CBESS Discussion Paper 08-01

Is there a distinction between morality and convention?

by Robert Sugden

University of East Anglia

Abstract

In Humean analyses of the emergence and stability of moral rules, ideas of justice and reciprocity originate in non-moral, conventional solutions to conflicts of interest in human interaction. This theory seems contrary to an empirical claim made by some developmental psychologists: that, from early childhood, human beings perceive a distinction between (universal) 'moral' and (relative) 'conventional' rules, and that moral rules apply to matters of welfare, fairness and trust. I review the psychological literature and argue that, properly understood, it is compatible with a Humean analysis of morality.

Acknowledgements This is a revised version of a paper presented at a conference on Norms and Values at ZiF (Zentrum für interdisziplinäre Forschung), Universität Bielefeld, in May 2008. I thank conference participants for their comments. My work was supported by the Economic and Social Research Council of the UK (award no. RES 051 27 0146).

University of East Anglia

CBESS

Centre for Behavioural and Experimental Social Science

Is there a distinction between morality and convention?

Robert Sugden

School of Economics
University of East Anglia
Norwich NR4 7TJ
England
r.sugden@uea.ac.uk

12 May 2008

Abstract (provisional) In Humean analyses of the emergence and stability of moral rules, ideas of justice and reciprocity originate in non-moral, conventional solutions to conflicts of interest in human interaction. This theory seems contrary to an empirical claim made by some developmental psychologists: that, from early childhood, human beings perceive a distinction between (universal) 'moral' and (relative) 'conventional' rules, and that moral rules apply to matters of welfare, fairness and trust. I review the psychological literature and argue that, properly understood, it is compatible with a Humean analysis of morality.

Acknowledgements This is a revised version of a paper presented at a conference on Norms and Values at ZiF (Zentrum für interdisziplinäre Forschung), Universität Bielefeld, in May 2008. I thank conference participants for their comments. My work was supported by the Economic and Social Research Council of the UK (award no. RES 051 27 0146).

I first read David Hume's *Treatise of Human Nature*, with its analysis of justice as an 'artificial virtue', about twenty-five years ago. I immediately felt that I was learning deep truths about the social world and about the human sense of justice. Ever since then, I have retained the conviction that Hume's analysis is essentially correct, and I have used it as a starting point in my attempts to model the emergence and stability of ideas of justice and reciprocity. Recently, however, I have come across findings in developmental psychology which, if taken at face value, seem to cast doubt on the Humean analysis. Conversely, if that analysis is correct, standard interpretations of the psychological evidence may be in error.

In Hume's account, human beings have natural sentiments of sympathy. These sentiments are directly responsible for the 'natural virtue' of benevolence, but are only indirectly implicated in the sense of justice. Justice originates in conventional solutions to conflicts of interest in human interaction. Initially, each person's motivation to act in accordance with such a convention is merely self-interest. Once a conventional practice is in place, however, mechanisms of natural sympathy come into play. Their effect is to induce, among those people who participate in the practice, a sense of general approval for actions which accord with the convention and of disapproval for actions which do not.

My book *The Economics of Rights, Cooperation and Welfare* (*ERCW*) develops a theory of the emergence of social norms which integrates Hume's ideas with Adam Smith's account of fellow-feeling and with the twentieth-century game theory of Thomas Schelling in economics, David Lewis in philosophy and John Maynard Smith in biology. *ERCW* was first published in 1986. In the second edition, published in 2004, I reviewed its arguments in the light of subsequent developments in economics, philosophy and psychology, and concluded that they remained fundamentally sound. But shortly after completing the second edition, my confidence was somewhat shaken. I discovered that there is a literature in developmental psychology which treats 'morality' and 'social convention' as distinct domains of reasoning and judgement.

According to this literature, moral and conventional rules are categorically different. Moral rules deal with issues of welfare, fairness and trust. These rules are perceived as objectively valid, independently of particular social practices and sources of authority (for children, typically parents and teachers). Conventional rules specify how interactions are to be structured within a given social system. These rules are perceived as socially contingent; the wrongness of violating them is removed if persons in authority give their permission, or if one moves to a social system in which they are not operative. There is a well-developed

experimental protocol, the *moral/ conventional distinction task*, for testing whether a subject can recognise this distinction, and this is standardly used to assess psychological development. Normal children can recognise the moral/ conventional distinction from about the age of three or four. For young children, typical examples of moral transgressions include hitting another child, stealing another child's property, and breaking promises; typical examples of conventional transgressions include talking in class without first raising one's hand, undressing in the playground, and going into toilets designated for the opposite sex. However, there is a significant exception to this finding: children with psychopathic tendencies have difficulty in making the moral/ conventional distinction, as do psychopathic adults.

From a Humean perspective, the whole idea of a moral/ conventional distinction is problematic. One might expect *natural* virtues to be supported by rules with the characteristics that have been attributed to the moral domain, but many of the rules that supposedly fall on the moral side of the moral/ conventional distinction (for example, rules against stealing and promise-breaking) are, in a Humean analysis, paradigm cases of convention. Viewed through the lens of this analysis, conflicts over the use of physical goods (which supposedly belong to the domain of morality) are not different in kind from conflicts over who speaks and who listens (which supposedly belong to the domain of convention). In most societies, each of these conflicts is resolved by rules whose precise specifications are arbitrary; these rules work because, in a given society, most people understand them in the same way and accept the normative obligations they impose. Such rules are not objective moral truths that are independent of social contingencies. But neither are they the commands of some constituted authority that is empowered to waive them at will.

If it is true that a moral/ conventional distinction is salient for children, it is tempting to interpret that fact as evidence of moral naïveté. Perhaps, as the Humean analysis maintains, principles of fairness and trust *are* fundamentally similar to conventional rules such as 'drive on the left', but this truth of social theory is too subtle for children (and probably most adults) to recognise. But for we Humeans, the anomalous behaviour of psychopaths gives pause for thought. It would be disturbing to have to conclude that psychopaths have a better understanding of the nature of morality than psychologically normal people do.

