There but for the grace of my orbitofrontal cortex... - Review of Thomas Nadelhoffer (ed.), “The Future of Punishment”, OUP

The human brain, with its one hundred billion neurons working in intricate collaborations to create the physical basis of the memories, perceptions, thoughts, and emotions that together make me, the person that I am, is surely one of the marvels of the world. We tend to forget how extraordinary it is, as well as the extent to which who we are and what we do depends on the brain, perhaps because it is such a superficially inert and uninteresting part of the body, immobile and hidden inside our heads. At least until something goes wrong.

Imagine the following scenario – doing so is likely to be uncomfortable, but bear with it for a minute: You are a happily married, ordinary, law-abiding citizen, but over the course of a few months you develop stronger and stronger sexual urges and fight a series of steadily more losing battles against them. You begin to use pornography almost addictively, and to your horror and shame you find yourself increasingly sexually attracted to pre-teen children. You struggle with it, developing headaches, while concentration and focus become more and more difficult. Your spouse learns that you have been flirting with your stepchild, and promptly kicks you out and presses charges. The police discover that there is child pornography among the enormous stash of porn on your computer. Your life is in ruins. Faced with a choice between prison and treatment you opt for the latter, only to find yourself expelled from the programme when you cannot stop yourself from making overt sexual advances towards the other clients and staff. Having failed to complete the programme, your prison sentence is effectuated, but the night before it is to commence the situation comes to a head. You are confused and dizzy, in the grip of enormously strong impulses, and terrified both at what has happened to you and what is to come. You give serious thought to taking your own life; there are moments when it seems clearly the best option. And you fight the temptation to assault and rape the elderly person who rents you a small room, the escalating sexual tension almost beyond your control. Desperate, and feeling what little remains of your self-control slipping, you drive to the nearest psychiatric ward and ask to be admitted. The staff give you a series of routine checks, and find that you have balance and cognitive problems, so they order an MRI scan. The scan shows a tumour almost the size of a chicken egg in the orbitofrontal cortex of your brain; the areas of the brain affected are known to be central to impulse control and practical judgement. On surgical removal of the tumour, the urges disappear instantly. You check into and complete the treatment programme without further problems. A year later you find the urges and headache returning, contact the neurologydepartment and have a new MRI, which shows that the tumour has regrown. Once this is removed, again the urges disappear, and you are able to resume picking up the pieces of your life.

Unlike many thought-experiments employed in philosophy, this scenario is no mere hypothetical. Roughly the above occurred in a much discussed case, involving a relatively ordinary 40-year old US male. (Burns & Swerdlow, 2003) What are we to think about a case like this? To what extent should you blame yourself? Many will probably, like me, feel intuitively ambivalent about it. On the one hand, your urges and actions are repugnant; as a decent person you will feel shame, regret, and self-loathing. On the other hand, a string of sympathy may be evoked by the thought that you were not properly in control of your actions, or perhaps even that you were not yourself. There seems to be an important difference between the two You's, or the You at the two points in time where you suffer overwhelming urges due to the growing tumour and where you lead a normal, respectable life without it. Surely this should have implications for what we think of you and what we can justify as a response to your actions? Would it not be patently unfair, for example, to blame or punish the healthy, ordinary you for whatever crimes your tumour-afflicted self committed? Indeed, would it not be unreasonable to blame or punish your tumour-afflicted self, given that the tumour had damaged, perhaps even disabled, your capacity for impulse control and practical judgement? What on earth were you supposed to have done, when the very parts of your brain necessary for being able to make decisions in the right way had been put out of action?

However, even if you feel the pull of these intuitions, you may worry that the slippery slope down this rabbit-hole goes very deep indeed. Consider: if your tumour-inflicted self ought not be punished because your brain limited your ability to do the right thing, then what about the ordinary you, not to mention the rest of us? Certainly it seems true that we are all limited in our ability to act by the way that our brain is organised and operates; even among those who resist full-fledged reductive physicalism, that is, the idea that the physical brain is all there is to the mind, very few people today believe that we can think or act independently of the neural basis of the brain. But if nobody can ever act in a way other than what the limits of their brain allow, should we ever blame or punish anyone?

Intuitions like the above are not, of course, conclusive, and there is a longstanding and rich debate in both philosophy and law about the conditions and possibility of moral and criminal responsibility in the face of such apparent counter-intuitions. The underlying problems we face in cases like Burns & Swerdlow have, however, been garnering increasing attention from cognitive scientists, philosophers and legal thinkers, particularly in the last decade, because it is still an area of many questions and few answers, and one where advancements in the cognitive sciences continually raise new issues. So the recent anthology edited by

Thomas A. Nadelhoffer, “The Future of Punishment”, part of Oxford University Press’ “Neuroscience, Law, and Philosophy”-series, is a timely publication.

