Northern Ireland Legal Quarterly

Volume 60 Number 3

EDITOR
PROFESSOR SALLY WHEELER

Queen's University Belfast
School of Law

Contents

NILQ 60(3): 279–304

Criminal law as a preventative tool of environmental regulation: compliance versus deterrence

PATRICK BISHOP*

University of Wales, Swansea

ABSTRACT

The effective enforcement of environmental law is an issue which continues to engender considerable academic debate. The conclusions that may be drawn from such debate will have implications not only for the future of environmental law but also the wider regulatory reform agenda. This article commences with several non-contentious propositions. First, any regulatory regime ought to place considerable emphasis on preventing harm; within the context of environmental law, this view is encapsulated by the preventative principle. Secondly, one may be tempted to treat the criminal law as a purely reactive or curative mechanism were it not for the deterrent effect associated with the imposition of criminal sanctions. Therefore, the initial premise is that if environmental law is to become more preventative in scope, policymakers ought to consider how the deterrent effect of environmental criminal law may be bolstered. Academic publications and official reports are replete with assertions that the fines imposed by the courts for environmental crime are lamentably low. A preoccupation with the level of fine may lead one to overlook the fundamental importance of the likelihood of apprehension as an essential ingredient of the deterrence formula. Mainstream criminological discourse provides clear evidence that the probability of apprehension has a greater influence on deterrence than severity of sanction. In this context, the enforcement style adopted by the relevant enforcement agency is crucial; in particular, whether the style adopted augments the perception that apprehension is more probable. This article proceeds to argue that many of the assumptions on which the compliance style is based may be questioned, if not totally undermined. It is contended that a transition from compliance-orientated enforcement towards an approach more focused on deterrence has the potential to enhance the deterrent, and by extension the preventative effect, of environmental criminal law.

* Centre for Environmental and Energy Law and Policy, School of Law, Swansea University, Singleton Park, Swansea SA2 8PP, p.bishop@swansea.ac.uk. The author would like to thank Stuart Macdonald, Karen Morrow and Mark Stallworthy for their advice and comments on an earlier draft of this article. Naturally, any errors or omissions are entirely my own.

Introduction

The effective enforcement of environmental law continues to engender considerable debate.[1] Academic discourse on the subject has been supplemented by the publication of the Hampton[2] and Macrory Reports,[3] which ultimately led to the enactment of the Regulatory Enforcement and Sanctions Act 2008. The underlying premise of these publications is that the use of conventional criminal sanctions alone can be an ineffectual mechanism for ensuring compliance with regulatory law. To a greater or lesser extent, the criticisms levelled at the criminal law are based on the assertion that the imposition of traditional sanctions does not adequately *prevent* regulatory breaches. While Macrory argues convincingly in favour of an enhanced sanctioning "tool kit", the new penalties advocated by the report and contained in the Act are seen as a supplement to (rather than replacement of) conventional criminal sanctions. Thus, while increased flexibility in the range of sanctions can be lauded for attenuating the hitherto binary nature of regulatory enforcement, it seems that the imposition of traditional criminal sanctions will remain an important weapon in the regulator's armoury. It is not suggested here that a greater range of activities which have the potential to damage the environment or cause environmental harm ought to be criminalised;[4] this article will simply seek to explore how the enforcement style adopted by a regulator may be modified so that the potential preventative effect of the criminal law is increased substantially. In essence, it will be argued that a move away from compliance-orientated enforcement strategies to an approach based on deterrence has the potential to alleviate many of the problematic aspects of regulatory non-compliance. This article will make a number of arguments; first, it will be uncontentiously submitted that the imposition of *ex post facto* criminal liability is able to operate in a preventative manner via the concept of deterrence. It will be further argued that simply increasing the severity of sanctions can only ever represent a partial solution to non-compliance. Finally, as a matter of enforcement policy, it is contended that the use of a strategy which places greater emphasis on deterrence (as opposed to compliance) has the potential to bolster the preventative aspects of the criminal law.

