

# Responsibility and punishment: whose mind? A response

**Oliver R. Goodenough**

*Vermont Law School, Chelsea Street, South Royalton, VT 05091, USA (ogoodenough@vermontlaw.edu)*

Cognitive neuroscience is challenging the Anglo-American approach to criminal responsibility. Critiques, in this issue and elsewhere, are pointing out the deeply flawed psychological assumptions underlying the legal tests for mental incapacity. The critiques themselves, however, may be flawed in looking, as the tests do, at the psychology of the offender. Introducing the strategic structure of punishment into the analysis leads us to consider the psychology of the punisher as the critical locus of cognition informing the responsibility rules. Such an approach both helps to make sense of the counterfactual assumptions about offender psychology embodied in the law and provides a possible explanation for the human conviction of the existence of free will, at least in others.

**Keywords:** punishment; insanity defence; criminal law; neuroscience; game theory

## 1. INTRODUCTION

The essays collected in this issue demonstrate how the discoveries of cognitive neuroscience are rapidly expanding our understanding of the workings of the human brain. As our models of human mental processes improve, they are beginning to inform debates in other fields. In the law, the rules for assessing criminal responsibility have become the object of such an examination, both in essays within this issue (Baird & Fugelsang 2004; Greene & Cohen 2004; Jones 2004; Sapolsky 2004) and elsewhere (e.g. Lewis 1998; Reider 1998; Winslade 2002; see also Morse 2004).

Although these critiques are persuasive in discrediting their target, i.e. the psychological model of 'free will' that informs the legal tests for responsibility, they are directed at the wrong locus of cognition. Although the legal tests are phrased in terms of the psychology of the person to be punished, I believe that the critical psychology is that of the punisher. The law of responsibility makes much more sense if it is looked at from the strategic position of an agent assessing whether to inflict punishment on a transgressor in a context of social interaction. In this brief essay I will sketch an alternative frame of analysis based on this starting point, arguing that, to be optimally effective, a potential punisher will take a committed position, that the agent standing in threat of punishment has a capacity for choice about action that we might well describe as 'free will', and that the potential punisher will be persuaded out of this position only in the face of overwhelming evidence. The results of such an approach may be subject to criticism, but they are not illogical. Furthermore, this approach provides a possible explanation for the human conviction that free will exists. Finally, for those who make the distinction between a 'positive' analysis and a 'normative' one, this essay may be viewed as a positive account of the psychology underlying commonly held normative views. It must leave

to another occasion a discussion of the useful scope of that distinction itself.

## 2. THE LAW OF RESPONSIBILITY

In the Anglo-American legal tradition, one of the predicates for criminal punishment is a showing that the accused meets a test for being able to act responsibly. The failure to meet that requirement is a possible defence to a criminal prosecution. In the United States, this test has come in five principal 'flavours': the *M'Naghten* rule, the 'irresistible impulse' test, the *Durham* standard, the American Law Institute Model Penal Code definition and the federal statutory definition of insanity (Reider 1998; Dressler 2001; Sapolsky 2004). There are two basic components to the test: a cognitive requirement and a volitional requirement. The cognitive component focuses on whether the offender had the capacity to understand the wrongful and/or unlawful nature of the criminal act. The volitional component asks whether or not the offender had the ability to control whether or not he committed the criminal act. The variations in the five flavours revolve largely around the degree to which each of the two requirements is taken into account and the severity of the deficit necessary to provide a legal excuse.

No matter which test is invoked, the US courts have taken a relatively parsimonious approach to accepting the defence, although there have been, over the years, a few widely publicized exceptions to this parsimony; exceptions that often lead to a backlash and a return to even greater parsimony (Reider 1998; Dressler 2001). Generally, only people suffering from extreme and obvious deficits are able successfully to invoke the defence; and often not even them (Lewis 1998). Even success is not a 'get-out-of-jail free' card—the alternative is often a long period of civil commitment for mental illness.