In any case, a naturalistic theory of morality should be capable of explaining how moral concepts are learned. It is clear that the moral/ conventional distinction task has revealed empirical regularities in people's perceptions of social and moral rules. A reader of *ERCW* is entitled to ask whether these regularities are consistent with the hypotheses advanced in that book.

Among psychologists who discuss the moral/ conventional distinction task, there is disagreement about how far the distinction between morality and convention is innate. Some commentators interpret the evidence from this task as supporting the very general hypothesis that human beings are innately equipped with the capacity to reason morally, and that moral reasoning uses distinct systems or 'modules' of the mind. This idea has been put forward particularly forcefully by Marc Hauser (2006) in a book whose title encapsulates his central claim: *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. If Hauser is right, there is a further problem for the Humean account of justice. According to that account (and as I shall explain in more detail later), rules of justice can begin as unintended regularities in human behaviour. Gradually, these regularities evolve into conventions and then into moral rules. The first stage in the emergence of a convention occurs when an individual begins to recognise that other people's behaviour shows some pattern, whether intended or not, and that his self-interest is served by aligning his own behaviour with that pattern. At this stage, many of the mechanisms of rule-recognition and rule-following may be common to the learning and following of straightforwardly prudential rules. And even when conventions have become moralised as rules of justice, there is considerable overlap between the dictates of morality and of prudence (as expressed by the proverb that honesty is the best policy). It is difficult to reconcile this analysis with the hypothesis that the human mind uses different processes for moral and non-moral reasoning.

My aim in this paper is to investigate the moral/ conventional distinction, as discussed by developmental psychologists, in relation to the Humean theory of justice I present in *ERCW*.

1. A Humean model of the emergence of rules of justice

In *Leviathan*, Thomas Hobbes imagines a state of nature in which men are permanently at war with one another. One of the forms that this warfare takes is that

> ... if any two men desire the same thing, which nevertheless they cannot both enjoy, they become enemies; and in the way to their end, which is principally their own conservation, and sometimes their delectation only, endeavour to destroy, or subdue one another. (1651, Ch. 13)

Hobbes's conclusion is that the only way to escape this state of war is by the creation of a 'common power' which can force individuals to respect property rights. But one might think that this kind of conflict could be resolved without needing to call on an external enforcer. That is what Hume claims. After discussing the problems that are caused by the lack of property rights in 'external goods', Hume argues that the solution is 'a convention enter'd into by all the members of the society to bestow stability on the possession of those external goods'. But:

> This convention is not of the nature of a *promise*: For even promises themselves, as we shall see afterwards, arise from human conventions. It is only a general sense of common interest; which sense all the members of the society express to one another, and which induces them to regulate their conduct by certain rules. I observe, that it will be for my interest to leave another in the possession of his goods, *provided* he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common interest is mutually express'd, and is known to both, it produces a suitable resolution and behaviour. ... Nor is the rule concerning stability of possession the less deriv'd from human conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. (489-490)

In *ERCW* I develop a family of models of Hobbes's case of the two men who desire the same thing, and show how conventions of the kind described by Hume can emerge spontaneously in from recurrent human interaction. Much of my analysis is adapted from models used by theoretical biologists to investigate the behaviour of animals of the same species which come into conflict over resources, such as nesting sites or mating opportunities. I now describe the simplest of these models.[1]

Consider a large human population in which pairs of individuals recurrently come into conflict over valuable resources. Each individual faces such conflicts many times, but against different opponents. Each such interaction is represented by the *Hawk–Dove game* shown in Table 1; payoffs can be interpreted as measures of individual self-interest. In each game there are two roles. One player, the *Possessor*, is in possession of the resource when the interaction begins; the other, the *Challenger*, is not. Each player has two alternative pure strategies. The *Hawk* strategy is to act with increasing aggression until the other contestant backs down or until one contestant is seriously injured; the *Dove* strategy is to back down at

the first sign of aggression by the other contestant. V is the value of the resource and D is the cost of injury to the loser of a fight; $D > V > 0$, implying that *Hawk* is the best response to *Dove* and vice versa. The symmetry of the payoff matrix represents the assumption that the asymmetry between *Possessor* and *Challenger* is uncorrelated with payoffs or fighting ability; it is assumed that if both contestants choose *Hawk*, each is equally likely to win the fight, and that if both choose *Dove*, each has an equal chance of getting the resource without having to fight.

Table 1: the Hawk-Dove game

		Challenger	
		Dove	*Hawk*
Possessor	*Dove*	$V/2, V/2$	$0, V$
	Hawk	$V, 0$	$(V-D)/2, (V-D)/2$

$D > V > 0$

The asymmetry between the roles of *Possessor* and *Challenger* is arbitrary as far as payoffs are concerned, and one might think that it could have no significance for how the game is played. But, as I now show, that would be a mistake.

First, suppose that no one is conscious of the asymmetry between the two roles, and that individuals learn which strategy is better for them by trial and error. Then the only equilibrium is one in which, in the population as a whole, *Hawk* is played with probability V/D. (There cannot be an equilibrium in which *Hawk* is always played, since every individual would do better by unilaterally deviating to *Dove*; and vice versa. So equilibrium requires a mix of *Hawk* and *Dove* play, such that each strategy has the same expected payoff. This implies a probability of V/D for *Hawk*.)

But now suppose instead that individuals recognise the asymmetry between the roles and entertain the possibility that behaviour might be role-dependent. In this case, the game has two additional and stable equilibria – one in which everyone follows the rule 'If *Possessor*, play *Hawk*; if *Challenger*, play *Dove*', and one in which everyone follows the opposite rule 'If *Possessor*, play *Dove*; if *Challenger*, play *Hawk*'. These equilibria can be

interpreted as de facto property rules for resolving conflicts over resources. Each rule is arbitrary; but each has the property that if everyone can expect everyone else to follow it, everyone has an incentive to follow it himself. Such a rule is a *convention*. Further, and crucially, the mixed-strategy equilibrium is now unstable. There can be a mixed-strategy equilibrium only if *Hawk* is played with probability *V*/*D*, not only by individuals in the aggregate, but also by *Possessors* and *Challengers* separately. If, for any reason, the probability of *Hawk* rises above *V*/*D* for, say, *Possessors* and falls below *V*/*D* for *Challengers*, then every individual has an incentive to play *Dove* as *Challenger* and *Hawk* as *Possessor*. Such deviations from the mixed-strategy equilibrium are self-reinforcing. The implication is that recurrent play of the game can be expected to lead to the eventual emergence of one or other of the conventions.