The preventative imperative

The preventative principle[5] can be described succinctly as the requirement that environmental harm/pollution be prevented in advance rather than remedied after the event and, as such, is simply an embodiment of the almost universally accepted adage that "prevention is better than cure". The desirability of preventing environmental damage rather than reacting to it is beyond doubt, to the extent that one commentator has described the preventative principle thus: "a beacon for environmental law at both the international

1 See e.g. K Hawkins, *Environment and Enforcement: Regulation and the social definition of pollution* (Oxford: Clarendon Press 1984); K Hawkins and J Thomas (eds), *Enforcing Regulation* (Boston: Kluwer-Nijhoff 1984); B Hutter, *Compliance: Regulation and Environment* (Oxford: Clarendon Press 1997); K Hawkins, *Law as a Last Resort: Prosecution in a regulatory agency* (Oxford: OUP 2002).

2 P Hampton, *Reducing Administrative Burdens: Effective inspection and enforcement* (the Hampton report) (Norwich: HMSO, March 2005).

3 R B Macrory, *Regulatory Justice: Making sanctions effective* (the Macrory report) (London: Better Regulation Executive, November 2006).

4 For an analysis of how some inherently anti-ecological activities are sanctioned by the law, see M Halsey, "Environmental crime: towards an eco-human rights approach" (1997) 8(3) *Current Issues in Criminal Justice* 217.

5 For an analysis of the principle of prevention, see N de Sadeleer, *Environmental Principles: From political slogans to legal rules* (Oxford: OUP 2002), pp. 61–91; for environmental principles in general see, R Macrory (ed.), *Principles of European Environmental Law* (Groningen: Europa Law Publishing 2004); S Tromans, "High talk and low cunning: putting environmental principles into legal practice" (1995) *JPL* 778.

level and national legal orders: a sort of golden rule".[6] On this basis it may be considered that the criminal law and the corresponding imposition of criminal sanctions is a paradigm reactive mechanism. Many environmental offences are inchoate; in such cases there is no requirement that harm has occurred as a condition precedent of criminal liability, for example, one may be guilty of the offence of causing poisonous, noxious or polluting matter to enter controlled waters[7] notwithstanding that the discharge has not actually harmed the river or life within the river.[8] Despite this, the criminal law is responsive in the sense that potential liability is "triggered" only where an environmental offence has been committed. As Hill has asserted: "A criminal prosecution is always a reactive event."[9] Despite this apparent reactive element to the criminal law, it is axiomatic that the imposition of *ex post* liability (criminal or civil) can operate in an *ex ante* manner by having a dissuasive or deterrent effect on potential polluters.[10] The ability of the criminal law to operate preventively has been explicitly recognised by the EU, a perusal of the directive on ship-source pollution and the introduction of penalties for infringements[11] and the proposed directive on the protection of the environment through the criminal law[12] reveals the recurrent use of the word "dissuasive", i.e. it is clearly recognised that the imposition of sufficiently punitive criminal sanctions may operate to prevent environmental crime. In this sense the designation of particular forms of conduct as "criminal" always entails a preventative element.[13] It is manifestly clear that the criminal law has both deterrent and retributive functions; it is equally apparent that both functions are not mutually exclusive, i.e. the imposition of a punitive sanction is capable of deterring others from engaging in the prohibited conduct. This article will proceed to analyse the extent to which regulatory criminal law and regulatory enforcement are capable of functioning in a manner which maximises the deterrent and, by extension, the preventative effect of environmental law.

Environmental criminal law

The terms "environmental criminal law" and "environmental crime" are wide ranging and could feasibly include casual fly-tipping or the consequences of a major industrial accident with almost innumerable substantive and administrative offences situated between these opposing ends of the spectrum.[14] The breadth of environmental crime is problematic for anyone wishing to identify and elucidate general principles. As such, most of the empirical

6 de Sadeleer, *Environmental Principles*, n. 5 above, p. 89. The preventative principle is a guiding principle of EU environmental law by virtue of Art. 174(2) of the EC Treaty.

7 Water Resources Act 1991, s. 85(1).

8 See the *dictum* of Bryant J in *National Rivers Authority* v *Egger* (1992) where it was stated: "One looks at the nature of the discharge and one says, is that discharge *capable* of causing harm to a river, in the sense of causing damage to the uses to which a river might be put; damage to animal, vegetable or other – if there is such other life which might live in a river, or damaging that river aesthetically." Reported in R Burnett-Hall, *Environmental Law* (London: Sweet & Maxwell 1995) pp. 351–4.