The challenge

The twelve contributions of the volume are divided into five themes, the first three of which all deal directly with how to justify punishment in the face of both the new problems raised by our increasing understanding of how the brain works and the hoary philosophical chestnuts of determinism and luck. Oddly enough, to my knowledge no one has actually presented a detailed version of the challenge to punishment in the academic literature, although it is frequently encountered in public debate. A version of it is sometimes attributed to Joshua Green and Jonathan Cohen (Greene & Cohen, 2004), but as several of the authors (Mele; Pardo & Patterson) in this volume rightly emphasise, the argument they make is subtly different. Greene and Cohen argue for what could be labelled a predictive sociological thesis, that as knowledge of the results of neuroscience disseminate through society, intuitive support for punishing according to desert will evaporate, and public support for retributive criminal justice will diminish to the point where we will replace it with other, more enlightened (according to Greene and Cohen) principles of criminal justice. This is a different argument than the normative, prescriptive argument, that retributive justice is unjustified because nobody can meet the necessary conditions for deserving to be punished. That argument might go something like the following:

1) Cognitive science shows facts X about the human mind.
2) Free will and/or moral responsibility are incompatible with facts X.
3) A person morally deserves something on the basis of an action F only if the person has freely willed and/or is morally responsible for F.
4) Persons cannot deserve anything on account of their actions (from 1-3).
5) Retributivism justifies punishment by persons deserving to be punished on the basis of their wrongdoing.

QED. Retributivism cannot justify punishment (from 4 and 5).

Call this “The Cognitive Science Challenge to Retributive Punishment”.

Cognitive science, free will, and moral responsibility

Starting with the first two premises of the challenge, what are the facts that are supposed to be incompatible with free will and moral responsibility? There are, broadly speaking, two lines of argument. Greene and Cohen gesture in the direction of neurological determinism, that is, the idea that the brain is a physical system whose state at a given time along with relevant agent-external stimuli and the laws of nature entails the state of the brain at any subsequent time. This, in combination with the idea that we cannot act independently of the state of the brain, might be taken to undermine free will. The other suggestion in the literature is based on the results of the famous Libet-experiments and their successors led by John-Dylan Haynes (Libet, Gleason, Wright, & Pearl, 1983; Libet, 1999; Soon, Brass, Heinze, & Haynes, 2008; Bode et al., 2011; Haynes, 2012). The most recent studies allow the scientists to predict which hand, left or right, a subject will voluntarily move, well before the subject consciously makes up her mind on the issue. Proponents of the argument claim that the experiments show that decisions (if we even want to call them that) are made by the brain and disseminate to consciousness, which then has the false experience of deciding – false because the choice has already been made. And furthermore, that an agent cannot have free will or moral responsibility if her conscious will is merely a sideshow with no real influence on which action the agent performs.

It is also possible, of course, to substitute a different or more general challenge for the challenges attributed to cognitive science, such as the incompatibilism of moral responsibility with universal determinism or with luck. Among the contributors to the volume Neil Levy has done so elsewhere while Derk Pereboom takes the opportunity to present both a critique of libertarianism and his four-stage manipulation case. (Pereboom, 2001, 2007; Levy, 2011) Pereboom's central argument relies on the idea that if two agents, or the same agent at two different points in time, are alike in all plausibly relevant respects and we are convinced that one of them is not morally responsible, then we must say the same about the other. But, Pereboom argues, we can start with a case that any reasonable person will agree involves an agent lacking moral responsibility – say, some poor test-subject of a sci-fi neuroscientist, who has had neural regulators implanted that fire impulses so as to direct her every thought and action – and move in incremental steps to a perfectly ordinary person. Suppose for instance that the scientist has merely shaped the neural structure to her designs, so that it will fire in exactly the way she wants without the use of constant signals from her implants. Presumably this person is no more responsible than the first, after all, her every thought and action have still been pre-determined by the scientist, even if the manipulation takes place further back in time from her actions. But then we can take another small step, and so on, and so on (Pereboom works with four cases, but we could individuate as much as we like). If we make the steps small and innocuous enough, none of the individual steps can plausibly make a difference to the agent's responsibility, but then by pair-wise comparison we can derive the surprising conclusion that ordinary agents are no more responsible than the agent in the original case. (p.58-59)

We have three arguments against free will and moral responsibility then. One that agents actions are tied to their determined physical brains, a second that the conscious will is causally inefficacious, and a third that there is no plausibly relevant difference between those who lack moral responsibility and those we think have it, and hence we must be mistaken about the latter. What are we to say in response?

One response is to deny the first premise, that cognitive science establishes facts allegedly incompatible with free will or moral responsibility. Among the contributors, Alfred Mele has published influential arguments to that effect (Mele, 2009), but here it is a point only briefly touched upon by Nancey Murphy, and Farah Focquart, Andrea Glenn and Adrian Raine, respectively.