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

### 3. CRITIQUING THE LAW OF RESPONSIBILITY: A PERSONAL ADDITION

The psychology of transgression underlying these tests is subject to challenge, and the picture of human thought emerging from the new neuroscience is strengthening these objections. In this issue, Greene & Cohen (2004) argue that the ideas of 'free will' and blame that bolster the traditional approach are fundamentally untenable, and that the law should shift from a retributive model of punishment to a forward-looking, consequentialist one, more interested in effective prevention than in assessing blame. Consequentialists, they argue, will 'hold people responsible for crimes simply because doing so has, on balance, beneficial effects through deterrence, containment, etc.' (Greene & Cohen 2004). Sapolsky (2004) takes a similar view, reinforcing it with a detailed examination of the role of the prefrontal cortex in decision-making. In particular, he focuses on problems with the *M'Naghten* test, which makes a lack of cognitive awareness of the nature of the action the critical point and largely devalues the volitional aspect. Baird & Fugelsang (2004) look at limitations in the ability of adolescents to fully consider the consequences of their actions, limitations based both in experience and developmental neurobiology. Reider (1998) called for a new test for the insanity defence, promoting an approach that would incorporate the discoveries of neuroscience into moral and legal theory.

In her book *Guilty by reason of insanity*, Lewis (1998) has offered a particularly telling critique of the law of responsibility and insanity, told with a passion and power born of a direct involvement with specific cases. Focusing on death-row inmates, she identifies a widely shared profile of organic brain injury, abuse as a child, and the denial of a loving, nurturing relationship with a parent or other caregiver. The cumulative effect of all of these insults is to remove layer after layer in the systems of desire, control and inhibition that keep most of us from committing capital crimes.

I say most of us, but I have only good luck and the fast reactions of my best friend in high school to thank for not being a statistic of conviction myself. Loss of control is not something that happens only to some distinguishable other. As an 18-year-old, I came close to injuring my best friend seriously, maybe even killing him, in a brief moment of furious rage. We were walking back from a squash game; we often played together. He had beaten me, and was not letting me forget it. I had had a bad day in other contexts, although nothing extraordinary. For some reason his teasing was too much for me. I just 'lost it'. I turned, raised my racquet, and swung it down at his head with all my strength. Luckily, his reactions were quicker than mine, here as well as on the court. He got his racquet up just in time to block my attack. My rage passed almost instantaneously. He swore at me and asked what I thought I was doing. I was unable to really answer; I had not been 'thinking' in any sense that he meant. After calling me an idiot, he kept on walking with me. We stayed friends; the experience has remained vivid in my memory.

As a subjective matter, I do not think any exercise of will would have stopped me from making that attack on my friend. I am not normally homicidal or violent; this person was my best friend at school; I was fully aware of the penalties for intentionally maiming or killing someone. For

whatever reason, at that short moment, I was essentially undeterable. Some form of very hard determinism was at work. In Sapolsky's analysis, my prefrontal cortex had disappeared of the picture. I was, briefly, in the state of volitional free fall that must afflict Lewis' subjects every day.

Suppose I had connected with my friend's skull and killed him? Should I have been punished by the law? I have described my experience and posed this question to a variety of academic listeners, most of whom, responding with good intuitive promptness, say yes, of course. A few of the more thoughtful apply the legal tests and agree. Only a very few say no, you had no responsibility. What is going on?

### 4. TURNING THE PERSPECTIVE INSIDE-OUT: THE BRAIN OF A PUNISHER

The law of criminal responsibility begins to make sense if we turn it inside out. Although the legal test for incapacity is phrased in terms of the psychology of the transgressor (and in terms that fall apart under the sophisticated scrutiny suggested in this issue by Sapolsky (2004) and Greene & Cohen (2004)), it is really a proxy for a theory of mind test by the punisher: does the transgressor fall within the class of agents on whom the strategic threat of punishment might have an effect? If so, then the punisher will maximize the effectiveness of the threat of punishment, by making both a personal and a public commitment to the strategic presumption that the transgressor is free to choose a course of behaviour in the face of such a threat—i.e. has a form of free will. To understand the law, and its arguably counterfactual psychology of responsibility, we need to look at different brains—the brains of the punishers.

Humans can be viewed as relatively competent strategic actors. In game theory, a strategic actor is one who can take the probable actions and reactions of a different actor into account as he/she plans his/her own course of conduct. Game theory describes what happens when two or more strategic actors are paired in an interaction under a variety of conditions and constraints (see, for example, von Neumann & Morgenstern 1944; Camerer 2003, 2004; Dixit & Skeath 2004). A key prerequisite to success in strategic thinking is to know what you are dealing with—another strategic actor with whom a strategic game can be played and whose mind and intentions can be 'read' according to the experience available to the strategic actor, or, alternatively, some kind of mechanistic object or process. The nature of the response may vary depending on this knowledge (Sanfey *et al.* 2003).