9 M Hill, "Prosecuting environmental crime" in W Upton (ed.), *The Changing Role of Environmental Law* (London: United Kingdom Environmental Law Association 2000), p. 29.

10 M Faure (ed.), "Deterrence, insurability, and compensation in environmental liability", vol. 5 of *Torts and Insurance Law* (New York: Springer 2003), pp. 19–20.

11 Directive 2005/35/EC. See also the Council Framework Decision 2005/667/JHA to strengthen the criminal law framework for the law against ship-source pollution.

12 COM (2007) 51 final, available at: http://eurlex.europa.eu/LexUriServ/site/en/com/2007/com2007_0051en01.pdf.

13 By virtue of the Criminal Justice Act 2003, s. 142(b), one of the stated purposes of sentencing is the reduction of crime.

14 See, generally, K Brickey, "Environmental crime at the crossroads: the intersection of environmental and criminal law theory" (1996) 71 *Tulane Law Review* 487.

research in this area has focused on a particular offence or particular type of offender. The focus of the current enquiry is itself relatively broad but is largely limited to corporate offenders under typical pollution control regimes. Nevertheless, this article's central argument, that deterrence-based approaches augment the perception that apprehension is more likely and therefore enhances the deterrent effect of the law, is sufficiently malleable to apply, to a greater or lesser extent, to most environmental crimes.

Environmental offences have proliferated since the inception of modern environmental law and are an integral feature of most "command and control" regulatory systems. Despite the increased use of alternative methods to achieve environmental policy goals, such as incentives and economic instruments, the use of direct regulation[15] is likely to remain as the bedrock of environmental law in England and Wales. As such, an essential ingredient of any regulatory regime is the availability of sanctions for non-compliance.[16] Macrory, in his report on regulatory justice, has recommended, inter alia, increased flexibility in the range of sanctions available to enforcement agencies, some of his recommendations are given effect in the Regulatory Enforcement and Sanctions Act 2008. A common feature of the new sanctions, such as fixed monetary penalties, is that they can be unilaterally imposed by an enforcement agency without recourse to the criminal courts and the corresponding due process safeguards which can make such a course of action time-consuming and expensive.[17] Despite this, a sanction will only be imposed where a violation of a regulatory statute has occurred[18] and thus the new sanctioning tool kit envisaged by Macrory will continue to operate within the context of a criminal law framework. Further, the overall focus of the report is on regulatory sanctions and as such the new mechanisms suggested to ensure compliance are punitive in effect – they impose a form of punishment as opposed to providing incentives.[19]

Even if environmental law were to develop so that emphasis was shifted from command and control regulation to alternative methods (economic instruments for example), then it is submitted that the criminal law would remain of crucial importance. To illustrate by way of example, in England and Wales, a landfill tax was introduced in 1996,[20] one of the stated purposes of the tax (a classic economic instrument) was to encourage alternative methods of waste disposal, such as recycling and composting.[21] One unintended

15 For a persuasive defence of direct regulation, see R Macrory, "Regulating in a risky environment", (2001) 54 *Current Legal Problems* 619.

16 Ibid., at p. 619. For a discussion of alternative methods of enforcement such as the common law, judicial review and human rights legislation, see L Hagger, "Current environmental enforcement issues: some international developments and their implications for the UK" (2000) 2 *Environmental Law Review* 23.

17 The use of such penalties is not without precedent in environmental law. Examples include: fixed penalty notices are utilised in respect of the offence of allowing a dog to defecate on designated land under the Dogs (Fouling of Land) Act 1996, ss 3–4; the Noise Act 1996 (as amended by the Anti-Social Behaviour Act 2003) permits a local authority to issue a fixed penalty notice in relation to specified offences under the Act; the Environmental Protection Act 1990, s. 88, allows the use of fixed penalties in relation to the offence of littering under s. 87.

18 Regulatory Enforcement and Sanctions Act 2008, s. 39(2).

19 The distinction between coercive sanctions and non-coercive incentives can become blurred. See J Brigham and D Brown, "Distinguishing penalties and incentives" (1980) 2(1) *Law and Policy Quarterly* 5.