The second premise receives more attention: does moral responsibility genuinely require the conditions that the challenge claims are unmeetable? John Martin Fischer is perhaps the most prominent contemporary defender of (semi-)compatibilism, the theory that agents have moral responsibility, and that this is compatible with determinism, so it comes as no surprise that he argues here for a Strawsonian account of moral responsibility, where you can deserve reactive attitudes, and this in turn justifies punishment. Much of his chapter is devoted to rebutting a different set of charges against retributivism made by David Boonin and David Dolinko, and draws extensively on a previously published critique. (Boonin, 2008; Dolinko, 1991, 1992, 1997; John Martin Fischer, 2006) But towards the end he suggests that agents can deserve to suffer because they 1) "[offend] against a law that is designed to protect people against certain sorts of rights violations" or 2) "[fail] to fulfill at least a prima facie moral duty [to obey the law]" (p.19;20). Why think that agents deserve such things? Because, as Peter Strawson argued, moral responsibility simply *means* being the appropriate target for a set of reactive attitudes, including blame and resentment, that is, one is morally responsible if and when one is such an appropriate target. (Strawson, 2008) And, Fischer argues, being an appropriate target of an attitude simply *is* deserving that attitude. How, ultimately, does that translate into punishment? "It seems perfectly reasonable that, if a morally responsible agent deserves indignation or resentment on the basis of his behavior, he might also deserve harsh treatment for so behaving." (p.23)

Here, at least, the footwork looks a little too fast. If moral responsibility simply *is* being an appropriate target for reactive attitudes, and being such an appropriate target simply *is* deserving something, then the first half of the claim is tautological: a morally responsible agent does not "deserve indignation", or, at least, the claim that he does says nothing more than what labelling him morally responsible did. The temptation is to let our translation of the concepts run both ways, that is, to think that now we have defined 'desert' as "being an apt target of an attitude", 'being an apt target of an attitude' also means "deserving something" (in the old fashioned sense, where it can justify punishment). But that will not do.

Furthermore, the entailment to the second half of the claim looks distinctly unsupported. This, after all, is the problem of harsh treatment familiar from expressive and communicative theories of punishment (Duff, 2001; Von Hirsch & Ashworth, 2005). Being an appropriate target of a reactive attitude, even a negative attitude such as blame, does not in any obvious and uncontroversial way mean that others are also justified in harming one.

Other arguments for compatibilist grounds of retributive desert are presented by Stephen Morse, and Michael Pardo and Dennis Patterson. Morse presents his well-versed argument that positive law does not employ the concepts of moral responsibility, free will or retributive desert, and is therefore not directly threatened by the challenge, in a revised version of a 2013 article. (Morse, 2004, 2007, 2010, 2011, 2013) But he grudgingly recognizes that *justifying* the concepts employed in positive law does require an account of moral responsibility, and suggests that compatibilism will suffice. Pardo and Patterson's chapter is primarily an extended and insightful critique of Greene and Cohen's argument, but along the way they argue in favour of a reasons-responsive compatibilism, qua John Martin Fischer (J.M. Fischer & Ravizza, 2000), as an account of moral responsibility that can ground desert.

One difficulty here is that compatibilism is not one theory, but rather a family of accounts of moral responsibility whose only necessarily shared feature is that they affirm that moral responsibility is compatible with determinism. There are a plethora of compatibilisms in the literature, dividing on such essential issues as whether the agent must be capable of doing otherwise (the principle of alternate possibilities, or PAP), must have a certain history, such as being free from manipulation (historical vs non-historical), and the capacities required, including control, sourcehood, reasons-responsiveness, consistency between first- and second-order preferences, etc. (excellent overviews are provided by McKenna, 2009; Levy & McKenna, 2009) Which of these is the type of compatibilism that is meant to provide moral responsibility as the basis of desert? The question is not trivial, because different compatibilisms face different theoretical as well as practical challenges. That is, some of them may focus on a mistaken or insufficient set of conditions for moral responsibility – a set of conditions that even if they did obtain, would not provide free will or moral responsibility; given that extant versions of compatibilism are competing theories, this is necessarily the case for some of them – and some of them may focus on conditions that turn out to be impossible for agents to realise. While they are all designed to avoid the challenge of determinism, it is not obvious that they are all equally well-equipped to ward off other cognitive science challenges, such as the challenge to the causal efficacy of conscious will. The twin challenges makes the obvious reply of "just slotting in whatever account of compatibilism turns out to be the right one" unavailable, because we cannot take it for granted that there is an account which successfully meets both challenges.

