepticism about the possibility of free will in a deterministic or mechanistic world.

3. FREE WILL AND RETRIBUTIVISM

The problem of free will is old and has many formulations (Watson 1982). Here is one, drawing on a more detailed and exacting formulation by Peter Van Inwagen (1982): determinism is true if the world is such that its current state is completely determined by (i) the laws of physics and (ii) past states of the world. Intuitively, the idea is that a deterministic universe starts however it starts and then ticks along like clockwork from there. Given a set of prior conditions in the universe and a set of physical laws that completely govern the way the universe evolves, there is only one way that things can actually proceed.

Free will, it is often said, requires the ability do otherwise (an assumption that has been questioned; Frankfurt 1966). One cannot say, for example, that I have freely chosen soup over salad if forces beyond my control are sufficient to necessitate my choosing soup. But, the determinist argues, this is precisely what forces beyond your control do—always. You have no say whatsoever in the state of the universe before your birth; nor do you have any say about the laws of physics. However, if determinism is true, these two things together are sufficient to determine your choice of soup over salad. Thus, some say, if determinism is true, your sense of yourself and others as having free will is an illusion.

There are three standard responses to the problem of free will. The first, known as 'hard determinism', accepts the incompatibility of free will and determinism ('incompatibilism'), and asserts determinism, thus rejecting free will. The second response is libertarianism (again, no relation to the political philosophy), which accepts incompatibilism, but denies that determinism is true. This may seem like a promising approach. After all, has not modern physics shown us that the universe is *in*deterministic (Hughs 1992)? The problem here is that the sort of indeterminism afforded by modern physics is not the sort the libertarian needs or desires. If it turns out that your ordering soup is completely determined by the laws of physics, the state of the universe 10 000 years ago, *and* the outcomes of myriad subatomic coin flips, your appetizer is no more freely chosen than before. Indeed, it is *randomly* chosen, which is no help to the libertarian. What about some other kind of indeterminism? What if, somewhere deep in the brain, there are mysterious events that operate independently of the ordinary laws of physics and that are somehow tied to the will of the brain's owner? In light of the available evidence, this is highly unlikely. Say what you will about the 'hard problem' of consciousness (Shear 1999), there is not a shred of scientific evidence to support the existence of *causally effective* processes in the mind or brain that violate the laws of physics. In our opinion, any scientifically respectable discussion of free will requires the rejection of what Strawson (1962) famously called the 'panicky metaphysics' of libertarianism.[1]

Finally, we come to the dominant view among philosophers and legal theorists: compatibilism. Compatibilists concede that some notions of free will may require indefensible, panicky metaphysics, but maintain that the kinds of free will 'worth wanting', to use Dennett's (1984) phrase, are perfectly compatible with determinism. Compatibilist theories vary, but all compatibilists agree that free will is a perfectly natural, scientifically respectable phenomenon and part of the ordinary human condition. They also agree that free will can be undermined by various kinds of psychological deficit, e.g. mental illness or 'infancy'. Thus, according to this view, a freely willed action is one that is made using the right sort of psychology—rational, free of delusion, etc.

Compatibilists make some compelling arguments. After all, is it not obvious that we have free will? Could science plausibly deny the obvious fact that I am free to raise my hand *at will*? For many people, such simple observations make the reality of free will non-negotiable. But at the same time, many such people concede that determinism, or something like it, is a live possibility. And if free will is obviously real, but determinism is debatable, then the reality of free will must not hinge on the rejection of determinism. That is, free will and determinism must be compatible. Many compatibilists sceptically ask what would it mean to give up on free will. Were we to give it up, wouldn't we have to immediately reinvent it? Does not every decision involve an implicit commitment to the idea of free will? And how else would we distinguish between ordinary rational adults and other individuals, such as young children and the mentally ill, whose will—or whatever you want to call it—is clearly compromised? Free will, compatibilists argue, is here to stay, and the challenge for science is to figure out how exactly it works and not to peddle silly arguments that deny the undeniable (Dennett 2003).

The forward-looking–consequentialist approach to punishment works with all three responses to the problem of free will, including hard determinism. This is because consequentialists are not concerned with whether anyone is really innocent or guilty in some ultimate sense that might depend on people's having free will, but only with the likely effects of punishment. (Of course, one might wonder what it means for a hard determinist to justify any sort of choice. We will return to this issue in § 8.) The retributivist approach, by contrast, is plausibly regarded as requiring free will and the rejection of hard determinism. Retributivists want to know whether the defendant truly *deserves* to be punished. Assuming one can deserve to be punished only for actions that are freely willed, hard determinism implies that no one really deserves to be punished. Thus,

hard determinism combined with retributivism requires the elimination of all punishment, which does not seem reasonable. This leaves retributivists with two options: compatibilism and libertarianism. Libertarianism, for reasons given above, and despite its intuitive appeal, is scientifically suspect. At the very least, the law should not depend on it. It seems, then, that retributivism requires compatibilism. Accordingly, the standard legal account of punishment is compatibilist.

4. NEUROSCIENCE CHANGES NOTHING

The title of a recent paper by Stephen Morse (2004), 'New neuroscience, old problems', aptly summarizes many a seasoned legal thinker's response to the suggestion that brain research will revolutionize the law. The law has been dealing with issues of criminal responsibility for a long time, Morse argues that there is nothing on the neuroscientific horizon that it cannot handle.

The reason that the law is immune to such threats is that it makes no assumptions that neuroscience, or any science, is likely to challenge. The law assumes that people have a general capacity for rational choice. That is, people have beliefs and desires and are capable of producing behaviour that serves their desires in light of their beliefs. The law acknowledges that our capacity for rational choice is far from perfect (Kahneman & Tversky 2000), requiring only that the people it deems legally responsible have a *general* capacity for rational behaviour.