20 Introduced by the Finance Act 1996: the detailed regime is provided for by the Landfill Tax Regulations 1996, SI 1996/1527; Landfill Tax (Qualifying Materials) Order 1996, SI 1996/1529; and Landfill Tax (Contaminated Land) Order 1996, SI 1996/1528. Landfill tax for active waste is currently set at £18 per tonne with anticipated increases of £3 per annum up to a maximum of £35 per tonne.

21 J Morris, P Phillips and A Read, "The UK Landfill Tax: an evaluation of the first three years" (2000) 2 *Environmental Law Review* 150, pp. 153–6.

consequence of the introduction of the tax was an increase in fly-tipping.[22] In effect, the success of this alternative method of achieving an environmental goal is dependent on the criminalisation of any illegitimate method[23] which can be utilised to circumvent the landfill tax. In summary, the criminal law is of crucial importance and as such, if environmental law is to become more preventative in scope it is essential that policymakers consider how the deterrent effect of the law may be bolstered.

Deterrence

The concept of deterrence[24] is fundamentally premised on the notion that the infliction of a punitive sanction is capable of influencing the future conduct of potential lawbreakers. Sentencing based on deterrence has an instrumental or consequentialist rationale where the paramount objective is the prevention of future crime. A utilitarian formula provides a justification for punishing individuals severely to deter effectively others from criminal acts; in one sense the classic "end justifies the means" argument where proportionality has little relevance.[25] This ability to instigate behavioural change is capable of operating directly, where a natural or legal person is deterred from future violations via the imposition of a sanction and the desire to avoid future sanctions (specific deterrence and incapacitation)[26] or indirectly, where an example is made of specific deviants so as to deter other individuals from future deviance (general deterrence).[27] Two further sub-categories of deterrence may be identified, namely initial deterrence and marginal deterrence. The former concerns the effect of prohibiting conduct previously lawful, while the latter is concerned with the deterrent effect of changes in enforcement policy and/or sanctions in relation to conduct already prohibited. While the simple term "deterrence" is used throughout this article, the central focus is on marginal deterrence. The extent of the deterrent effect associated with the imposition of criminal sanctions is a matter of fierce academic debate. An extreme viewpoint is that the criminal law has no deterrent effect.[28] However, the general consensus would seem to be that punitive sanctions are capable of deterring and thus the debate has largely focused on the degree of any potential deterrent effect, as Ehrlich has commented: "it is perhaps the extent of the deterrent effect of law enforcement, not its mere existence, that is principally at issue".[29]

The question of why people commit crime has proved to be as unavoidable as any other and requires, inter alia, sociological and psychological examination. Criminological discourse on the subject has concentrated, to a large extent, on non-environmental crimes. Whereas

22 M Watson, "Environmental offences: the reality of environmental crime" (2005) 7 *Environmental Law Review* 190, p. 195; S Bell and D McGillivray, *Environmental Law* 6th edn (Oxford: OUP 2006), p. 614; P. Stookes, *A Practical Approach to Environmental Law* (Oxford: OUP 2005), p. 344.

23 S. 33(1) of the Environmental Protection Act 1990 and the Waste Management Licensing Regulations 1994, reg. 1(3) and Sch. 4, create various criminal offences associated with the disposal etc of waste.

24 See, generally, A Ashworth and A von Hirsch (eds), *Principled Sentencing: Theory and policy* 2nd edn (Oxford: Hart Publishing 1998), ch. 2.

25 A Ashworth, "Sentencing" in M Maguire, R Morgan and R Reiner (eds), *The Oxford Handbook of Criminology* 2nd edn (Oxford: OUP 1997), p. 1098.

26 Incapacitation refers to the removal of an individual actor's ability to commit crime, usually via imprisonment. In most environmental crime prosecutions, the defendant will be a corporation and is therefore imprisonment is not an option. However, incapacitation can still occur via the removal of a licence where the activity pursued by the defendant requires prior authorisation.

27 See, generally, K Williams and R Hawkins, "Perceptual research on general deterrence: a critical review" (1986) 20(4) *Law and Society Review* 545.