Furthermore, it is not given that any type of moral responsibility can support moral desert. It may be tempting to take this as true by definition, that is, to say that moral responsibility simply *is* that quality of the agent which supports moral desert, but that does not remove the problem. It is a further step, for any substantial account of moral responsibility, to argue that being morally responsible in the sense at stake *also* means that the agent can accrue moral desert. If this looks dubious, then bear in mind the different senses of moral responsibility we might employ. Derk Pereboom rightly underscores that there is nothing in either the cognitive science challenges or the traditional philosophical challenges to free will and moral responsibility that denies the ability of ordinary persons to act according to reasons, and that we are thus uncontroversially morally responsible in the sense of being in a position to answer for our actions by accounting for our reasons. (p.51) "What the hell do you think you are doing?" remains a sensible question to ask of the idiot cutting ahead in the queue, on any of the plausible views, even the skeptical views, of free will and moral responsibility. But this is not the same as being morally responsible in the desert-generating sense, and we need an argument to the effect that one entails the other: "Although it may be natural for us to respond to reasons-responsive actions with reactive attitudes, and for us to refrain from or withdraw such responses when actions fail to meet this standard, it is a further task to show that the reasons responsiveness of an action legitimates the attribution of basic desert moral responsibility and the attitudes that presuppose it." (p.60)

So the retributive compatibilist must supply an account of compatibilism which at once looks theoretically plausible, that is, supplies what is on reflection a set of conditions that would make an agent morally responsible for an action, which is practically feasible, that is, employs conditions which are consistent with our current best understanding of the properties human agents are endowed with, and which supports moral desert. Do Morse, or Pardo and Patterson succeed?

Morse's argument consists of a series of claims: first, most philosophers are compatibilists, which indicates that it is at least a plausible theory. Second, the intuitions that drive philosophical arguments are often based on "complicated, unrealistic hypotheticals". Morse does not spell out the implication of this, but from the context it appears he takes it to mean that the intuitions are uncertain, and that this weakens the challenges to free will and moral responsibility. Third, criminal justice cannot wait for a consensus to emerge in the free will debate, and most people involved on a practical level with criminal justice do not, as a matter of fact, care about it. And fourth, to recognize that agents lack moral responsibility would be to impoverish the world. Thus: "it is inconceivable that a practical institution such as the criminal law would transform itself in apparently negative ways as the result of an insoluble theoretical, metaphysical debate." (p.125)

There are a number of serious problems here. First, as is clear, Morse does not supply any positive account of compatibilism, and as such does not even attempt to solve the task indicated above. He suggests in the same section that the burden of proof is on the skeptic, but this strikes me as unpersuasive both because some (e.g. Pereboom) have actually supplied the positive arguments he requires, which shifts the burden of response onto Morse and other proponents of compatibilism, and because it is not clear to me that we should accept that the burden of proof is as uneven as Morse suggests. After all, he, and those inclined as he is, are defending a system which imposes harm on persons, and the justification for doing so is that the harm is deserved. Surely it is not unreasonable to ask that an alleged justification for causing harm be spelled out in an accurate, plausible and consistent way? But doing so entails providing the account of moral responsibility and desert at stake. Second, Morse takes the artificiality of intuitions in the philosophical debate to mean that they are uncertain. I am inclined to be moderately sympathetic to this point, but it is not clear how this could be a problem for one party to that debate, that is, the free will skeptics, rather than for all involved, including defenders of compatibilism. Third, the fact that we cannot postpone criminal justice until we obtain sufficient consensus in the free will debate cannot in itself constitute an argument for any particular position in that debate. It merely suggests that we must take stock and adopt the theory that currently looks most plausible, in spite of the uncertainty. This could be compatibilism, but Morse gives us no reason to think that it is, and whether politicians, judges, lawyers and others involved with criminal justice on a practical level happen to care about the free will debate seems to me mostly besides the point. Morse's final claim is that "interpersonal life would be exceptionally impoverished if concepts of responsibility, including genuinely deserved praise or blame, were extirpated from our lives". (p.125) While superficially appealing, this sounds odd on reflection. Consider that, if the skeptic is right, then our attributions of desert are mistaken, and our praise or blame has been, is, and will be unjustified as responses to an agent's desert (there can be other justifications, or reasons, for blame or praise, but they are not the ones at stake). But surely, realising this is an improvement not an impoverishment? After all, if we ordinarily blame someone for having done wrong, and then discover that the blame was unwarranted, we take ourselves to have wronged that person, and relish the chance of retracting our blame. And we ought similarly to rejoice at a realisation that improves our ability to avoid blaming persons who do not deserve blame in the future. Morse's reasoning appears to presuppose that blaming and praising people on account of desert is valuable, and that losing opportunities to do so is therefore an impoverishment. But that only works when people genuinely deserve blame or praise, which is exactly the point denied by the challenge.