Thus, questions about who is or is not responsible in the eyes of the law have and will continue to turn on questions about rationality. This approach was first codified in the *M'Naghten* standard according to which a defence on the ground of insanity requires proof that the defendant laboured under 'a defect of reason, from disease of the mind' (Goldstein 1967). Not all standards developed and applied since *M'Naghten* explicitly mention the need to demonstrate the defendant's diminished rationality (e.g. the *Durham* standard; Goldstein 1967), but it is generally agreed that a legal excuse requires a demonstration that the defendant 'lacked a general capacity for rationality' (Goldstein *et al.* 2003). Thus, the argument goes, new science can help us figure out who was or was not rational at the scene of the crime, much as it has in the past, but new science will not justify any fundamental change in the law's approach to responsibility unless it shows that people in general fail to meet the law's very minimal requirements for rationality. Science shows no sign of doing this, and thus the basic precepts of legal responsibility stand firm. As for neuroscience more specifically, this discipline seems especially unlikely to undermine our faith in general minimal rationality. If any sciences have an outside chance of demonstrating that our behaviour is thoroughly irrational or arational it is the ones that study behaviour directly rather than its proximate physical causes in the brain. The law, this argument continues, does not care if people have 'free will' in any deep metaphysical sense that might be threatened by determinism. It only cares that people in general are minimally rational. So long as this appears to be the case, it can go on regarding people as free (compatibilism) and holding ordinary people responsible for their misdeeds while making exceptions for those who fail to meet the requirements of general rationality.

In light of this, one might wonder what all the fuss is about. If the law assumes nothing more than general minimal rationality, and neuroscience does nothing to undermine this assumption, then why would anyone even *think* that neuroscience poses some sort of threat to legal doctrines of criminal responsibility? It sounds like this is just a simple mistake, and that is precisely what Morse contends. He calls this mistake 'the fundamental psycholegal error' which is 'to believe that causation, especially abnormal causation, is *per se* an excusing condition' (Morse 2004, p. 180). In other words, if you think that neuroscientific information about the causes of human action, or some particular human's action, can, by itself, make for a legitimate legal excuse, you just do not understand the law. Every action is caused by brain events, and describing those events and affirming their causal efficacy is of no legal interest in and of itself. Morse continues, '[The psycholegal error] leads people to try to create a new excuse every time an allegedly valid new '' syndrome '' is discovered that is thought to play a role in behaviour. But syndromes and other causes do not have excusing force unless they sufficiently diminish rationality in the context in question' (Morse 2004, p. 180).

In our opinion, Morse and like-minded theorists are absolutely correct about the relationship between current legal doctrine and any forthcoming neuroscientific results. For the law, as written, neuroscience changes nothing. The law provides a coherent framework for the assessment of criminal responsibility that is not threatened by anything neuroscience is likely to throw at it. But, we maintain, the law nevertheless stands on shakier ground than the foregoing would suggest. The legitimacy of the law itself depends on its adequately reflecting the moral intuitions and commitments of society. If neuroscience can change those intuitions, then neuroscience can change the law.

As it happens, this is a possibility that Morse explicitly acknowledges. However, he believes that such developments would require radical new ideas that we can scarcely imagine at this time, e.g. a new solution to the mind–body problem. We disagree. The seeds of discontent are already sown in common-sense legal thought. In our opinion, the 'fundamental psycholegal error' is not so much an error as a reflection of the gap between what the law officially cares about and what people really care about. In modern criminal law, there has been a long tense marriage of convenience between compatibilist legal principles and libertarian moral intuitions. New neuroscience, we argue, will probably render this marriage unworkable.

5. WHAT REALLY MATTERS FOR RESPONSIBILITY? MATERIALIST THEORY, DUALIST INTUITIONS AND THE 'BOYS FROM BRAZIL' PROBLEM

According to the law, the central question in a case of putative diminished responsibility is whether the accused was sufficiently rational at the time of the misdeed in question. We believe, however, that this is not what most people really care about, and that for them diminished rationality is just a presumed correlate of something deeper. It seems that what many people really want to know is: was it really *him*? This question usually comes in the form of a disjunction, depending on how the excuse is constructed: was it *him*, or was it his *upbringing*? Was it *him*, or was it his *genes*?

Was it *him*, or was it his *circumstances*? Was it *him*, or was it his *brain*? But what most people do not understand, despite the fact that naturalistic philosophers and scientists have been saying it for centuries, is that there is no 'him' independent of these other things. (Or, to be a bit more accommodating to the supernaturally inclined, there is no 'him' independent of these things that shows any sign of affecting anything in the physical world, including his behaviour.)

Most people's view of the mind is implicitly *dualist* and *libertarian* and not *materialist* and *compatibilist*. Dualism, for our purposes, is the view that mind and brain are separate, interacting, entities.[2] Dualism fits naturally with libertarianism because a mind distinct from the body is precisely the sort of non-physical source of free will that libertarianism requires. Materialism, by contrast, is the view that all events, including the operations of the mind, are ultimately operations of matter that obeys the laws of physics. It is hard to imagine a belief in free will that is materialist but not compatibilist, given that ordinary matter does not seem capable of supplying the non-physical processes that libertarianism requires.

Many people, particularly those who are religious, are explicitly dualist libertarians (again, not in the political sense). However, in our estimation, even people who do or would readily endorse a thoroughly material account of human action and its causes have dualist, libertarian intuitions. This goes not only for educated people in general, but for experts in mental health and criminal behaviour. Consider, for example, the following remarks from Jonathan Pincus, an expert on criminal behaviour and the brain.

> When a composer conceives a symphony, the only way he or she can present it to the public is through an orchestra... If the performance is poor, the fault could lie with the composer's conception, or the orchestra, or both... Will is expressed by the brain. Violence can be the result of volition only, but if a brain is damaged, brain failure must be at least partly to blame.
>
> (Pincus 2001, p. 128)

To our untutored intuitions, this is a perfectly sensible analogy, but it is ultimately grounded in a kind of dualism that is scientifically untenable. It is not as if there is *you*, the composer, and then *your brain*, the orchestra. You *are* your brain, and your brain is the composer and the orchestra all rolled together. There is no little man, no 'homunculus', in the brain that is the real you behind the mass of neuronal instrumentation. Scientifically minded philosophers have been saying this *ad nauseum* (Dennett 1991), and we will not belabour the point. Moreover, we suspect that if you were to ask Dr Pincus whether he thinks there is a little conductor directing his brain's activity from within or beyond he would adamantly deny that this is the case. At the same time, though, he is comfortable comparing a brain-damaged criminal to a healthy conductor saddled with an unhealthy orchestra. This sort of doublethink is not uncommon. As we will argue in § 7, when it comes to moral responsibility in a physical world, we are all of two minds.