28 I Ehrlich, "The deterrent effect of criminal law enforcement" (1972) 1 *Journal of Legal Studies* 259, at p. 260.

29 Ibid., at p. 261.

traditional *mens rea* crimes such as theft, burglary and murder are committed by individuals,[30] environmental crimes can be committed by individuals and corporations and it is the latter who are most often prosecuted by enforcement agencies. When an environmental crime is committed by a corporation then the obvious problem from a deterrence perspective is that incarceration is seldom an option. However, while the individual and corporation may both exercise bounded rationality, the inherent difference in character between natural and legal persons arguably enhances the deterrent effect of the criminal law in a corporate context. Many typical *mens rea* crimes are committed by individuals as a result of an impulse which is difficult, if not impossible to control. The heroin addict who steals to maintain a supply of drugs and the archetypal "crime of passion" are just two such examples. Thus, for crimes committed in such circumstances the offender is unlikely to reflect upon the potential consequences of his or her actions – this viewpoint is bolstered to some extent by empirical studies on the deterrent effect of capital punishment.[31] Corporations, on the other hand, are more likely to operate as rational economic calculators, heuristically weighing up the benefits and disadvantages of transgressing the law and are thus more likely to be deterred by the real or perceived consequences of criminal sanctions. In distinguishing individual and corporate criminality in this manner, one needs to avoid oversimplification. First, in comparison to individuals, corporations will generally possess greater resources with which to pay fines and combat adverse publicity. Secondly, not all individual offenders commit crime because of their innate tendencies; career criminals whose crimes are motivated by economic gain clearly exist. Similarly, it cannot automatically be assumed that a corporation will act as a dispassionate economic maximiser; particular firms may well propagate an ethos or personality which tends towards criminal conduct. These exceptions notwithstanding, it is contended that in many instances corporations are more likely to be susceptible to deterrence-based strategies than is the case with individual offenders.

The potential deterrent effect of the criminal law is a subject which requires both theoretical examination and empirical investigation; this article concentrates on the theoretical debate for the following reason. From a methodological perspective, empirical studies in this area are problematic. First, when one attempts to measure the deterrent effect of the criminal law one is faced with measuring an event that has not occurred.[32] Second, if it can be satisfactorily determined that a particular crime is not being committed, a causal relationship between a particular sanction and the paucity of a specific offence is notoriously difficult to establish. One is faced with the difficulty of safely ascribing non-offending to the deterrent effect of a criminal sanction and not to one of the other myriad reasons that inhibit the commission of crime, such as an inherent respect for the law, peer disapproval etc.[33]

30 In certain circumstances corporations are capable of committing crimes with a *mens rea* element. N.B. Corporate Manslaughter and Corporate Homicide Act 2007.

31 In a well-known study conducted by sociologist Thorsten Sellin, in which the correlation of homicide rates in retentionist and abolitionist states was analysed, it was concluded that the existence of the death penalty has no enhanced deterrent effect. See T Sellin, "Homicides in retentionist and abolitionist states" in T Sellin (ed.), *Capital Punishment 135* (New York: Harper & Row 1967). For a more contemporary review of the relevant literature see J Donohue and J Wolfers, "Uses and abuses of empirical evidence in the death penalty debate" (2005) 58 *Stanford Law Review* 791. Due to the methodological difficulties of studies in this area, one needs to treat empirical data with caution, see n. 33 below and accompanying text.

32 H Jacob, "Deterrent effects of formal and informal sanctions" (1980) 2(1) *Law and Policy Quarterly* 61, at p. 61.

33 Ashworth, "Sentencing", n. 25 above, at p. 1098.

From an economic analysis perspective, the deterrent effect of the criminal law was famously expressed by Becker[34] in terms of the relationship between the gain expected from regulatory breach,[35] the severity of sanction and the likelihood of apprehension. Thus, in basic terms, compliance will be ensured where the benefits of failing to comply with the law are exceeded by the costs of apprehension.[36] Thus, the more severe the sanction and the higher the probability of apprehension, the greater the likelihood that an actor will be deterred from regulatory breach. Such a formulation presupposes that the object of deterrence is a rational profit maximiser,[37] a presupposition that could be questioned most readily in the case of individual offenders. In relation to corporate entities, where often the prime objective is the creation of profit, it can more easily be assumed that increases in the economic costs of non-compliance will bolster the deterrent effect of the law.

An obvious response to Becker's analysis would be simply to increase the severity of sanction. From a public policy perspective, such a course of action may be seen as an attractive option; increasing the severity of punishment is a relatively low-cost option when compared with the costs associated with increasing the probability of apprehension, in terms of increases in surveillance, law enforcement activity etc.[38] In essence, the economic costs of increasing the severity of sanction are marginal to the cost of increasing the probability of apprehension.