The ability to distinguish another mentally competent actor is often called a 'theory of mind' (e.g. Baron-Cohen *et al.* 1993; Frith & Frith 2003), and it has been suggested that theory of mind capacities are linked to the function of recognizing a strategic partner (Coricelli *et al.* 2000). History, anecdote and science suggest that the capacity may be over-applied, and that humans may err on the side of attributing strategic agency to non-strategic phenomena. The ancient Greeks saw lightning, heard thunder, and postulated Zeus as an explanation. Most of us have felt the urge to slap a computer that has frozen with a key document unsaved, or to kick a recalcitrant soda machine that has eaten our money and dispensed nothing; many of us have succumbed. Researchers both note the tendency for computer users to personify their machines (Reeves & Nass

1996) and work with that tendency to create psychologically effective interfaces (e.g. Dryer 1999). This apparent over-application suggests that the costs of a false attribution of strategic agency have been less onerous than those related to missing a strategic agent that in fact exists.

Punishment is a well-examined strategic move (Binmore 1998; Dixit & Skeath 2004), and one that is a deeply human trait. Whatever our longing for a utopia where the lion and the lamb lie down together in non-coerced peace and harmony, punishment for the transgression of norms is an important element in maintaining relatively cooperative human societies. Its effectiveness, and in some cases necessity, has been demonstrated both theoretically and as a matter of experimental study (Binmore 1998; Fehr & Gächter 2002; Fehr & Fischbacher 2004a). Punishment by a third party is particularly effective at stabilizing cooperative social structures (Bendor & Swistak 2001; Fehr & Fischbacher 2004b). The apparent ubiquity of punishment as an element in normative systems gives 'common sense' support to the importance of its role.

To be effective, threats of punishment must involve commitment. 'Tying yourself to a rule, which you would not want to follow if you were completely free to act at a later time, is an essential part of this process' (Dixit & Skeath 2004, p. 231). As parents discover, an empty threat is no threat at all. (A corollary of this is that threats need to be chosen carefully.) At its most extreme, this commitment can be assured by creating a mechanistic 'doomsday' device (a 'grim reaper' strategy), where a terrible consequence is simply unavoidable if the targeted default occurs (id). The adolescent game of 'chicken' involving driving two cars at each other and seeing who turns away first can often be won by the person who observably ties the wheel straight ahead and jumps into the back seat. Of course, if both jump, there is no winner. A possible corollary of this is that an actor will benefit strategically by forcefully and publicly committing him/herself to the proposition that the other driver cannot make the jump, and therefore will always be at the wheel. The convinced projection of a 'free will' model of choice on a potential offender as part of a punishment rule can be seen as a similar move to maximize the effect of a punishment threat.

Punishment by a third party is typically not without cost, however; it often requires the 'altruistic' bearing of a cost without any direct material gain (Fehr & Gächter 2002). It will be to the punisher's advantage not to waste punishment on those for whom it could never act as a deterrent.<sup>1</sup> This utilitarian logic is recognized in legal scholarship as one of the justifications for the insanity defence (Dressler 2001). In this, as in other strategic contexts where the outcome for an actor is influenced by the behaviour and choices of another and vice versa, it is costly to waste punishment on the truly un-influencible. In trust and ultimatum game experiments, the differential neurological and behavioural reactions to defections by a known computer as opposed to by a perceived human agent (Coricelli *et al.* 2000; Sanfey *et al.* 2003) suggest that human brains sort the world on precisely this kind of basis.

As discussed above in the context of the game of chicken, there appears to be a psychological tilt toward imputing 'free will' agency in doubtful cases; the countervailing costs of mistaken punishment, however, should constrain the extent of the tilt. This constraint can be gamed in return by

a competent actor *faking* incapacity. Whereas most defendants claiming lack of mental capacity probably have at least some basis for their claim, Mafia don Vincent 'The Chin' Gigante notoriously wandered around New York's Greenwich Village in his pyjamas, muttering to himself, in what a federal court determined to be a calculated attempt to forestall prosecution (Tyre 1997). Furthermore, it is a rare defendant who is not at least somewhat intimidated by threats of legal punishment; almost anyone, in fact, walking the streets has some kind of partial control over impulsive, illegal actions (the psychology of the transgressor is of interest in this context). In a cooperative game where both sides play within these expectations, an agent threatening punishment will seek to avoid being deceived about incapacity and will work to make the other agent as suggestible as possible.