It is essential to Hume's argument that conventions are prior to the sense of justice: 'Without such a convention [of property], no one wou'd ever have dream'd, that there was such a virtue as justice, or have been induc'd to conform his actions to it' (498). This is the point of his distinction between natural and artificial virtues. A natural virtue such as sympathy induces actions whose value is immediately evident to the actor, independently of social context. (Consider a mother who hears her baby crying with hunger, and feeds him. If the mother is motivated by natural sympathy, her sense of the baby's contentment on being fed is an immediate emotional reward, and confirms the value of her action. She does not need to think about what other mothers do.) But the rules of justice are too arbitrary for individual actions to have this kind of direct emotional feedback. Indeed, as Hume points out, individual acts of justice, considered in isolation, often have adverse effects on general welfare (his example is a beneficent man who repays a large debt to a miser [497]). The moral value of justice can be perceived only if justice is understood as a set of rules that are followed *generally*. But if that is so, the rules must be in operation *before* they are perceived as morally obligatory. According to Hume, the emergence of rules of justice is to be explained by 'interest', as it is in the Hawk–Dove model. The question of why we 'annex the idea of virtue to justice' requires a separate analysis.

In *ERCW* I present such an analysis. This is not exactly that of Hume, since it draws on Smith's later analysis of fellow-feeling, but it is Humean in spirit.[2] The essential idea is that human beings tend to feel resentment when they are conscious of being harmed by other people's unexpected actions. Resentment is a negative emotion which is directed against another person; it compounds disappointment that an expectation has been frustrated with

anger towards the person who has frustrated it. I postulate that this is a primitive emotion, prior to any sense of moral entitlement. I suggest that the evolutionary value of resentment might be as a cue for aggression in situations in which this is likely to pay. Think of the Hawk–Dove game. Suppose the convention 'If *Possessor*, play *Hawk*; if *Challenger*, play *Dove*' has become established. Then individuals must be motivated to act accordingly – to be aggressive in the role of *Possessor* and submissive in the role of *Challenger*. In other words, the perception of being *Possessor* must cue a desire to fight against any *Challenger* who tries to take the disputed resource. Is there a general emotional mechanism that would generate such a specialised desire, irrespective of the nature of the dispute and of the convention that is being violated? One obvious answer is a mechanism which responds to the frustration of an expectation by directing aggression towards the person who has frustrated it.

Some philosophers have criticised this part of *ERCW* on the grounds that the emotion of resentment has moral content: for you to feel resentment against another person, they argue, you must be conscious that he has *wronged* you, or at least that he has deliberately and knowingly frustrated your expectations. I can accept that these factors tend to increase the anger and aggression of resentment, but (for the reasons I have already given) I maintain that resentment does not require them. In any event, it should be clear why a Humean analysis needs an assumption of this kind. The aim is to show that the *prior* existence of a property convention *subsequently* induces a moral sentiment of justice. If a moral sentiment is to emerge, it must surely do so as the transformation of some prior emotion; and to require that prior emotion to be itself moral would lead to an infinite regress.

If my argument so far is correct, a property convention will tend to be associated with regularities in people's emotional repertoires: people will predictably feel resentment when they are harmed by other people's deviations from that convention. This is where Smith's mechanism of fellow-feeling takes effect. Smith's hypothesis – one that has been confirmed by recent psychological and neurological research – is that emotions are contagious.[3] If person A's action in breach of a convention provokes person B's resentment, then A's consciousness of B's hostility towards him will tend to induce a negative affective response in A. Further, if person C observes A's action and B's response, she can be expected to experience some reflection of B's resentment. (The reader might ask why C sympathises with B's resentment rather than with A's aggression. If C normally conforms to the convention, she will find it easier to imagine herself in B's position than in A's. Again, it is

necessary to assume the prior existence of the convention in order to explain the emotions it induces.) Thus, A's breach of the convention will have a general tendency to induce negative emotions directed at A, which in turn will induce unease in A. And this, according to Smith, is just what morality *is*. Morality is a system of rules specifying the propriety of emotions. To judge that an emotion has propriety is to approve it as appropriate to the circumstances in which it is experienced. And, as Smith puts it: 'To approve of the passions of another ... as suitable to their objects, is the same thing as to observe that we entirely sympathise with them' (1759/ 1976: 16). Morality is just the generalisation of fellow-feeling.

Notice how, in this account, the sense of justice is an emergent property of a social process; it is not a property of the individual human mind. Notice too that in the earlier stages of this process, the relevant interactions are between individuals engaging in non-moral reasoning. Because of this, the rules of justice inherit properties of non-moral conventions. In particular, they inherit whatever properties favoured the emergence of the conventions from which they have grown. For example, suppose there is some small asymmetry in the Hawk–Dove payoffs, so that the role of *Possessor* confers a marginal advantage in a fight. This will favour the emergence of the convention 'If *Possessor*, play *Hawk*; if *Challenger*, play *Dove*' rather than its opposite.[4] Once the convention is fully established, this asymmetry in fighting ability is irrelevant, since fights do not occur. Nevertheless, its effects are imprinted on the convention and hence on the sense of justice that is induced: people come to feel that possession is associated with moral entitlement.

This sense of moral entitlement may be generalised beyond the cases in which following the convention is prescribed by self-interest. For example, if the Hawk–Dove payoffs vary across games, an individual may sometimes find himself as *Challenger* in situations in which he can take the resource at minimal cost to himself (say, because the *Possessor* is unusually weak). Although self-interest dictates *Hawk* in this non-standard case, this motivation may be inhibited by the perception that *Challengers* who play *Hawk* are generally the object of disapproval – and that he himself disapproves of such behaviour on the part of others. This inhibition is a form of fellow-feeling. In this sense, natural sympathy plays a part in the moralisation of rules of justice. But it remains true that, in standard cases, the actions prescribed by justice are also in the self-interest of the person to whom they are prescribed, given that other people can be expected to follow the convention.