Pardo and Patterson come closer. In their words, “moral desert may be grounded in the control people have over their actions through the exercise of their practical rationality”. (p.145) What kind of control is this and why is it sufficient? The kind that, in spite of determinism, allows persons to do otherwise (PAP), specifically by having both the ability and the opportunity to act in several ways. And as they point out, agents can possess the ability to perform an action, say, ride a bicycle, even on the occasions where they do not perform that action, just as agents can possess an opportunity to do something even when they do not seize it – that after all is part of the meaning of it being an opportunity. How do we know whether agents have opportunities? “One has an opportunity to act (or not to act) if conditions *external* to the person are not forcing or preventing the exercise of the ability on a particular occasion.” (p.147)

This is non-historical, reasons-responsive, leeway compatibilism, which has been defended in the past by Daniel Dennett and others. (Dennett, 1984, 2003) But that does not make it an unproblematic position. Pardo and Patterson preempt one type of objection, which we have already encountered. For would not the agent’s particular brain state at a particular time constrain the actions available to the agent? After all, as previously discussed we do not want to introduce the shaky assumption that agents can act independently of their brain states. But no, say Pardo and Patterson, because the agent’s brain state would have been different, if she had *wanted* to act differently. This is a variation on the classical compatibilist notion that an agent can act differently if, had she wanted to act differently, she *would* have acted differently. In that similarity lies also the evident challenge, for as critics of classical compatibilism have keenly pointed out, it seems to merely shift the question of opportunity back one step: could the agent have wanted something different? (cf. e.g. Van Inwagen, 1983, pp. 114-121) Only, it seems, if either elements of her constitution that she did *not* have the ability and opportunity to affect (say, her motivational structure) or external circumstances had been different.

Consider our unfortunate original case. Are you morally responsible for the actions you take while the tumour grows in your frontal cortex? It seems Pardo and Patterson must say yes, because you have both the ability to do otherwise, and the opportunity to do so. After all, if you had wanted to do otherwise, then your brain state would have been different, say, tumour-free. Note that this is not a theory of wishful thinking – Pardo and Patterson are not saying that your wanting something else could miraculously cure your tumor. It is rather a counterfactual – if you had wanted to do otherwise, then necessarily you would also have been tumour-free, since you could only want otherwise without the tumour – and it is the existence of such a counterfactual which is supposed to make you responsible for your actual, tumour-influenced actions. Is this plausible? Contemporary dispositionalists are still developing and defending versions of the position, but I think it is safe to say that it faces very serious objections. (Vihvelin, 2004; Clarke, 2009; Berofsky, 2011)

Desert, retributivism and punishment

A second type of objection to the challenge might start with the question: why the particular focus on retributivism? One reason is that, even if, as several authors (Morse; Pardo and Patterson) claim, penal policies over the past decades are in fact contrary to what retributivism would prescribe, it, and mixed-theories that include retributivist considerations, has arguably been the dominant theory of criminal justice ethics. The negative retributivist principle in particular, which holds that *pace* consequentialist penal ethics undeserved punishment is unjustifiable, has enjoyed widespread acceptance, but the positive retributivist principle, that there are moral reasons to punish those who deserve to suffer, has also had influential proponents (e.g. Moore, 1997). A second reason is that the problems seem to be particular to retributivism, or at least particularly serious for retributivism. Consequentialists, whose sole concern is the costs and benefits achieved by the available policy alternatives, can largely shrug their shoulders and carry on, even if it turns out that persons are not morally responsible and cannot deserve.

But then how serious is the challenge for retributivism? There could be at least two types of response. One would be that retributivism does just fine, because persons can deserve punishment on account of wrongdoing even when they are not morally responsible for doing wrong (attacking premise three). This line, however, is not pursued. A second is that, even if desert-based retributivism is in trouble, we can get something which works in roughly the same way, that is, prescribes more or less the same punishment in more or less the same situations, perhaps even for more or less the same (i.e. non-consequentialist) type of reasons, without relying on desert; call it quasi-retributivism if you like. This grants premise five, but suggests that it is too narrowly focused, and that the wider alternative, that no satisfactory retributive'ish principle is available without desert, is false. It is the strategy pursued by Shaun Nichols and Michael Louis Corrado in their respective two chapters.

The quasi-retributivist strategy will be satisfactory for those who are motivated not so much to preserve retributivism as to avoid the twin spectres of consequentialism and therapeutic imprisonment looming in the background once it is gone. In Corrado's formulation this is "the awful outcome": "Those who are competent to guide their own behavior [...] are not to be treated differently from those who are incompetent because of mental illness, or because of severe addiction; indeed, they will not be treated differently from misbehaving children. All who present a danger will be subjected to therapy until they are no longer dangerous, or, if therapy is ineffective, they will remain in detention." (p.81)

This is no mere strawman. Pereboom's suggested alternative to retributive and consequentialist punishment is incapacitation, analogous to the strategies employed in cases of quarantine: we seclude some persons under conditions where they are incapable of doing harm that they would otherwise have done. He wants, however, to avoid allowing the detention of persons who have as yet done no wrong, since "the right to liberty must carry weight...as should the concern about using people merely as means" and "the risk posed by a state policy that allows for preventative detention of non-offenders", i.e. through misuse. (p.73) But the argument here is unpersuasive. If we are detaining so as to incapacitate, then *all* detention is preventative, and whether the person has or has not committed a crime can have no relevance other than as a possible indicator of future risk. We deprive liberty and treat as means or ends in exactly the same way for both groups, and there is plenty of risk for misuse of incapacitation-policies even where they are limited to those who have committed a crime, just as states likely to misuse such policies seem well capable of misusing officially retributive policies too. If not quite "the awful outcome", the resulting policy is so close as to probably make no difference to those inclined as Corrado is.