A recent article by Laurence Steinberg and Elizabeth Scott (Steinberg & Scott 2003), experts respectively on adolescent developmental psychology and juvenile law, illustrates the same point. They argue that adolescents do not meet the law's general requirements for rationality and that therefore they should be considered less than fully responsible for their actions and, more specifically, unsuitable candidates for the death penalty. Their main argument is sound, but they cannot resist embellishing it with a bit of superfluous neuroscience.

> Most of the developmental research on cognitive and psychosocial functioning in adolescence measures behaviors, self-perceptions, or attitudes, but mounting evidence suggests that at least some of the differences between adults and adolescents have neuropsychological and neurobiological underpinnings.
>
> (Steinberg & Scott 2003, p. 5)

Some of the differences? Unless some form of dualism is correct, *every* mental difference and *every* difference in behavioural tendency is a function of some kind of difference in the brain. But here it is implicitly suggested that things like 'behaviours, self-perceptions, or attitudes' may be grounded in something other than the brain. In summing up their case, Steinberg and Scott look towards the future.

> Especially needed are studies that link developmental changes in decision making to changes in brain structure and function... In our view, however, there is sufficient indirect suggestive evidence of age differences in capacities that are relevant to criminal blameworthiness to support the position that youths who commit crimes should be punished more leniently then their adult counterparts.
>
> (Steinberg & Scott 2003, p. 9)

This gets the order of evidence backwards. If what the law ultimately cares about is whether adolescents can behave rationally, then it is evidence concerning adolescent behaviour that is *directly* relevant. Studying the adolescent brain is a highly *indirect* way of figuring out whether adolescents in general are rational. Indeed, the only way we neuroscientists can tell if a brain structure is important for rational judgement is to see if its activity or damage is correlated with (ir)rational *behaviour*.[3]

If everyone agrees that what the law ultimately cares about is the capacity for rational behaviour, then why are Steinberg and Scott so optimistic about neuroscientific evidence that is only indirectly relevant? The reason, we suggest, is that they are appealing not to a legal argument, but to a moral intuition. So far as the law is concerned, information about the physical processes that give rise to bad behaviour is irrelevant. But to people who implicitly believe that real decision-making takes place in the mind, not in the brain, demonstrating that there is a brain basis for adolescents' misdeeds allows us to blame adolescents' brains instead of the adolescents themselves.

The fact that people are tempted to attach great moral or legal significance to neuroscientific information that, according to the letter of the law, should not matter, suggests that what the law cares about and what people care about do not necessarily coincide. To make this point in a more general way, we offer the following thought experiment, which we call '*The Boys from Brazil* problem'. It is an extension of an argument that has made the rounds in philosophical discussions of free will and responsibility (Rosen 2002).

In the film *The Boys from Brazil*, members of the Nazi old guard have regrouped in South America after the war. Their plan is to bring their beloved *führer* back to life by raising children genetically identical to Hitler (courtesy of some salvaged DNA) in environments that mimic that of Hitler's upbringing. For example, Hitler's father died while

young Adolph was still a boy, and so each Hitler clone's surrogate father is killed at just the right time, and so on, and so forth.

This is obviously a fantasy, but the idea that one could, in principle, produce a person with a particular personality and behavioural profile through tight genetic and environmental control is plausible. Let us suppose, then, that a group of scientists has managed to create an individual—call him 'Mr Puppet'—who, by design, engages in some kind of criminal behaviour: say, a murder during a drug deal gone bad. The defence calls to the stand the project's lead scientist: 'Please tell us about your relationship to Mr Puppet. . .'

> It is very simple, really. I designed him. I carefully selected every gene in his body and carefully scripted every significant event in his life so that he would become precisely what he is today. I selected his mother knowing that she would let him cry for hours and hours before picking him up. I carefully selected each of his relatives, teachers, friends, enemies, etc. and told them exactly what to say to him and how to treat him. Things generally went as planned, but not always. For example, the angry letters written to his dead father were not supposed to appear until he was fourteen, but by the end of his thirteenth year he had already written four of them. In retrospect I think this was because of a handful of substitutions I made to his eighth chromosome. At any rate, my plans for him succeeded, as they have for 95% of the people I've designed. I assure you that the accused deserves none of the credit.

What to do with Mr Puppet? Insofar as we believe this testimony, we are inclined to think that Mr Puppet cannot be held fully responsible for his crimes, if he can be held responsible for them at all. He is, perhaps, a man to be feared, and we would not want to return him to the streets. But given the fact that forces beyond his control played a dominant role in causing him to commit these crimes, it is hard to think of him as anything more than a pawn.

But what does the law say about Mr Puppet? The law asks whether or not he was rational at the time of his misdeeds, and as far as we know he was. For all we know, he is psychologically indistinguishable from the prototypical guilty criminal, and therefore fully responsible in the eyes of the law. But, intuitively, this is not fair.