Thus, even when the sense of justice has emerged, it has a belt-and-braces relationship with prudential reasoning.

A somewhat similar analysis can be made of principles of reciprocity, such as promise-keeping, mutual aid, and playing one's part in the provision of public goods. In these cases, it is necessary to analyse *repeated* interactions among individuals who can recognise one another and can remember who did what in previous interactions. In small groups of individuals who interact relatively frequently, conventions of reciprocity are likely to emerge. Mechanisms of resentment and fellow-feeling then induce a corresponding system of interrelated sentiments of approval and disapproval. This is what we (as members of such groups) perceive as the morality of fairness and reciprocity. Thus, in the small groups in which practices of reciprocity first emerge, reciprocity is supported by motivations of both interest and morality.[5] These practices may then spread to larger groups of potential cooperators, or persist as initially small groups increase in size. Large-group cooperation usually requires institutionalised systems of incentives, for which the morality of reciprocity is merely a back-up[6]; but occasionally moral sentiments alone may provide sufficient motivation.

So, if the Humean analysis of justice and reciprocity is correct, there is no discontinuity between the domains of prudential and moral reasoning. There seems to be no role for mental modules dedicated to the processing of moral reasoning. We do not need moral minds, only prudence and fellow-feeling.

2. The moral/ conventional distinction

The moral/ conventional distinction, as a concept in developmental psychology, seems to have been first proposed by Elliot Turiel; other major contributors to the topic (and sometime collaborators with Turiel) include Larry Nucci and Judith Smetana. My principal sources are Turiel et al (1987), Smetana (1993) and Nucci (2001). It seems that the distinction was discovered when developmental psychologists investigated a stage-based theory of moral development proposed by Jean Piaget (1932) and Lawrence Kohlberg (1969). According to that theory, young children (up to the age of about ten) treat all rules as stemming from the commands of authority, and so have no sense of morality as an autonomous domain. However, the moral/ conventional distinction task revealed that

children as young as three or four were capable of distinguishing between authority and morality.

Smetana gives the following summary of the moral/ conventional distinction, as it is understood by its proponents. She begins with two prototypical scenarios of events in a daycare centre for pre-school children. In 'Event A', Lisa, Michael and David are rocking in a rocking boat and Jenny is waiting for her turn. As the rocking boat slows down, Jenny (presumably out of frustration at waiting) bites Lisa. In 'Event B', the children are being taken to a pool. Jason has forgotten his bathing costume and is asked to choose one from a communal stock. He insists he wants to wear a pink costume even though the teachers tell him that the pink ones are for girls. Smetana says:

> Event A is an example of a *moral* transgression. Moral rules pertain to issues such as others' welfare (or harm), trust, or the fair distribution of resources. ... Moral knowledge is thought to be constructed from the intrinsic consequences of acts for persons. Because moral events have consequences for others' rights and welfare, moral rules are hypothesized to be obligatory, non-alterable, and applicable across situations, and the wrongness of moral acts [i.e. transgressions of moral rules] is thought to be non-contingent on specific social rules or authority dictates... In contrast, Event B is an example of a *social convention*. Conventional knowledge is constructed from an understanding of the social system and refers to the arbitrary and consensually agreed-upon behavioural uniformities that structure social interactions within social systems... In contrast to moral rules, conventional rules are hypothesized to be contextually relative, arbitrary and changeable, and the wrongness of conventional acts is hypothesized to be contingent on the rules and dictates of authority. Morality and social convention are hypothesized to be distinct types of social knowledge that develop in parallel out of different types of social interaction (112-113).

Smetana uses a three-way classification of social rules, in which the third type is *prudential* (for example: do not touch the electricity outlets, do not jump from a moving swing); but most work has focussed on the distinction between moral and conventional rules.

Notice that Smetana is classifying rules along two conceptually independent dimensions. One dimension concerns their *content*. Here the distinction is between rules that deal with welfare, fairness and trust and rules that deal with 'arbitrary and consensually agreed-upon behavioural uniformities'. I take it that Smetana is defining the first class of rules as 'moral' and the second as 'conventional', and I will follow this practice when discussing the moral/conventional distinction. The second distinction concerns the *structural properties* of rules. Rules in one class, which I shall call *non-contingent*, are perceived by individuals as 'obligatory, non-alterable, and applicable across situations'.

Rules in the other class, which I shall call *socially contingent*, are perceived as 'contextually relative, arbitrary and changeable'.

Smetana is proposing two empirical hypotheses. The weaker hypothesis is that, within any given social group, psychologically normal individuals recognise a distinction between non-contingent and socially contingent rules, and their categorisations of specific rules as one or the other are in broad agreement. The stronger hypothesis is that the content-based distinction between moral and conventional maps onto the structural distinction between non-contingent and socially contingent: moral rules are non-contingent while conventional rules are socially contingent. This mapping is hypothesised to hold across all societies.

The most common mode of investigating the moral/ conventional distinction is to take some act which transgresses a rule, to describe this to subjects, and then to ask subjects to express certain judgements about it. In the case in which the subject is a child or adolescent and the act occurs at school, typical questions have the form:

Q1. Would it be all right to commit the act if there were no school rule against it?

Q2. Would it be all right to commit the act at home, or in a different school?

Q3. Would it be permissible to change the rule?

Q4. Would it be all right to commit the act if the teacher gave permission?

Q5. How serious is the transgression?

Q6. How much punishment does the transgressor deserve?

The usual finding is that moral rules (that is, rules that are classified by investigators as being concerned with welfare, fairness or trust) elicit 'No' answers to Q1–Q4, while conventional rules (that is, rules intended to secure coordination in social organisation) elicit 'Yes' answers to the same questions. Q5 and Q6 elicit a less sharp distinction. There is some tendency for moral transgressions to be judged more serious and more deserving of punishment than conventional transgressions, but there are some conventional rules, in particular ones concerned with decency, sexual behaviour and gender roles, that subjects construe as socially contingent (i.e. answer 'Yes' to Q1–Q4) while treating transgressions as serious.