Increasing the severity of sanction may be seen as a deceptively attractive solution to regulatory non-compliance; however, more draconian penalties per se are far from a panacea. To concentrate extensively on the level of fines or the severity of sanctions in general is to adopt a rather blinkered approach to the question of how to reduce regulatory breach. An integral aspect of the deterrence formula is the probability of apprehension, thus, no matter how severe the sanction, if the probability of apprehension is zero, the cost of non-compliance would also be zero so that an actor would be provided with an economic justification for regulatory breach even where the expected gains from such violations are insignificant. In reality, no matter how remote the probability of apprehension, it will be greater than zero. Nevertheless, to concentrate simply on the severity of sanctions without also considering the probability of apprehension is a short-sighted approach. The importance of the likelihood of apprehension is illustrated by the following example: let us suppose that the expected gain from regulatory non-compliance is £100,000. If the risk of apprehension is 0.5 (an implausibly high probability) then the penalty would have to be raised to £200,000 so that the expected punishment cost equals the anticipated gain.[39] The formula enunciated above is a simplification; the cost of non-compliance cannot be analysed simply in terms of the fine imposed but would also include

34 G Becker, "Crime and punishment: an economic approach" (1968) 76 *Journal of Political Economy* 161.

35 Becker's analysis is concerned with crime in general, however, the approach is capable of utilisation in all contexts, including regulatory non-compliance: "Although the word 'crime' is used in the title to minimize terminological innovations, the analysis is intended to be sufficiently general to cover all violations, not just felonies like murder, robbery, and assault, which receive so much newspaper coverage, but also tax evasion, the so-called white-collar crimes, and traffic and other violations." Ibid., p. 170.

36 This model has been varied by A Ogus and C Abbot, "Sanctions for pollution: do we have the right regime?" (2003) 14(3) *JEL* 283. Becker's formulation is expressed as $U < pD$ where U is the profit from the offending activity, p is the probability of apprehension and D is the cost to the offender resulting from apprehension (at pp. 289–90).

37 J Boswell and R Lee (eds), *Economics, Ethics and the Environment* (London: Cavendish 2002), at p. 1.

38 G Stigler, "The optimum enforcement of laws" (1970) 78 *Journal of Political Economy* 526, at p. 527.

39 J Coffee, "No soul to damn: no body to kick: an unscandalized inquiry into the problem of corporate punishment" (1980) 79 *Michigan Law Review* 386, at p. 389.

costs less amenable to exact quantification, such as loss of reputation[40] and legal representation expenditure. Further, the formula may be adjusted to take into account the fact that some actors may be particularly risk averse.[41] In this example, deterrence is ensured where the cost of non-compliance at least equals the expected gain. Even if the probability of apprehension is extremely low, this magic figure could be achieved by an astronomically high fine, given the costs of increasing the likelihood of apprehension, Posner has postulated that an efficient outcome could be achieved by "a probability arbitrarily close to zero and a fine arbitrarily close to infinity".[42] Such an approach is problematic for two reasons: first, an extremely high fine may discourage socially beneficial activities where there is a risk of accidental violation of the criminal law.[43] A problem further compounded by the existence of strict liability for most environmental offences. Secondly, such high fines will often exceed the solvency of the particular offender.

It has been determined that the two variables of the deterrence formula are severity of sanction and likelihood of apprehension. As such, a detailed consideration of each is required.

Sanctions for environmental crime

Fines

The range of corporate sanctions available to the courts is limited; enforcement agencies are stymied to a considerable extent by the absence of imprisonment as a viable punishment. It is not the purpose of this article to consider in any detail the scope of the new sanctions advocated by Macrory and contained in the Regulatory Enforcement and Sanctions Act 2008: while most of the new sanctioning tools are novel in the sense that they may be imposed administratively without recourse to the courts, they fundamentally remain a form of financial penalty.