Thus, it seems that the law's exclusive interest in rationality misses something intuitively important. In our opinion, rationality is just a presumed correlate of what most people really care about. What people really want to know is if the accused, as opposed to something else, is responsible for the crime, where that 'something else' could be the accused's brain, genes or environment. The question of someone's ultimate responsibility seems to turn, intuitively, on a question of internal versus external determination. Mr Puppet ought not be held responsible for his actions because forces beyond his control played a dominant role in the production of his behaviour. Of course, the scientists did not have complete control—after all, they had a 5% failure rate—but that does not seem to be enough to restore Mr Puppet's free will, at least not entirely. Yes, he is as rational as other criminals, and, yes, it was his desires and beliefs that produced his actions. But those beliefs and desires were rigged by external forces, and that is why, intuitively, he deserves our pity more than our moral condemnation.[4]

The story of Mr. Puppet raises an important question: what is the difference between Mr Puppet and anyone else accused of a crime? After all, we have little reason to doubt that (i) the state of the universe 10 000 years ago, (ii) the laws of physics, and (iii) the outcomes of random quantum mechanical events are together sufficient to determine everything that happens nowadays, including our own actions. These things are all clearly beyond our control. So what is the real difference between us and Mr Puppet? One obvious difference is that Mr Puppet is the victim of a diabolical plot whereas most people, we presume, are not. But does this matter? The thought that Mr Puppet is not fully responsible depends on the idea that his actions were externally determined. Forces beyond his control constrained his personality to the point that it was 'no surprise' that he would behave badly. But the fact that these forces are connected to the desires and intentions of evil scientists is really irrelevant, is it not? What matters is only that these forces are beyond Mr Puppet's control, that they're not really *his*. The fact that someone could deliberately harness these forces to reliably design criminals is an indication of the strength of these forces, but the fact that these forces are being guided by other minds rather than simply operating on their own seems irrelevant, so far as Mr Puppet's freedom and responsibility are concerned.

Thus, it seems that, in a very real sense, we are all puppets. The combined effects of genes and environment determine all of our actions. Mr Puppet is exceptional only in that the intentions of other humans lie behind his genes and environment. But, so long as his genes and environment are intrinsically comparable to those of ordinary people, this does not really matter. We are no more free than he is.

What all of this illustrates is that the 'fundamental psycholegal error' is grounded in a powerful moral intuition that the law and allied compatibilist philosophies try to sweep under the rug. The foregoing suggests that people regard actions only as fully free when those actions are seen as robust against determination by external forces. But if determinism (or determinism plus quantum mechanics) is true, then no actions are truly free because forces beyond our control are always sufficient to determine behaviour. Thus, intuitive free will is libertarian, not compatibilist. That is, it requires the rejection of determinism and an implicit commitment to some kind of magical mental causation.[5]

Naturalistic philosophers and scientists have known for a long time that magical mental causation is a non-starter. But this realization is the result of philosophical reflection about the nature of the universe and its governance by physical law. Philosophical reflection, however, is not the only way to see the problems with libertarian accounts of free will. Indeed, we argue that neuroscience can help people appreciate the mechanical nature of human action in a way that bypasses complicated arguments.

6. NEUROSCIENCE AND THE TRANSPARENT BOTTLENECK

We have argued that, contrary to legal and philosophical orthodoxy, determinism really does threaten free will and responsibility as we intuitively understand them. It is just that most of us, including most philosophers and legal

theorists, have yet to appreciate it. This controversial opinion amounts to an empirical prediction that may or may not hold: as more and more scientific facts come in, providing increasingly vivid illustrations of what the human mind is really like, more and more people will develop moral intuitions that are at odds with our current social practices (see Robert Wright (1994) for similar thoughts).

Neuroscience has a special role to play in this process for the following reason. As long as the mind remains a black box, there will always be a donkey on which to pin dualist and libertarian intuitions. For a long time, philosophical arguments have persuaded some people that human action has purely mechanical causes, but not everyone cares for philosophical arguments. Arguments are nice, but physical demonstrations are far more compelling. What neuroscience does, and will continue to do at an accelerated pace, is elucidate the 'when', 'where' and 'how' of the mechanical processes that cause behaviour. It is one thing to deny that human decision-making is purely mechanical when your opponent offers only a general, philosophical argument. It is quite another to hold your ground when your opponent can make detailed predictions about how these mechanical processes work, complete with images of the brain structures involved and equations that describe their function.[6]

Thus, neuroscience holds the promise of turning the black box of the mind into a *transparent bottleneck*. There are many causes that impinge on behaviour, but all of them—from the genes you inherited, to the pain in your lower back, to the advice your grandmother gave you when you were six—must exert their influence through the brain. Thus, your brain serves as a bottleneck for all the forces spread throughout the universe of your past that affect who you are and what you do. Moreover, this bottleneck contains the events that are, intuitively, most critical for moral and legal responsibility, and we may soon be able to observe them closely.

At some time in the future we may have extremely high-resolution scanners that can simultaneously track the neural activity and connectivity of every neuron in a human brain, along with computers and software that can analyse and organize these data. Imagine, for example, watching a film of your brain choosing between soup and salad. The analysis software highlights the neurons pushing for soup in red and the neurons pushing for salad in blue. You zoom in and slow down the film, allowing yourself to trace the cause-and-effect relationships between individual neurons—the mind's clockwork revealed in arbitrary detail. You find the tipping-point moment at which the blue neurons in your prefrontal cortex out-fire the red neurons, seizing control of your pre-motor cortex and causing you to say, 'I will have the salad, please'.

At some further point this sort of brainware may be very widespread, with a high-resolution brain scanner in every classroom. People may grow up completely used to the idea that every decision is a thoroughly mechanical process, the outcome of which is completely determined by the results of prior mechanical processes. What will such people think as they sit in their jury boxes? Suppose a man has killed his wife in a jealous rage. Will jurors of the future wonder whether the defendant acted in that moment *of his own free will*? Will they wonder if it was *really him* who killed his wife rather than his *uncontrollable anger*? Will they ask whether he *could have done otherwise*? Whether he really *deserves* to be punished, or if he is just a victim of unfortunate circumstances? We submit that these questions, which seem so important today, will lose their grip in an age when the mechanical nature of human decision-making is fully appreciated. The law will continue to punish misdeeds, as it must for practical reasons, but the idea of distinguishing the truly, deeply guilty from those who are merely victims of neuronal circumstances will, we submit, seem pointless.

At least in our more reflective moments. Our intuitive sense of free will runs quite deep, and it is possible that we will never be able to fully talk ourselves out of it. Next we consider the psychological origins of the problem of free will.