I believe that these goals lead to a commitment and to a bias in human psychology. *The commitment* is to treating the other agent as if he/she had the capacity to fully integrate the threat of punishment into its decision-making calculus, and to act accordingly, i.e. as if he/she had a kind of free will. Declaring this committed position both neutralizes attempts at deception by the transgressor and to some degree forces the role of a considering agent on the other player. However counterfactual the free will proposition may be in a deterministic world (Greene & Cohen 2004), it is a strategic fiction that underlies the productivity of a punishment rule, and is a fiction that may be deeply lodged in human cognitive and emotional psychology.<sup>2</sup> Our free will intuitions may be false in the world of deterministic science and yet nonetheless effective in the world of strategic interaction.

*The bias* will be against surrendering this commitment, even in the face of evidence to the contrary. The transgressor will seek to invoke the 'don't-waste-punishment' proposition, claiming, in the often-repeated words of children to their parents, 'I couldn't help it'. In fact, in many instances where punishment is inflicted, the transgressor indeed cannot help it—exactly the point made by Greene and Cohen, by Sapolsky, and by my own lapse into attempted manslaughter. Even in a fully accurate criminal system, those being punished will be either the inattentive or the undeterred, and many of the undeterred will be those who, for whatever reason, were, at the time of the criminal action, effectively the undeterable. Relaxing the bias against being persuaded of this fact, however, appears likely to lead to an increase in deception and to a decrease in attention to consequence, i.e. a decrease in deterability. In either case, the effectiveness of the punishment scheme would begin to unravel, a consequence described by Dixit & Skeath (2004) for a professor willing to listen to student excuses about paper deadlines. Only in the clearest, most common-sense recognized cases of near-total incapacity will the bias be overridden.

Holmes adopted this conclusion in his treatment of responsibility in criminal law:

[The tests for liability] take no account of incapacities, unless the weakness is so marked as to fall into well known exceptions, such as infancy or madness. They assume that every man is as able as every other to behave as they command.

(Holmes 1963)

Sadly, the efficacy of a punishment system may rest on a willingness to punish many people who really could not

help it. For better or for worse, the Anglo-American approach to the law of responsibility, parsimonious in definition and in application, is consistent with both the commitment and the bias suggested here.

### 5. POSSIBILITIES FOR REFORM?

What do we make of a system like this? From an individual fairness standpoint, it does cry out for reform. Holmes, with his customary moral coolness (Alschuler 2000; Hoffman 2004), suggested that the convicted awaiting capital punishment should be viewed as dying for the good of society, not unlike soldiers on the battlefield (Holmes 1963). I am sceptical as to whether such an approach would provide much comfort to those on death row, and it is intuitively disturbing for society when honestly faced. This may be an issue where judgement at a macro-level sometimes reverses at the micro-level. Perhaps the insights of neuroscience about the development of reasoning in adolescents (Baird & Fugelsang 2004) will help to reinforce and reinstate rules prescribing less harsh treatment for offenders who are minors. Youth is a marker, which folk psychology, popular acceptance and the law have all recognized, at various points and in various degrees, as a reliable, unlikely to be faked, impairment of competence (Robinson & Darley 1995). Perhaps the increased understanding that neuroscience can provide about profound, but non-obvious, mental illness (Lewis 1998; Sapolsky 2004) will come to replace the common-sense estimations of mental illness that are rooted in our shared theory of mind capabilities. Greene and Cohen foresee such a change. However, any such reforms will face resistance from deeply rooted human psychology, based in the strategic logic of punishment: the psychology of the punisher.

### 6. CONCLUSIONS

The critiques in this issue and elsewhere, based in cognitive neuroscience, of the Anglo-American approach to criminal responsibility are correct in pointing out the deeply flawed psychological assumptions underlying the legal tests. The critiques themselves, however, may be flawed in looking, as the tests do, at the psychology of the offender. Introducing the strategic structure of the punishment decision into the analysis leads us to consider the psychology of the punisher as the critical locus of cognition informing the rule. Such an approach both helps make sense of the counterfactual assumptions about offender psychology embodied in the law and provides a possible explanation for the human conviction of the existence of free will, at least in others.

The author is grateful to the Gruter Institute for Law and Behavioral Research and to the Vermont Law School for support in this research, to Paul Zak, Morris Hoffman and Semir Zeki for comments in the draft stage, and to Sarah Sun Beale and the International Society for the Reform of Criminal Law for the opportunity and impetus to develop the approach described here.

### ENDNOTES

<sup>1</sup> Binmore (1998) suggests that there is also an incentive for the participants in a social order to seek light punishments, growing out of the possibility that each may end up, if only through inattention, on the wrong side of the law.

<sup>2</sup> Fehr & Gächter (2002) have argued that emotional engagement is an important proximate mechanism for initiating altruistic punishment. See also Sanfey *et al.* (2003).