Academic literature and official publications alike are replete with assertions that the level of fines imposed by the courts for environmental crimes is insufficient to provide a meaningful deterrent; that is to say, polluters are not paying enough. To make such a claim is rather a trite assertion, as Coffee has opined, much of the literature on corporate punishment in general seems to consist of little more than such a contention.[44] Indeed, within the specific context of environmental protection, it has become somewhat de rigeur to lament the low level of fines imposed by the courts.[45] While increasing the severity of penalty would theoretically increase deterrence, the imposition of more punitive sanctions (increasing the level of fine for a particular regulatory infringement is the most obvious example) can only ever represent a partial solution to non-compliance for several reasons, not least, policymakers will invariably be subject to political constraints. Increasing the maximum fine for fly-tipping to £500,000 is unlikely to be a vote winner! A common explanation for the relatively low fines imposed for environmental crime is based on the fact that most regulatory breaches are subject to the jurisdiction of the magistrates' court or

40 P de Prez, "Beyond judicial sanctions: the negative impact of conviction for environmental offences" (2000) 2 *Environmental Law Review* 11. The effect of damaged reputation/public image will largely depend on the level of competition within the particular market. For example, a company which enjoys a monopoly is unlikely to be affected by adverse publicity.

41 See M Polinsky and S Shavell, "The optimal trade off between probability and magnitude of fines" (1979) 69(5) *American Economic Review* 880.

42 R Posner, "An economic theory of the criminal law" (1985) 85 *Columbia Law Review* 1193, at p. 1206.

43 Ibid.

44 Coffee, "No soul", n. 39 above, at p. 388.

45 R Malcolm, "Prosecuting for environmental crime: does crime pay?" (2002) 14(5) *Environmental Law and Management* 289.

crown court: courts which have little experience of both the technical aspects of many environmental laws and the nature and extent of potential environmental harm.[46] Further, for many environmental offences, the maximum fine which may be imposed by a magistrates' court is £20,000, a sum four times higher than the value of the statutory maximum fine on summary conviction for most other crimes.[47] Thus, magistrates familiar with a maximum of £5000 may hesitate to impose greater fines, even where permitted to do so, in an area in which they have little experience.[48]

If one considers the approach of the higher echelons of the judiciary, the picture is somewhat mixed. In a handful of decisions, the Court of Appeal has concluded that the fine imposed by the first instance court was excessive: in *R* v *Milford Haven Port Authority (The Sea Empress)*,[49] a fine of £4 million imposed by Cardiff Crown Court was reduced to £750,000; in *R* v *Anglian Water Services Ltd*,[50] an initial fine of £200,000 was reduced to £60,000; and in *R* v *Cemex Cement Ltd*,[51] a £400,000 fine was reduced to £50,000. On one level, these decisions may be seen as symptomatic of the lack of clear sentencing guidelines for environmental offences.[52] Whether the approach of the Court of Appeal is unduly lenient is an arguable point; it could be contended that such drastic reductions undermine the message that environmental crime should be taken seriously. On the other hand, the reduced fines remain substantial. What is clear from the above cases is that the courts are influenced by the financial status of the offender.[53] In *R* v *Milford Haven Port Authority*, the Court of Appeal was influenced by the fact that the port authority performed public functions and had limited avenues open to increase revenue.[54] Similarly, in *R* v *Cemex Cement*, the appellant company had made a profit of just £39,000 in 2006. Setting the fine at a level which has serious financial consequences without forcing the defendant company into insolvency may achieve the objectives of specific deterrence but is more questionable in relation to general deterrence.

Setting the fine cognisant of the defendant's financial circumstances implicitly recognises the view that the imposition of substantial financial penalties creates externalities, summed up by the metaphor: "when the corporation catches a cold, someone else sneezes".[55] To punish a company is to also punish a range of stakeholders; the externalities of fines may be justified on the basis that to a greater or lesser extent, stakeholders (for example, shareholders) can at least sometimes be said to have received "unjust enrichment" from the benefits of corporate crime.[56] However, in a world of imperfect competition, the costs of financial penalties will ultimately be transferred to

46 It has been contended that a magistrate may only see an environmental case every seven years: M Grekos, "Environmental fines: all small change" (2004) *JPL* 1330, at p. 1335.

47 *Environmental Crime and the Courts*, House of Commons Environmental Audit Committee, Sixth Report of Session 2003–04, at p. 9.

48 Ibid, at p. 11.

49 [2000] Env LR 632. Noted by M Davies, "Sentencing for environmental offences" (2000) 2 *Environmental Law Review* 195.