7. FOLK PSYCHOLOGY AND FOLK PHYSICS COLLIDE: A COGNITIVE ACCOUNT OF THE PROBLEM OF ATTRIBUTIVE FREE WILL

Could the problem of free will just melt away? This question begs another: why do we have the problem of free will in the first place? Why does the idea of a deterministic universe seem to contradict something important in our conception of human action? A promising answer to this question is offered by Daniel Wegner in *The illusion of conscious will* (Wegner 2002). In short, Wegner argues, we feel as if we are uncaused causers, and therefore granted a degree of independence from the deterministic flow of the universe, because we are unaware of the deterministic processes that operate in our own heads. Our actions appear to be caused by our mental states, but not by physical states of our brains, and so we imagine that we are metaphysically special, that we are non-physical causes of physical events. This belief in our specialness is likely to meet the same fate as other similarly narcissistic beliefs that we have cherished in our past: that the Earth lies at the centre of the universe, that humans are unrelated to other species, that all of our behaviour is consciously determined, etc. Each of these beliefs has been replaced by a scientific and humbling understanding of our place in the physical universe, and there is no reason to believe that the case will be any different for our sense of free will. (For similar thoughts, see Wright (1994) on Darwin's clandestine views about free will and responsibility.)

We believe that Wegner's account of the problem of free will is essentially correct, although we disagree strongly with his conclusions concerning its (lack of) practical moral implications (see below). In this section we pick up on and extend one strand in Wegner's argument (Wegner 2002, pp. 15–28). Wegner's primary aim is to explain, in psychological terms, why we attribute free will to ourselves, why we feel free from the inside. Our aim in this section is to explain, in psychological terms, why we insist on attributing free will to *others*—and why scientifically minded philosophers, despite persistent efforts, have managed to talk almost no one out of this practice. The findings we review serve as examples of how psychological and neuroscientific data are beginning to characterize the mechanisms that underlie our sense of free will, how these mechanisms can lead us to assume free will is operating when it is not, and how a scientific understanding of these mechanisms can serve to dismantle our commitment to the idea of free will.

Looking out at the world, it appears to contain two fundamentally different kinds of entity. On the one hand, there are ordinary objects that appear to obey the ordinary laws of physics: things like rocks and puddles of water and blocks of wood. These things do not get up and move around on their own. They are, in a word, inanimate. On the other hand, there are things that seem to operate by some kind of magic. Humans and other animals, so long as they are alive, can move about at will, in apparent defiance of the physical laws that govern ordinary matter. Because things like rocks and puddles, on the one hand, and mice and humans, on the other, behave in such radically different ways, it makes sense, from an evolutionary perspective, that creatures would evolve separate cognitive systems for processing information about each of these classes of objects (Pinker 1997). There is a good deal of evidence to suggest that this is precisely how our minds work.

A line of research beginning with Fritz Heider illustrates this point. Heider and Simmel (Heider & Simmel 1944) created a film involving three simple geometric shapes that move about in various ways. For example, a big triangle chases a little circle around the screen, bumping into it. The little circle repeatedly moves away, and a little triangle repeatedly moves in between the circle and the big triangle. When normal people watch this movie they cannot help but view it in social terms (Heberlein & Adolphs 2004). They see the big triangle as *trying* to harm the little circle, and the little triangle as trying to *protect* the little circle; and they see the little circle as *afraid* and the big triangle as *frustrated*. Some people even spontaneously report that the big triangle is a *bully*. In other words, simple patterns of movement trigger in people's minds a cascade of complex social inferences. People not only see these shapes as 'alive'. They see beliefs, desires, intentions, emotions, personality traits and even moral blameworthiness. It appears that this kind of inference is automatic (Scholl & Tremoulet 2000). Of course, you, the observer, know that it is only a film, and a very simple one at that, but you nevertheless cannot help but see these events in social, even *moral*, terms.

That is, unless you have damage to your amygdala, a subcortical brain structure that is important for social cognition (Adolphs 1999). Andrea Heberlein tested a patient with rare bilateral amygdala damage using Heider's film and found that this patient, unlike normal people, described what she saw in completely asocial terms, despite that fact that her visual and verbal abilities are not compromised by her brain damage. Somehow, this patient is blind to the 'human' drama that normal people cannot help but see in these events (Heberlein & Adolphs 2004).

The sort of thinking that is engaged when normal people view the Heider–Simmel film is sometimes known as 'folk psychology' (Fodor 1987), 'the intentional stance' (Dennett 1987) or 'theory of mind', (Premack & Woodruff 1978). There is a fair amount of evidence (including the work described above) suggesting that humans have a set of cognitive subsystems that are specialized for processing information about intentional agents (Saxe *et al.* 2004). At the same time, there is evidence to suggest that humans and other animals also have subsystems specialized for 'folk physics', an intuitive sense of how ordinary matter behaves. One compelling piece of evidence for the claim that normal humans have subsystems specialized for folk physics comes from studies of people with autism spectrum disorder. These individuals are particularly bad at solving problems that require 'folk psychology', but they do very well with problems related to how physical objects (e.g. the parts of machine) behave, i.e. 'folk physics' (Baron Cohen 2000). Another piece of evidence for a 'folk physics' system comes from discrepancies between people's physical intuitions and the way the world actually works. People say, for example, that a ball shot out of a curved tube resting on a flat surface will continue to follow a curved path outside the tube when in fact it will follow a straight path (McCloskey *et al.* 1980). The fact that people's physical intuitions are slightly, but systematically, out of step with reality suggests that the mind brings a fair amount of implicit theory to the perception of physical objects.

Thus, it is at least plausible that we possess distinguishable cognitive systems for making sense of the behaviour of objects in the world. These systems seem to have two fundamentally different 'ontologies'. The folk physics system deals with chunks of matter that move around without purposes of their own according to the laws of intuitive physics, whereas the folk psychology system deals with unseen features of minds: beliefs, desires, intentions, etc. But what, to our minds, is a mind? We suggest that a crucial feature, if not the defining feature, of a mind (intuitively understood) is that it is an uncaused causer (Scholl & Tremoulet 2000). Minds animate material bodies, allowing them to move without any apparent physical cause and in pursuit of goals. Moreover, we reserve certain social attitudes for things that have minds. For example, we do not resent the rain for ruining our picnic, but we would resent a person who hosed our picnic (Strawson 1962), and we resent picnic-hosers considerably more when we perceive that their actions are intentional. Thus, it seems that folk psychology is the gateway to moral evaluation. To see something as morally blameworthy or praiseworthy (even if it is just a moving square), one has to first see it as 'someone', that is, as having a mind.