Another mode of investigation is to ask subjects to justify the judgements expressed in answers to questions such as Q1–Q6. Moral rules are typically justified by appeals to

fairness and the welfare of others, while conventional rules are justified in terms of the commands of authority, the threat of punishment, social expectations or the need for social coordination.

The proponents of the moral/conventional distinction point to a very large number of investigations which have found the regularities summarised in the preceding two paragraphs. There have been several studies of American children who have had devout religious upbringings – as Catholics, Mennonites or Conservative or Orthodox Jews. These children generally differentiate between the standard 'moral' rules (not hitting, not stealing, and so on) and the specific rules prescribed by their religion; they perceive the former as universally obligatory, but are more likely to judge the latter to be alterable by religious authorities and obligatory only within the relevant community. Although most studies have used Western populations, a considerable number of studies have been made of children and adolescents in other cultures, often with similar results. However, some investigations have produced apparently conflicting findings. It seems that subjects in many non-Western cultures have a stronger tendency to justify rules by appeal to customs and traditions, and to perceive those rules as non-alterable. This tendency is more marked among groups of relatively low socio-economic status (perhaps because those groups have had less exposure to 'Westernising' influences). In some cultures (for example, in Korea) subjects are more likely to offer justifications which appeal to social status, social roles and courtesy.[7]

Turiel et al (1987: 172-174) summarise the findings of this line of research by presenting lists of rules that have been classified as non-contingent or socially contingent through use of the moral/ conventional distinction task, and by claiming that this classification separates moral from conventional rules. The list of non-contingent (and moral) transgressions includes hitting, name-calling, stealing, destroying another's property, breaking a promise, and not taking turns. The much longer list of socially contingent (and conventional) transgressions includes chewing gum in class, undressing in the playground, calling a teacher by her first name, dressing casually in a business office, public nudity, swearing and cross-gender dressing.

Psychopaths and individuals diagnosed as having antisocial personality disorders give anomalous responses to the moral/ conventional distinction task.[8] Their judgements of the relative seriousness of moral and conventional transgressions are similar to those of normal populations, but when justifying why moral transgressions are bad they are much less likely to refer to the effects of the transgression on the victim, and much less likely to

judge that moral transgressions would remain wrong if rules prohibiting them were removed. James Blair, Derek Mitchell and Karina Blair (2005: 59), in an authoritative discussion of psychopathy, treat this anomaly as an 'impairment in moral reasoning': even when adult, psychopaths are unable to perform a reasoning task that is within the capacity of pre-school children. It seems that psychopaths can understand social rules, but lack a normal understanding of morality.

Blair et al relate this phenomenon to other known characteristics of psychopaths. Psychopaths have impaired ability to recognise and to respond empathetically to other people's sadness or fear (but have normal responses to anger, happiness and surprise). They also show abnormally weak emotional responses to anticipations of threats and punishments. Since the imagination of one's own future emotions may use similar mental processes as the imagination of other people's emotions, this is further evidence of a failure of empathy. Psychopaths are also deficient in the capacity to understand situations which, for normal individuals, would induce guilt, even though they have a normal understanding of embarrassment. Presumably guilt is more dependent than is embarrassment on the perception of others' sadness and fear. Blair et al argue that empathetic responses to others' sadness and fear is crucial for socialisation. In the playground, for example, a normal child is conditioned not to make unprovoked attacks on other children by responding empathetically to the emotional reactions of children who are attacked. I take Blair et al to be suggesting that this form of aversive conditioning is implicated in children's perception that rules prohibiting harm are independent of context and authority.

3. Is the moral/ conventional distinction conventional?

According to the standard interpretation of the moral/ conventional distinction, the distinction between rules that are perceived as universally obligatory and rules that are perceived as socially contingent maps onto the distinction between rules that pertain to welfare, fairness and trust and rules that do not. I am not fully convinced about the reliability of this mapping.

In part, my scepticism arises from my work as a social theorist of morality. The methodology of social theory is very different from that of developmental psychology, and each approach focuses on some explanatory mechanisms at the expense of others. I acknowledge that social theory in the tradition of Hume and Smith, Schelling and Lewis,

uses simplifying psychological assumptions that specialists in the field will see as naïve. But, in compensation, it has a much richer analysis of the mechanisms by which social conventions and moral rules emerge and become self-sustaining as unintended consequences of human interaction. In the light of this analysis, the idea that rules of welfare, fairness and trust can be defined independently of convention seems equally naïve.

A further ground for scepticism is the reflection that welfare, fairness and trust are central components of *liberal* morality. This is a morality that is probably accepted by most developmental psychologists and perhaps also by most teachers in the schools that have been studied. In some work on the moral/ conventional distinction, there seems to be a political subtext of secular liberalism, viewed in opposition to the politics of identity. Thus, Nucci (2001: 50-51) presents his findings in support of a proposal that moral education in the (American) public school system should cover those elements of substantive morality that are 'central concerns' of all religious and ethical systems, without 'hiding behind a smoke screen of value relativism', but should exclude all teaching of 'particular doctrinal values'. I cannot help feeling that it is too neat, too suggestive of wishful thinking, to suppose that a universal property of developmental psychology maps on to one's own particular moral code.

In my (admittedly superficial) reading of the relevant literature, I have come across various discussions which articulate, and provide supporting evidence for, my initial reservations about the objectivity of the moral/ conventional distinction. I now turn to these.