But then what is the alternative? We need a justification for drawing a line between those traditionally labelled responsible and those not, one that simultaneously justifies "intentional harsh treatment" for one group and therapy for the other, but that does not rely on desert. Corrado's suggestion is that instituting such a policy is justified by the benefits it provides, and that the capacity to benefit will roughly follow the traditional distinction along responsibility. As he recognizes this is apparently an incredible claim. In what possible sense could being imprisoned, for months, years, or even decades at a time, and under the brutal conditions prevalent in most prisons, constitute a benefit to the convict? His answer is that the institution provides persons with the benefits of 1) an opportunity to "learn how to live within the rules of society", 2) an opportunity to "learn the significance of rules", and 3) avoiding "the awful outcome", including the possibility of indefinite, therapeutic detention.

I confess to finding this an incredible claim not merely at first glance but also upon careful consideration. Consider the last benefit first: for whom, exactly, will it be a benefit? Under "the awful outcome" (TAO), some persons who would be imprisoned under Corrado's scheme will instead be released because they do not constitute a danger. Clearly, these benefit from TAO, rather than Corrado's scheme. Others will be released under Corrado's scheme, who would have been detained under TAO, but these are the ones who we have strong reason to keep imprisoned, because they constitute a real and present danger. Although releasing them clearly constitutes a benefit for them, it is not clear that providing them with this benefit constitutes an all-things-considered point in favour of Corrado's scheme. Similar problems afflict the first two benefits. Corrado is careful not to insist that the benefit is to *actually* learn one or the other, since many of those subjected to intentional harsh treatment will likely not learn anything. But how much of an opportunity is supposed to be enough? For many, these opportunities may be opportunities in a very formal sense only – they may be literally incapable of seizing that opportunity due to lack of motivation, knowledge, or skills. But in that case, is it plausible that the bare opportunity is a benefit? If on the other hand we restrict opportunity to those situations where there is a sufficient likelihood that someone will learn, then only a sub-group, perhaps even a minority, of those subjected to intentional harsh treatment will be given the benefit.

Nichols shoulders an even heavier burden than that of simply giving up the concept of desert, for he wants to provide an argument that is not grounded in moral realism, and needs not assume that there are objective values or reasons. His solution is to focus on what he labels a "bare retributive norm", which is essentially the retributivist principle that we should punish a person because but only because of her past wrong-doing without the desert-based justification (p.26-27) The norm is, he argues, both widespread and inferentially basic, which is to say that it is a norm that people hold without having derived it from other norms, values or facts, and which is therefore not subject to revision through challenging other elements of a person's value or belief-system. It is better understood, he claims, as a cultural mediation of the primitive emotion of anger. And respecting it will give us a system of punishment that balances our other concerns, such as to avoid causing harm, in the way that many find intuitively attractive about retributivism.

An obvious problem is that once you have given up on objective values and reasons, it becomes hard to phrase an argument as anything other than either a blunt descriptive statement or pathos-driven persuasion. Thus, much of Nichols text has an oddly passive character, as when he suggests that "we might adopt" what he calls "ethical conservatism", the idea that we preserve some norms, and give priority to those that are widespread, inferentially basic and emotionally driven. (p.38) Yes, we might, and if we did, we could consistently follow the argument through. But why should we? For the sake of mere convenience, that is, simply *because* it will allow us to hold the retributive norm? That does not sound like the right kind of reason to adopt a norm-selecting principle (or, perhaps, any principle). Similarly, Nichols writes that "we don't *need* to have a justification for the retributive norm in order for it to retain its legitimacy for us." (p.37) But note how oddly ambiguous this claim is. On a psychological reading, it may well be true – perhaps the retributive norm is a kind of norm that persons will find legitimate even though they have no reason to do so. But this is not a particularly pertinent feature in this context. And on a more ambitious reading, it far from clear that the claim is true, for surely it is hard to see what could make the norm actually, as opposed to apparently, legitimate for a person, except the presence of just the kind of justification that Nichols wants to do without. The closest we get to an argument, apart from the reference to a forthcoming publication which may clarify the matter, is the suggestion that once moral realism is out, we can either chuck all norms with it, or we can keep and give priority to some: "If *none* of our ethical beliefs has an ultimate justification, then, barring a complete upheaval of commonsense ethics, we are bound to grant normative clout to *some* moral norms that lack any ultimate justification." (p.38) While an accurate description of the alternatives, this leaves it unclear both why we should prefer preserving some norms to moral nihilism (obviously, I do prefer one to the other, but do I have any reason to do so?), as well as why we should choose to preserve the norms that Nichols favours. Just under the surface of the discussion lurk heavy-duty metaethical issues, and perhaps it is inevitable both that they cannot be covered in a short chapter, and that one's position on these may make one more or less sympathetic to Nichols' argument. But as an independent defence of quasi-retributivism it seems to me insufficient.