With all of this in the background, one can see how the problem of attributive free will arises. To see something as a responsible moral agent, one must first see it as having a mind. But, intuitively, a mind is, among other things, an uncaused causer. Consequently, when something is seen as a mere physical entity operating in accordance with deterministic physical laws, it ceases to be seen, intuitively, as a mind. Consequently, it is seen as an object unworthy of moral praise or blame. (Note that we are not claiming that people automatically attribute moral agency to anything that appears to be an uncaused causer. Rather, our claim is that seeing something as an uncaused causer is a *necessary but not sufficient* condition for seeing something as a moral agent.)

After thousands of years of our thinking of one another as uncaused causers, science comes along and tells us that there is no such thing—that all causes, with the possible exception of the Big Bang, are caused causes (determinism). This creates a problem. When we look at people as physical systems, we cannot see them as any more blameworthy or praiseworthy than bricks. But when we perceive people using our intuitive, folk psychology we cannot avoid attributing moral blame and praise.

So, philosophers who would honour both our scientific knowledge and our social instincts try to reconcile these two competing outlooks, but the result is never completely

satisfying, and the debate wears on. Philosophers who cannot let go of the idea of uncaused causes defend libertarianism, and thus opt for scientifically dubious, 'panicky metaphysics'. Hard determinists, by contrast, embrace the conclusions of modern science, and concede what others will not: that many of our dearly held social practices are based on an illusion. The remaining majority, the compatibilists, try to talk themselves into a compromise. But the compromise is fragile. When the physical details of human action are made vivid, folk psychology loses its grip, just as folk physics loses its grip when the morally significant details are emphasized. The problem of free will and determinism will never find an intuitively satisfying solution because it arises out of a conflict between two distinct cognitive subsystems that speak different cognitive 'languages' and that may ultimately be incapable of negotiation.

8. FREE WILL, RESPONSIBILITY AND CONSEQUENTIALISM

Even if there is no intuitively satisfying solution to the problem of free will, it does not follow that there is no correct view of the matter. Ours is as follows: when it comes to the issue of free will itself, hard determinism is mostly correct. Free will, as we ordinarily understand it, is an illusion. However, it does not follow from the fact that free will is an illusion that there is no legitimate place for responsibility. Recall from § 2 that there are two general justifications for holding people legally responsible for their actions. The retributive justification, by which the goal of punishment is to give people what they really deserve, does depend on this dubious notion of free will. However, the consequentialist approach does not require a belief in free will at all. As consequentialists, we can hold people responsible for crimes simply because doing so has, on balance, beneficial effects through deterrence, containment, etc. It is sometimes said that if we do not believe in free will then we cannot legitimately punish anyone and that society must dissolve into anarchy. In a less hysterical vein, Daniel Wegner argues that free will, while illusory, is a necessary fiction for the maintenance of our social structure (Wegner 2002, ch. 9). We disagree. There are perfectly good, forward-looking justifications for punishing criminals that do not depend on metaphysical fictions. (Wegner's observations may apply best to the personal sphere: see below.)

The vindication of responsibility in the absence of free will means that there is more than a grain of truth in compatibilism. The consequentialist approach to responsibility generates a derivative notion of free will that we can embrace (Smart 1961). In the name of producing better consequences, we will want to make several distinctions among various actions and agents. To begin, we will want to distinguish the various classes of people who cannot be deterred by the law from those who can. That is, we will recognize many of the 'diminished capacity' excuses that the law currently recognizes such as infancy and insanity. We will also recognize familiar justifications such those associated with crimes committed under duress (e.g. threat of death). If we like, then, we can say that the actions of rational people operating free from duress, etc. are free actions, and that such people are exercising their free will.

At this point, compatibilists such as Daniel Dennett may claim victory: 'what more could one want from free will?'. In a word: retributivism. We have argued that commonsense retributivism really does depend on a notion of free will that is scientifically suspect. Intuitively, we want to punish those people who truly deserve it, but whenever the causes of someone's bad behaviour are made sufficiently vivid, we no longer see that person as truly deserving of punishment. This insight is expressed by the old French proverb: 'to know all is to forgive all'. It is also expressed in the teachings of religious figures, such as Jesus and Buddha, who preach a message of universal compassion. Neuroscience can make this message more compelling by vividly illustrating the mechanical nature of human action.

Our penal system is highly counter-productive from a consequentialist perspective, especially in the USA, and yet it remains in place because retributivist principles have a powerful moral and political appeal (Lacey 1988; Tonry 2004). It is possible, however, that neuroscience will change these moral intuitions by undermining the intuitive, libertarian conceptions of free will on which retributivism depends.

As advocates of consequentialist legal reform, it behoves us to briefly respond to the three standard criticisms levied against consequentialist theories of punishment. First, it is claimed that consequentialism would justify extreme overpunishing. As noted above, it is possible in principle that the goal of deterrence would justify punishing parking violations with the death penalty or framing innocent people to make examples of them. Here, the standard response is adequate. The idea that such practices could, in the real world, make society happier on balance is absurd. Second, it is claimed that consequentialism justifies extreme underpunishment. In response to some versions of this objection, our response is the same as above. Deceptive practices such as a policy of faking punishment cannot survive in a free society, and a free society is required for the pursuit of most consequentialist ends. In other cases consequentialism may advocate more lenient punishments for people who, intuitively, deserve worse. Here, we maintain that a deeper understanding of human action and human nature will lead people—more of them, at any rate—to abandon these retributivist intuitions. Our response is much the same to the third and most general criticism of consequentialist punishment, which is that even when consequentialism gets the punishment policy right, it does so for the wrong reasons. These supposedly right reasons are reasons that we reject, however intuitive and natural they may feel. They are, we maintain, grounded in a metaphysical view of human action that is scientifically dubious and therefore an unfit basis for public policy in a pluralistic society.