A number of writers have raised doubts about the hypothesis that context-independence and authority-independence are attributed to, and only to, rules concerned with welfare, fairness and trust. Rules of other kinds have sometimes been found to be perceived as universally obligatory. For example, Jonathan Haidt, S. Koller and M. Dias (1993) have found that actions that evoke feelings of disgust tend to be perceived as universally prohibited. Examples include incest (even under conditions in which there is no risk of procreation), sexual intercourse with dead animals, and (for Americans) cutting up the American flag and using the pieces as rags to clean the bathroom. The flag example is interesting because of the way it evokes ideas of sacredness and pollution in a particular Western population. As Mary Douglas (1966) has argued, the symbolism of purity and pollution seems to be universal to human societies, even though different societies have different systems of symbols. Richard Shweder, Manamohan Mahapatra and Joan Miller (1987) investigate the moral/ conventional distinction in a population of Brahman and

Untouchable individuals in Bhubaneswar, an old temple town in Orissa. They find that Brahmans perceive many rules of purity and pollution as universal, and treat transgressions of these as much more serious than everyday cases of hurting, stealing or breach of trust. For example, Brahman children perceive as universally obligatory the rules that women must not cook during their menstrual periods (children know about women's recurring periods of 'uncleanness', even though they do not know about menstrual bleeding) and that women must change clothes between defecation and cooking.

It seems that in just about all societies, *gratuitous* hurting of others is perceived as non-contingently wrong. That is surely not surprising: it is hard to see how any social organisation could function effectively if all its members were allowed to assault one another at will and without reason. And since sadness and fear are strongly susceptible to emotional contagion, one would expect a general, cross-cultural tendency for hurting to be disapproved of, *other things being equal*. But once one goes beyond trivial cases, one finds many cultures and social organisations in which specific kinds of physical assault are permitted in specific circumstances. One finds too that the rules that delimit legitimate assault are socially contingent.

The Brahman children studied by Haidt et al saw no transgression in a scenario in which a husband beats his wife for repeatedly going alone to movies without his permission, or in one in which a father canes a son who repeatedly truants from school. Up to the nineteenth century, flogging was routinely used as a punishment for adult workers on naval ships and on slave plantations, and in many Western countries, harsh physical punishments of children and adolescents continued to be considered normal well into the twentieth century. Within certain boundaries, violence between children was accepted too. For example, I went to primary school in the 1950s in a predominantly working-class area of northern England. If one child reported to a teacher that, in the playground, he or she had been hit by a child of the same age and sex, the standard response was: 'Hit him (or her) back'. I now think that this attitude expressed concepts of honour and respect. A child who had to appeal to a teacher to deal with a dispute with another child was showing himself to be 'soft', and would lose the respect of his fellow-children. By teaching children to stand up for themselves, teachers were inculcating what were generally seen as morally appropriate standards of behaviour. My guess is that, in this culture, most children judged 'hitting back' as universally permissible (and perhaps even obligatory), and not something whose rightness

was dependent on authority. Similarly, perhaps, the men of Bhubaneswar who beat their wives for disobedience perceived themselves as upholding their respect as husbands.

It is also interesting to look at the issue of socially accepted violence from the opposite side. In societies in which most forms of physical violence are no longer seen as acceptable, how do people think about practices of violence that are accepted in other societies, or were accepted at other times? This question has been investigated by Daniel Kelly, Stephen Stich, Serena Eng and Daniel Fessler (2007), using a web-based questionnaire with adult (and mostly American) participants. Respondents were asked to consider various scenarios of non-gratuitous violence. In one example, an officer on a ship finds a subordinate sailor drunk on watch and punishes him with five lashes with a whip. In one version, the event took place three hundred years ago; in another, it takes place on an American cargo ship in 2004. In another example, a sergeant in charge of training American commandos subjects them to serious physical abuse in simulated interrogations, to prepare them to deal with interrogation by enemy forces. In one version, this practice is not prohibited by army regulations and is approved by the sergeant's superiors. In another, the Pentagon has just issued new regulations prohibiting this previously common practice. Respondents were much more likely to say that the violent act was 'OK' in the first version of each scenario (that is, in which violence is socially accepted or approved by authority) than in the second (in which it is not). The implication seems to be that, for adult Western subjects, prohibitions on non-gratuitous violence are context- and authority-dependent.

I was particularly struck by a critique of the moral/ conventional distinction by Carolyin Pope Edwards (1987), based on fieldwork among the Oyugi people of Kenya, carried out in the 1970s. Edwards argues that, from an early age, Oyugi children have a strong sense of the wrongness of transgressing rules which structure social interaction – rules of politeness, respect, etiquette, the division of labour and the performance of work tasks. They perceive these rules as just as obligatory as rules concerning care for others and the control of aggression. Edwards explains this difference between Oyugi and American children as the result of differences in their experiences of economic and social organisation:

> The African children described in this study live in large, rural households that are economic as well as social units; children are given responsibilities at an early age that may increase their identification with social rules and help them appreciate their value. In such a cultural context, the distinction between the obligatory/ interpersonal and organizational/ regulatory domains may be less prominent than

> in American classrooms where 'school rules' so obviously come 'from outside'. (126)
>
> The 'good purposes' or 'reasons for' many kinds of rules can be made comprehensible to young children. Just as a child who receives a hit or kick can directly experience the purpose of rules prohibiting aggression, so too a child who experiences a delay in her supper because her older sibling did not collect the firewood can see for herself the inherent 'rightness' of rules about obedience to parents. Similarly, just as a child who cares for an infant can apprehend the purpose of moral rules prescribing nurturance, so too a child who tries to control an unruly, unhygienic, and unmannerly toddler can construct the inherent need for cleanliness and etiquette standards. (139)

Edwards's argument, as I understand it, is that breaches of social conventions typically *are* harmful to others, but the mechanisms which generate this harm are not always transparent. A child who has been brought up in a small nuclear family and who has not taken part in cooperative productive labour has an impoverished understanding of the implications of 'conventional' transgressions. It is only because the modern child's realm of experience is so restricted that the moral/ conventional distinction is so salient for her.

The proponents of the moral/ conventional distinction recognise some of the problems exhibited by examples such as these, and try to deal with them by using the idea of 'rule overlap' – the idea that a rule or action may be *both* moral *and* conventional. Smetana (1993: 119), following Turiel (1983), describes three kind of overlap.