Intuitions and borderline cases

The final contribution under the third theme signals the coming shift in the focus of the fourth and fifth themes. Nancey Murphy sketches an account of the capacities involved in responsible agency informed by the insights of the cognitive sciences, and argues that our understanding of these capacities count in favour of a shift towards restorative justice, which would serve to enhance such capacities, instead of damaging them, as is likely true of current forms of punishment.

The fourth theme then focuses on empirical investigation and explanation of the intuitions central to the debate. Does the average person feel that free will is incompatible with determinism? What about moral responsibility or blame? Since intuitions play such an important role in the discussion, with countless thought-experiments designed to elicit intuitions in support of one view or another, and since everyone agrees that intuitions should be, well, intuitive, whether and to what extent the general public actually share a certain intuition is undoubtedly pertinent. The experimental philosophy work being done in that field is new and promising, and the results presented in these chapters are to my mind among the most interesting in the volume. Alfred Mele presents empirical evidence against the thesis that folk-psychology is substance-dualist libertarian, as alleged by Greene and Cohen. Meanwhile Thomas Nadelhoffer, Dena Gromet, Geoffrey Goodwin, Eddy Nahmias, Chandra Sripada, and Walter Sinnott-Armstrong's collective work illustrates the potential of interdisciplinary collaboration. The group presents a wealth of studies that show the various and often subtle ways that intuitions about free will and moral responsibility can be affected to argue that skepticism is ultimately triggered by challenges to the "deep self". Briefly, the idea is that persons consist of both their immediate judgements, desires, and decisions, and the cognitive and evaluative structure which under normal circumstances is used to review and revise such judgements, desires, and decisions. Intuitively, it appears, challenges to the former do much less to shake confidence in free will and moral responsibility than challenges to the latter. Finally, Eyal Aharoni and Alan J. Fridlund attempt a socio-biological explanation of how retributive intuitions could emerge, arguing that they serve to protect social collaborators valuable to the individual, a hypothesis for which they present two types of supportive evidence.

In the fifth and last theme Neil Levy, and Farah Focquart, Andrea Glenn, and Adrian Raine address a more concrete question: given that we want to hold persons responsible to differing degrees, how and where should we draw that line? Everyone agrees that there are persons who should not be held responsible at all – infants and the severe mentally handicapped, at the very least – but how much further does that list extend? What about children and adolescents? Senile elders? Schizophrenics and psychotics? These are important questions in their own right, but they also feed back into the more fundamental discussion because adopting a particular stance on one this issue will have implications for what we can think about the other. Levy considers the particular problem of addiction, to argue against Gene Heyman's influential thesis that addicts possess sufficient choice to be accountable for the actions they take while in the grips of addiction. (Heyman, 2009) Focquart, Glenn and Raine focus on psychopaths to argue that our best current understanding of the condition indicates that they are insensitive to moral reasons, and as such cannot be considered responsible.

In conclusion

As is hopefully apparent from the above discussion this is a rich debate with many engaging issues. Given the wealth of topics and arguments of the volume, I cannot hope to do the individual chapters justice – even those chapters I have discussed in some length all contain much that I have not had a chance to present, evaluate, and particularly praise.

Part of the challenge of working with these questions lies in their strongly interdisciplinary nature, and it is gratifying to see prominent representatives from a diversity of fields as contributors to the volume. After all we need to simultaneously grapple with questions such as 1) under what conditions (if any) we can justify punishing a person for her wrongdoing, 2) what it means to deserve to be punished, and under what conditions (if any) a person can do so, 3) what it means to be morally responsible, and under what conditions (if any) a person can be so, 4) what the connection (if any) is between free will and moral responsibility, and 5) what developments in the cognitive sciences tell us about agency and the mind, particularly as this concerns the conditions for free will and/or moral responsibility. Each of these is a complex topic in its own right, which raises a slew of more specific questions. Given the constraints imposed by the need for specialisation, few scholars can claim to be experts on criminal justice ethics, desert, free will and moral responsibility, philosophy of mind, and the cognitive sciences all at once. For many of us it is a daunting enough task to achieve the expertise required to contribute to one of these fields while holding sufficiently informed opinions on the others. One obvious solution is division of labour and collaboration, as in this volume.