Finally, as defenders of hard determinism and a consequentialist approach to responsibility, we should briefly address some standard concerns about the rejection of free will and conceptions of responsibility that depend on it. First, does not the fact that you can raise your hand 'at will' prove that free will is real? Not in the sense that matters. As Daniel Wegner (2002) has argued, our first-person sense of ourselves as having free will may be a systematic illusion. And from a third-person perspective, we simply do not assume that anyone who exhibits voluntary control over his body is free in the relevant sense, as in the case of Mr Puppet.

A more serious challenge is the claim that our commitments to free will and retributivism are simply inescapable

for all practical purposes. Regarding free will, one might wonder whether one can so much as make a decision without implicitly assuming that one is free to choose among one's apparent options. Regarding responsibility and punishment, one might wonder if it is humanly possible to deny our retributive impulses (Strawson 1962; Pettit 2002). This challenge is bolstered by recent work in the behavioural sciences suggesting that an intuitive sense of fairness runs deep in our primate lineage (Brosnan & De Waal 2003) and that an adaptive tendency towards retributive punishment may have been a crucial development in the biological and cultural evolution of human sociality (Fehr & Gachter 2002; Boyd *et al.* 2003; Bowles & Gintis 2004). Recent neuroscientific findings have added further support to this view, suggesting that the impulse to exact punishment may be driven by phylogentically old mechanisms in the brain (Sanfey *et al.* 2003). These mechanisms may be an efficient and perhaps essential, device for maintaining social stability. If retributivism runs that deep and is that useful, one might wonder whether we have any serious hope of, or reason for, getting rid of it. Have we any real choice but to see one another as free agents who deserve to be rewarded and punished for our past behaviours?

We offer the following analogy: modern physics tells us that space is curved. Nevertheless, it may be impossible for us to see the world as anything other than flatly Euclidean in our day-to-day lives. And there are, no doubt, deep evolutionary explanations for our Euclidean tendencies. Does it then follow that we are forever bound by our innate Euclidean psychology? The answer depends on the domain of life in question. In navigating the aisles of the grocery store, an intuitive, Euclidean representation of space is not only adequate, but probably inevitable. However, when we are, for example, planning the launch of a spacecraft, we can and should make use of relativistic physical principles that are less intuitive but more accurate. In other words, a Euclidean perspective is not necessary for *all* practical purposes, and the same may be true for our implicit commitment to free will and retributivism. For most day-to-day purposes it may be pointless or impossible to view ourselves or others in this detached sort of way. But—and this is the crucial point—it may not be pointless or impossible to adopt this perspective when one is deciding what the criminal law should be or whether a given defendant should be put to death for his crimes. These may be special situations, analogous to those routinely encountered by 'rocket scientists', in which the counter-intuitive truth that we legitimately ignore most of the time can and should be acknowledged.

Finally, there is the worry that to reject free will is to render all of life pointless: why would you bother with anything if it has all long since been determined? The answer is that you will bother because you are a human, and that is what humans do. Even if you decide, as part of a little intellectual exercise, that you are going to sit around and do nothing because you have concluded that you have no free will, you are eventually going to get up and make yourself a sandwich. And if you do not, you have got bigger problems than philosophy can fix.

9. CONCLUSION

Neuroscience is unlikely to tell us anything that will challenge the law's stated assumptions. However, we maintain that advances in neuroscience are likely to change the way people think about human action and criminal responsibility by vividly illustrating lessons that some people appreciated long ago. Free will as we ordinarily understand it is an illusion generated by our cognitive architecture. Retributivist notions of criminal responsibility ultimately depend on this illusion, and, if we are lucky, they will give way to consequentialist ones, thus radically transforming our approach to criminal justice. At this time, the law deals firmly but mercifully with individuals whose behaviour is obviously the product of forces that are ultimately beyond their control. Some day, the law may treat all convicted criminals this way. That is, humanely.

The authors thank Stephen Morse, Andrea Heberlein, Aaron Schurger, Jennifer Kessler and Simon Keller for their input.

ENDNOTES

[1] Of course, scientific respectability is not everyone's first priority. However, the law in most Western states is a public institution designed to function in a society that respects a wide range of religious and otherwise metaphysical beliefs. The law cannot function in this way if it presupposes controversial and unverifiable metaphysical facts about the nature of human action, or anything else. Thus, the law must restrict itself to the class of intersubjectively verifiable facts, i.e. the facts recognized by science, broadly construed. This practice need not derive from a conviction that the scientifically verifiable facts are necessarily the only facts, but merely from a recognition that verifiable or scientific facts are the only facts upon which public institutions in a pluralistic society can effectively rely.

[2] There are some forms of dualism according to which the mind and body, although distinct, do not interact, making it impossible for the mind to have any observable effects on the brain or anything else in the physical world. These versions of dualism do not concern us here. For the purposes of this paper, we are happy to allow the metaphysical claim that souls or aspects of minds may exist independently of the physical body. Our concern is specifically with interactionist versions of dualism according to which non-physical mental entities have observable physical effects. We believe that science has rendered such views untenable and that the law, insofar as it is a public institution designed to serve a pluralistic society, must not rely on beliefs that are scientifically suspect (see previous endnote).

[3] It is conceivable that rationality could someday be redefined in neurocognitive rather than behavioural terms, much as water has been redefined in terms of its chemical composition. Were that to happen, neuroscientific evidence could then be construed as more direct than behavioural evidence. But Steinberg and Scott's argument appears to make use of a conventional, behavioural definition of rationality and not a neurocognitive redefinition.

[4] This is not to say that we could not describe Mr Puppet in such a way that our intuitions about him would change. Our point is only that, when the details are laid bare, it is very hard to see him as morally responsible.