### REFERENCES

- Alschuler, A. W. 2000 *Law without values: the life, work and legacy of Justice Holmes*. University of Chicago Press.
- Baird, A. A. & Fugelsang, J. A. 2004 The emergence of consequential thought: evidence from neuroscience. *Phil. Trans. R. Soc. Lond. B* **359**, 1797–1804. (doi:10.1098/rstb.2004.1549)
- Baron-Cohen, S., Tager-Flusberg, H. & Cohen, D. J. 1993 *Understanding other minds*. Oxford University Press.
- Bendor, J. & Swistak, P. 2001 The evolution of norms. *Am. J. Sociol.* **106**, 1493–1547.
- Binmore, K. 1998 *Just playing: game theory and the social contract*. Cambridge, MA: MIT Press.
- Camerer, C. F. 2003 Behavioral studies of strategic thinking in games. *Trends Cogn. Sci.* **7**, 225–231.
- Camerer, C. F. 2004 Behavioral game theory: prediction human behavior in strategic situations. In *Advances in behavioral economics* (ed. C. F. Camerer, G. Loewenstein & M. Rabin), pp. 374–392. Princeton University Press.
- Coricelli, G., McCabe, K. & Smith, V. 2000 Theory-of-mind mechanism in personal exchange. In *Affective minds* (ed. G. Hatano, N. Okada & H. Tanabe), pp. 249–259. Amsterdam: Elsevier.
- Dixit, A. & Skeath, S. 2004 *Games of strategy*, 2nd edn. New York: W. W. Norton.
- Dressler, J. 2001 *Understanding criminal law*. New York: Lexis.
- Dryer, D. C. 1999 Getting personal with computers: how to design personality for agents. *Appl. Artif. Intell.* **13**, 273–295.
- Fehr, E. & Fischbacher, U. 2004a Social norms and human cooperation *Trends Cogn. Sci.* **8**, 185–190.
- Fehr, E. & Fischbacher, U. 2004b Third party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87. Available at [www.iew.unizh.ch/wp/iewwp106.pdf](http://www.iew.unizh.ch/wp/iewwp106.pdf).
- Fehr, E. & Gächter, S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140.
- Frith, U. & Frith, C. D. 2003 Development and neurophysiology of mentality. *Phil. Trans. R. Soc. B* **358**, 459–473. (doi:10.1098/rstb.2002.1218)
- Greene, J. & Cohen, J. 2004 For the law, neuroscience changes nothing and everything. *Phil. Trans. R. Soc. Lond. B* **359**, 1775–1785. (doi:10.1098/rstb.2004.1546)
- Hoffman, M. B. 2004 The neuroeconomic path of the law. *Phil. Trans. R. Soc. Lond. B* **359**, 1667–1676. (doi:10.1098/rstb.2004.1540)
- Holmes, O. W. 1963 *The Common law*. Boston: Little Brown & Co. [Re-edited edition published 1963, original edition published 1881.]
- Jones, O. D. 2004 Law, evolution and the brain: applications and open questions. *Phil. Trans. R. Soc. Lond. B* **359**, 1697–1707. (doi:10.1098/rstb.2004.1543)
- Lewis, D. O. 1998 *Guilty by reason of insanity: a psychiatrist explores the minds of killers*. New York: Fawcett.
- 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.
- Reeves, B. & Nass, C. 1996 *The media equation: how people treat computers, television, and new media like real people and places*. New York: Cambridge University Press.
- Reider, L. 1998 Toward a new test for the insanity defense: incorporating the discoveries of neuroscience into moral and legal theories. *UCLA Law Rev.* **46**, 289–342.

- Robinson, P. H. & Darley, J. M. 1995 *Justice, liability, and blame*. Boulder, CO: Westview Press.
- 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.
- Sapolsky, R. M. 2004 The frontal cortex and the criminal justice system. *Phil. Trans. R. Soc. Lond. B* **359**, 1787–1796. (doi:10.1098/rstb.2004.1547)
- Tyre, P. 1997 Mob bosses took a beating last year. *CNN Interactive U.S. News Story Page*, 3 January 1997, available at [www.cnn.com/US/9701/03/mobster.wrap](http://www.cnn.com/US/9701/03/mobster.wrap).
- von Neumann, J. & Morgenstern, O. 1944 *Theory of games and economic behavior*. Princeton University Press.
- Winslade, W. J. 2002 Traumatic brain injury and legal responsibility. In *Neuroethics: mapping the field* (ed. S. Marcus), pp. 74–82. New York: Dana Press.