The first type of overlap occurs when 'conventional concerns for social organization entail injustices (such as in a caste system)'. The implication seems to be that there can be arbitrary rules of social organisation and stratification that are in some sense functional, while also being unjust. But what is meant by saying that, say, the rules of a caste system are unjust? If the claim that the caste system is unjust is made from the external viewpoint of a theorist, then that seems an irrelevant interjection: we are supposed to be investigating people's perceptions of rules, not moralising about them. Alternatively, if people in the caste-based society *perceive* these rules to be unjust, then (according to the proponents of the moral/ conventional distinction) they should also perceive that injustice to be independent of authority and social contingency; and so 'conventional concerns for social organisation' should have no force in legitimising it. (Compare the case of the teacher who says that it is all right to steal.)

The second type of overlap concerns 'second-order events in which a violation of a convention results in psychological or physical harm to others adhering to the convention';

Smetana's example is driving on the left when it is conventional to drive on the right. This example illustrates perfectly why, from a Humean perspective, the moral/ conventional distinction is problematic. Whenever a convention of justice or reciprocity is in operation, an individual who deviates from it will tend to harm others. That is not an exceptional case in which harm and convention happen to overlap; it is a characteristic property of rules of justice and reciprocity. Of course, the harm caused by unilateral deviations is not always immediately obvious. For example, it may be difficult for young children to understand the purpose of a rule requiring them to ask the teacher's permission before speaking in class; such a rule may be *perceived* as an arbitrary regulation, imposing a pointless form of social order. It requires some degree of moral maturity to understand how conventions work. (Recall Edwards's interpretation of the differences between American and Oyugi children: according to this account, the Oyugi children are more mature.)

The third type of overlap consists of what Smetana calls 'ambiguous multidimensional events, where individuals make different domain attributions about the same event'. I take her to mean that there can be rules that some people perceive as moral (that is, as pertaining to welfare, fairness or trust) but which others perceive as merely conventional. Smetana suggests that debates about abortion and homosexuality provide 'examples of such multifaceted issues'. Turiel et al (1987: 212-215) use this concept of multidimensionality to try to counter Shweder et al's interpretations of the Bhubaneswar data. Discussing the scenario in which a husband beats his disobedient wife, Turiel et al say:

> It is likely that Indian subjects view the maintenance of familial role differentiations as a matter of great importance, permitting the use of corporal punishment to ensure its enforcement. ... [The] finding that Indians did not view the act as a transgression suggests that they judged the moral feature of physical harm as subordinate to the social-organizational consideration of authority and maintenance of sex-role differentiations.

Well, yes, but how does that support the hypothesis of a moral/ conventional distinction? Turiel et al seem to be saying that, in this case, the perception of a moral prohibition against harm is overruled by the perception that wife-beating is a socially approved practice which maintains the family as a form of social organisation. But that should not happen if moral prohibitions are perceived as independent of social contingencies.

Turiel et al introduce a different form of argument when they discuss the scenario of the father who canes the son. In this case, they suggest, respondents who see nothing wrong in the father's action may be evaluating the event 'on the basis of naive theories of child-

rearing and discipline practices', according to which physical punishment is in the long-term interest of the child, and so not really a case of harm at all. Similarly, they offer instrumental, welfare-based rationalisations for such Brahman customs as the rule that widows should not eat fish. Here the claim is that we are dealing with an 'unearthly-belief-mediated moral event': the widow who eats fish is believed to be offending the spirit of her dead husband and so causing harm (207).[9] I have to say that these explanatory manoeuvres read like ad hoc attempts to reformulate a hypothesis so as to avoid disconfirming evidence. (I wonder what instrumental reasons Turiel and Nucci would find to rationalise the rule against using the American flag to clean the toilet.)

In a more general manoeuvre of retreat, Turiel et al claim that their position is misrepresented by critics who assume that the distinction between moral and conventional rules is 'objective'. The correct reading of their position, they say, is that the concepts of welfare, fairness and trust which define the moral domain have to be interpreted subjectively, with 'consideration to context, circumstances, or interpretations and meanings given to them by individuals' (205-206). But if one takes this line, the moral/ conventional distinction starts to dissolve. We still have a distinction between rules that are perceived as non-contingent and rules that are perceived as socially contingent. But if the concepts of 'welfare', 'fairness' and 'trust' are themselves subjective and context-dependent, how does a person's knowledge of rules about welfare, fairness and trust differ from her knowledge of consensually agreed-upon behavioural uniformities that structure interactions within social systems? Turiel et al seem to be saying that the central 'moral' concepts of welfare, fairness and trust may, after all, be conventional.

4. Conclusion

So what implications can be drawn from the study of the moral/ conventional distinction? Does this research cause problems for the Humean analysis of the emergence and stability of rules of justice?

I think there can be no doubt that the moral/ conventional distinction task has identified systematic patterns in the human understanding of morality. In particular, it has established that, from a very young age, normal children acquire the subjective perception that some of the rules of behaviour that they learn are 'objectively' or 'universally' moral, while others are socially contingent, or contingent on the commands of authority. The fact

that children can distinguish between these two categories is evidence that they are developing the ability to engage in a simple but significant form of autonomous moral reasoning: they have a concept of moral obligation that is not merely a reproduction of commands from adults.

The evidence from psychopaths strongly suggests that emotional empathy plays a crucial role in our learning of (what we perceive as) non-contingent obligations. In the classroom and playground settings of a modern primary school, it is not surprising that the most salient emotional cues are associated with actions by which one child hurtfully contravenes the expectations of another, for example by hitting her, or taking or damaging her property. Such actions naturally induce immediate and transparent responses of fear and sadness, and those emotions are particularly susceptible to emotional contagion. Thus, normal children gradually internalise rules which prohibit these kinds of harm. At the same time, children are being made aware of various rules which they are expected to follow, but whose function is not so transparent. If breaches of these rules do not seem to hurt anyone very much, the obligation to follow them is not internalised.