[5] Compatibilist philosophers such as Daniel Dennett (2003) might object that the story of Mr Puppet is nothing but a misleading 'intuition pump'. Indeed, this is what Dennett says about a similar case of Alfred Mele's (1995). We believe that our case is importantly different from Mele's. Dennett and Mele imagine two women who are psychologically identical: Ann is a typical, good person, whereas Beth has been brainwashed to be just like Ann. Dennett argues, against Mele, that if you take seriously the claim that these two are psychologically identical and properly imagine that Beth is as rational, open-minded, etc. as Ann, you will come to see that the two are equally free. We agree with Dennett that Ann and Beth are comparable and that

Mele's intuition falters when the details are fleshed out. But does the same hold for the intuition provoked by Mr Puppet's story? It seems to us that the more one knows about Mr Puppet and his life the less inclined one is to see him as truly responsible for his actions and our punishing him as a worthy end in itself. We can agree with Dennett that there is a sense in which Mr Puppet is free. Our point is merely that there is a legitimate sense in which he, like all of us, is not free and that this sense matters for the law.

[6] We do not wish to imply that neuroscience will inevitably put us in a position to predict any given action based on a neurological examination. Rather, our suggestion is simply that neuroscience will eventually advance to the point at which the mechanistic nature of human decision-making is sufficiently apparent to undermine the force of dualist/libertarian intuitions.

REFERENCES

Adolphs, R. 1999 Social cognition and the human brain. *Trends Cogn. Sci.* **3**, 469–479.

Baron Cohen, S. 2000 Autism: deficits in folk psychology exist alongside superiority in folk physics. In *Understanding other minds: perspectives from autism and developmental cognitive neuroscience* (ed. S. Baron Cohen, H. Tager Flusberg & D. Cohen), pp. 78–82. New York: Oxford University Press.

Bentham, J. 1982 *An introduction to the principles of morals and legislation*. London: Methuen.

Bowles, S. & Gintis, H. 2004 The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theor. Popul. Biol.* **65**, 17–28.

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. 2003 The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535.

Brosnan, S. F. & De Waal, F. B. 2003 Monkeys reject unequal pay. *Nature* **425**, 297–299.

Dennett, D. C. 1984 *Elbow room: the varieties of free will worth wanting*. Cambridge, MA: MIT Press.

Dennett, D. C. 1987 *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. C. 1991 *Consciousness explained*. Boston, MA: Little Brown and Co.

Dennett, D. C. 2003 *Freedom evolves*. New York: Viking.

Fehr, E. & Gachter, S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140.

Fodor, J. A. 1987 *Psychosemantics: the problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Frankfurt, H. 1966 Alternate possibilities and moral responsibility. *J. Philosophy* **66**, 829–839.

Goldstein, A. M., Morse, S. J. & Shapiro, D. L. 2003 Evaluation of criminal responsibility. In *Forensic psychology*. vol. 11 (ed. A. M. Goldstein), pp. 381–406. New York: Wiley.

Goldstein, A. S. 1967 *The insanity defense*. New Haven, CT: Yale University Press.

Hart, H. L. A. 1968 *Punishment and responsibility*. Oxford Univeristy Press.

Heberlein, A. S. & Adolphs, R. 2004 Impaired spontaneous anthropomorphizing despite intact perception and social knowledge. *Proc. Natl Acad. Sci. USA* **101**, 7487–7491.

Heider, F. & Simmel, M. 1944 An experimental study of apparent behavior. *Am. J. Psychol.* **57**, 243–259.

Hughs, R. I. G. 1992 *The structure and interpretation of quantum mechanics*. Cambridge, MA: Havard University Press.

Kahneman, D. & Tversky, A. (eds) 2000 *Choices, values, and frames*. Cambridge University Press.

Kant, I. 2002 *The philosophy of law: an exposition of the fundamental principles of jurisprudence as the science of right*. Union, NJ: Lawbook Exchange.

Lacey, N. 1988 *State punishment: political principles and community values*. London and New York: Routledge & Kegan Paul.

McCloskey, M., Caramazza, A. & Green, B. 1980 Curvilinear motion in the absence of external forces: naive beliefs about the motion of objects. *Science* **210**, 1139–1141.

Mele, A. 1995 *Autonomous agents: from self-control to autonomy*. Oxford University Press.

Morse, S. J. 2004 New neuroscience, old problems. In *Neuroscience and the law: brain, mind, and the scales of justice* (ed. B. Garland), pp. 157–198. New York: Dana Press.

Pettit, P. 2002 *The capacity to have done otherwise. Rules, reasons, and norms: selected essays*. Oxford University Press.

Pincus, J. H. 2001 *Base instincts: what makes killers kill?* New York: Norton.

Pinker, S. 1997 *How the mind works*. New York: Norton.

Premack, D. & Woodruff, G. 1978 Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **4**, 515–526.

Rosen, G. 2002 The case for incompatibilism. *Philosophy Phenomenol. Res.* **64**, 699–706.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. 2003 The neural basis of economic decision-making in the ultimatum game. *Science* **300**, 1755–1758.

Saxe, R., Carey, S. & Kanwisher, N. 2004 Understanding other minds: liking developmental psychology and functional neuroimaging. *A. Rev. Psychol.* **55**, 87–124.

Scholl, B. J. & Tremoulet, P. D. 2000 Perceptual causality and animacy. *Trends Cogn. Sci.* **4**, 299–309.

Shear, J. (ed.) 1999 *Explaining consciousness: the hard problem*. Cambridge, MA: MIT Press.

Smart, J. J. C. 1961 Free will, praise, and blame. *Mind* **70**, 291–306.

Steinberg, L. & Scott, E. S. 2003 Less guilty by reason of adolescence: developmental immaturity, diminished responsibility, and the juvenile death penalty. *Am. Psychol.* **58**, 1009–1018.

Strawson, P. F. 1962 Freedom and resentment. *Proc. Br. Acad.* **xlviii**, 1–25.

Tonry, M. 2004 *Thinking about crime: sense and sensibility in American penal culture*. New York: Oxford University Press.

Van Inwagen, P. 1982 The incompatibility of free will and determinism. In *Free will* (ed. G. Watson), pp. 46–58. New York: Oxford University Press.

Watson, G. (ed.) 1982 *Free will*. New York: Oxford University Press.

Wegner, D. M. 2002 *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wright, R. 1994 *The moral animal: evolutionary psychology and everyday life*. New York: Pantheon.