The flip side of the solution is that the anthology includes specialised contributions on many topics across the breadth of the debate. Such diversity means that there is something for everyone, but perhaps not everything for anyone. Even so, the anthology collects a group of interesting and useful contributions, many of which will undoubtedly help to move the debate forward in the years to come.

References

Berofsky, B. (2011). Compatibilism Without Frankfurt: Dispositional Analyses of Free Will. In R. Kane (Ed.), *Handbook of Free Will, 2nd Ed.*

Bode, S., He, A. H., Soon, C. S., Trampel, R., Turner, R., & Haynes, J.-D. (2011). Tracking the Unconscious Generation of Free Decisions Using Ultra-High Field fMRI. *PLoS ONE, 6*(6), e21612.

Boonin, D. (2008). *The Problem of Punishment*. New York: Cambridge University Press.

Burns, J. H., & Swerdlow, R. H. (2003). Right Orbitofrontal Tumor with Pedophilia Symptom and Constructional Aprxia Sign. *Archives of Neurology, 60*, 437.

Clarke, R. (2009). Dispositions, Abilities to Act, and Free Will: The New Dispositionalism. *Mind, 118*(470), 323-351.

Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. Oxford: Oxford University Press.

Dennett, D. (2003). *Freedom Evolves*. London: Allen Lane The Penguin Press.

Dolinko, D. (1991). Some thoughts about retributivism. *Ethics, 101*(3), 537-559.

Dolinko, D. (1992). Three Mistakes of Retributivism. *UCLA Law Review, 39*(6), 1623-1658.

Dolinko, D. (1997). Retributivism, consequentialism, and the intrinsic goodness of punishment. *Law and Philosophy, 16*(5), 507-528.

Duff, A. (2001). *Punishment, Communication, and Community*. Oxford: Oxford University Press.

Fischer, J. M. (2006). Punishment and desert: A reply to Dolinko. *Ethics, 117*(1), 109-118.

Fischer, J. M., & Ravizza, M. (2000). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Greene, J., & Cohen, J. (2004). For the Law, Neuroscience Changes Nothing and Everything. *Philosophical Transactions of the Royal Society of London, 359*, 1775-1785.

Haynes, J.-D. (2012). Beyond Libet: Long-term Prediction of Free Choices from Neuroimaging Signals. In W. Sinnot-Armstrong & L. Nadel (Eds.), *Conscous Will and Responsibility: A Tribute to Benjamin Libet* (pp. 85-96). Oxford: Oxford University Press.

Heyman, G. M. (2009). *Addiction: A Disorder of Choice*: Harvard University Press.

Levy, N. (2011). *Hard Luck*. Oxford: Oxford University Press.

Levy, N., & McKenna, M. (2009). Recent Work on Free Will and Responsibility. *Philosophy Compass, 4*(1), 96-133.

Libet, B. W. (1999). Do we have free will? *Journal of Consciousness Studies, 6*, 47-57.

Libet, B. W., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain, 106 (Pt 3)*, 623-642.

McKenna, M. (Ed.) (2009) Stanford Encyclopedia of Philosophy.

Mele, A. R. (2009). *Effective Intentions: The Power of Conscious Will*: Oxford University Press.

Moore, M. S. (1997). *Placing blame: a general theory of the criminal law*: Oxford University Press, Incorporated.

Morse, S. J. (2004). New Neuroscience, Old Problems. In B. Garland (Ed.), *Neuroscience and the Law*. New York: Dana Press.

Morse, S. J. (2007). The Non-Problem of Free Will in Forensic Psychiatry and Psychology. *Behavioral Sciences and the Law, 25*, 203-220.

Morse, S. J. (2010). Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note. In M. J. Farah (Ed.), *Neuroethics - An Introduction with Readings*. Cambridge, MA: The MIT Press.

Morse, S. J. (2011). Lost in Translation? An Essay on Law and Neuroscience. In M. Freeman (Ed.), *Law and Neuroscience*. Oxford: Oxford University Press.

Morse, S. J. (2013). Common Criminal Law Compatibilism. In N. Vincent (Ed.), *Neuroscience and Legal Responsibility*. Oxford: Oxford University Press.

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

Pereboom, D. (2007). Hard Incompatibilism *Four Views on Free Will*. Oxford: Blackwell Publishing.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. [10.1038/nn.2112]. *Nat Neurosci, 11*(5), 543-545.

Strawson, P. (2008). Freedom and Resentment *Freedom and Resentment and Other Essays*. Abingdon: Routledge.

Van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Oxford University Press.

Vihvelin, K. (2004). Free Will Demystified: A Dispositional Account. *Philosophical Topics, 32*(1/2), 427-450.

Von Hirsch, A., & Ashworth, A. (2005). *Proportionate Sentencing*. Oxford: Oxford University Press.