However, emotional contagion need not be restricted to actions which cause 'objective' harm. Take the case of disgust. If a child behaves in a way that a parent thinks of as deeply disgusting, the parent's negative emotional response is likely to be easily perceived by the child, and liable to emotional contagion. The point is not that the disgusting action causes offence and that offence is a form of harm (thus belonging to the domain of morality). Rather, disgust – just like the fear felt by someone who has been assaulted – is a negative affective state which can be perceived by empathy. If certain kinds of action reliably induce disgust, and if children are sufficiently exposed to experiences of that regularity and sufficiently insulated from contrary experiences, we should expect rules against those actions to be internalised and to be perceived as universally obligatory.

Relative degrees of insulation may be significant in explaining the apparently conflicting evidence about whether rules that are specific to particular religions are perceived as non-contingent. I conjecture that, in contrast to the Brahmans of Bhubaneswar, even the most devout religious communities in America are insufficiently insulated for their rules to have this property. Thus, the Orthodox Jewish child in America does not learn that transgressions of religious rules induce emotions of disgust; he learns that transgressions of Orthodox rules by Orthodox Jews induce emotions of disgust in Orthodox Jews.

As we get older, and as we get more experience of the diversity of forms of social organisation, we may come to realise that many of what at first sight seem to be arbitrary conventions serve useful functions, and that each of us tends to be harmed when other people unilaterally deviate from such practices. One form that this understanding takes is illustrated by the Oyugi children who understood the function of hierarchical authority structures in organising cooperative work tasks. Another form is illustrated by the adult respondents who recognised that flogging as a punishment on board ships can be morally prohibited in one culture and morally permissible in another.

I see no fundamental tension between what is known about the moral/ conventional distinction and the Humean theory of justice and reciprocity as moralised convention. To the contrary, that distinction is the product of an essentially Humean form of moral learning, in which moral rules are codifications of predictable emotional responses to recurrent stimuli, and are learned by empathic contagion. The psychopath who cannot perceive morality as autonomous, who cannot see the difference between an internalised moral sentiment and an external regulation, is not a person who has a better understanding than the rest of us of the true nature of moral rules. He is someone whose capacities for learning moral rules is impaired.

References

Blair, James, Derek Mitchell and Karina Blair (2005). *The Psychopath: Emotion and the Brain.* Oxford: Blackwell.

Douglas, Mary (1966). *Purity and Danger: An Analysis of Conceptions of Pollution and Taboo*. Routledge and Kegan Paul.

Edwards, Carolyn Pope (1987). Culture and the construction of moral values: a comparative ethnography of moral encounters in two cultural settings. In Jerome Kagan and Sharon Lamb, *The Emergence of Morality in Young Children*, 123-151. University of Chicago Press.

Haidt, Jonathan, Koller, S. and Dias, M. (1993). Affect, culture and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65: 613-628.

Hauser, Marc D. (2006). *Moral Minds: How Nature Designed our Universal Sense of Right and Wrong*. Harper Collins.

Hobbes, Thomas (1651). *Leviathan*. Macmillan, 1962.

Hume, David (1739-40). *A Treatise of HumanNature*. Clarendon Press, 1978.

Kelly, Daniel, Stephen Stich, Serena Eng and Daniel Fessler (2007). Harm, affect, and the moral/ conventional distinction. *Mind and Language* 22: 117–131.

Kohlberg, Lawrence (1969). Stage and sequence: the cognitive-developmental approach to socialization. In David Goslin (ed), Handbook of Socialization Theory and Research, 347-380. New York: Wiley.

Maynard Smith, John and Geoffrey Parker (1976). The logic of asymmetric contests. *Animal Behaviour* 24: 159-175.

Nucci, Larry (2001). *Education in the Moral Domain*. Cambridge University Press.

Piaget, Jean (1932), *The Moral Judgment of the Child*. New York: Free Press, 1965.

Shweder, Richard, Manamohan Mahapatra and Joan Miller (1987). Culture and moral development. In Jerome Kagan and Sharon Lamb, *The Emergence of Morality in Young Children*, 1-83. University of Chicago Press.

Smetana, Judith (1993). Understanding of social rules. In Mark Bennett (ed), *The Child as Psychologist: An Introduction to the Development of Social Cognition*, 111-141. Harvester Wheatsheaf.

Smith, Adam (1759). *The Theory of Moral Sentiments*. Clarendon Press, 1976.

Sugden, Robert (2004). *The Economics of Rights, Co-operation and Welfare*, second edition. Palgrave Macmillan. (First edition published by Basil Blackwell, 1986.)

Sugden, Robert (2005). Fellow-feeling. In Benedetto Gui and Robert Sugden (eds), *Economics and Social Interaction*. Cambridge University Press, 52-75.

Turiel, Elliot (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press.

Turiel, Elliot, Melanie Killen and Charles Helwig (1987). Morality: its structure, functions, and vagaries. In Jerome Kagan and Sharon Lamb, *The Emergence of Morality in Young Children*, 155-243. University of Chicago Press.

Notes

[1] The analysis summarised in the following three paragraphs is presented in more detail in Sugden (2004: 58-107); it is adapted from a the model of asymmetric animal contests presented by Maynard Smith and Parker (1976).

[2] Hume postulates that our moral approval of justice derives from 'sympathy with public interest' (499-500). Smith (1759/ 1976: 85-91) thinks this psychologically implausible, and I agree. The following analysis is presented in more detail in Sugden (2004: 218-223).

[3] For more on Smith's hypothesis and the evidence which supports it, see Sugden (2005).

[4] This mechanism is explained in Sugden (2004: 93-95), which again adapts an earlier argument by Maynard Smith and Parker (1976).

[5] These mechanisms are explained in Sugden (2004: 108-182).

[6] Compare Hume's (1739-40/ 1978: 538-539) account of how two neighbours with a common interest in draining a meadow can achieve their objective by voluntary action, while a thousand people with a similar common interest will fall foul of the free-rider problem.

[7] My source here is Nucci (2001: 11-13, 21-51, 94-97).

[8] This paragraph relies on Blair et al (2005: 47-66, 124-128).

[9] See also Nucci (2001: 102-104), who discusses a similar range of what appear to be counter-examples to his hypothesis, and uses very similar arguments to try to eliminate them.