50 [2004] Env LR 10.

51 [2007] EWCA Crim 1759.

52 N Parpworth, "Environmental offences: the need for sentencing guidelines in the Crown Court" [2008] *Journal of Planning and Environmental Law* 18

53 By virtue of the Criminal Justice Act 2003, s. 164(3), the courts are required to take into account the financial status of the offender when setting the level of fine.

54 [2000] Env LR 632, at p. 647.

55 Coffee, "No soul", n. 39 above, at p. 401.

56 Ibid.

consumers.[57] A high degree of care needs to be taken when advancing such arguments. Any punishment administered to any type of offender will usually create externalities. For example, it would be nonsensical to argue against punishing a convicted murderer on the basis that there are family members who are financially dependent on him or her. Thus, taken to its logical extreme, the externalities argument could be used to justify never punishing a corporate offender. In respect of environmental regulation, Hawkins has caricatured the constituents of pollution control as "environmentalists" and "business", the former advocate greater restraints on economic activity in the interests of the environment, while the latter prefer a hands-off policy of non-intervention.[58] The implicit role of the courts may thus be seen in terms of achieving a suitable balance between these competing aims in order to "preserve the sometimes fragile balance between the interests of economic activity . . . and public welfare".[59] As such, the courts are fully aware of the socio-economic consequences of corporate punishment[60] and this creates serious difficulties if higher fines are seen as the main gambit in securing regulatory compliance.

While the general consensus of opinion is that fines for environmental crimes are commonly lower than they ought to be, the wisdom of such a view has been questioned by Parpworth et al. on the basis that any interest group is highly likely to regard the punishment administered by the courts to be inadequate where an offence is committed against an interest which the group seeks to promote or protect.[61] As such, a fine which is less than the prosecutor and other interested parties would have liked may be seen as a "corrective" of the prosecutor's sectional view.[62] Such a contention, it is submitted, has only limited force. While one would never advocate that the prosecutor (most often the Environment Agency) should be the ultimate arbiter of whether a particular sentence is adequate, one cannot simply dismiss the Environment Agency's opinion as biased or unbalanced. To look at the issue from a different perspective, it could equally be contended that the magistrates' assessment of the seriousness of an environmental crime will seldom, if ever, be fully informed given the paucity of environmental offences reaching the magistrates' court. In any event, three possible solutions to the problem of low fines have been suggested, namely the creation of a specialist environmental court[63] or regulatory tribunal,[64] increased training for the magistracy and judiciary on environmental issues[65] and the use of minimum fines. All of which present difficulties.

The desirability or otherwise of an environmental court has been mooted for a considerable time. Arguably the most fruitful area of debate in this context is the

57 Coffee, "No soul", n. 39 above, at p. 402.

58 See Hawkins, *Environment and Enforcement*, n. 1 above, at p. 9.

59 Ibid.

60 To claim that the courts are influenced by the socio-economic consequences of their decisions should not automatically be construed as a criticism. For example, in the context of nuisance it has been asserted that the courts pay insufficient regard to such considerations. See S Tromans, "Nuisance – prevention or payment?" (1982) *CLJ* 87.

61 N Parpworth, K Thompson and B Jones, "Environmental offences: utilising civil penalties" (2005) *JPL* 560, at p. 564.

62 Ibid.

63 H Woolf, "Are the judiciary environmentally myopic?" (1992) 4(1) *JEL* 1; R Carnwath, "Environmental enforcement: the need for a specialist court" (1992) *JPL* 799.

64 The Macrory report, n. 3 above, paras 3.54–60.

65 Parpworth, "Environmental offences", n. 52 above, at p. 30.

jurisdiction of any such specialist court or tribunal.[66] The logic underpinning the creation of an environmental court with criminal jurisdiction is simple. If it is accepted that the scarcity of environmental cases heard by magistrates is one reason for the low level of fines, then a court which exclusively dealt with environmental cases would fully appreciate the seriousness of environmental crime and administer a commensurate sentence. In a different regulatory context – health and safety – the issue has been succinctly summed up by Scott Baker J: "it is difficult for judges and magistrates, who only rarely deal with these cases, to have an instinctive feel for the appropriate level of penalty".[67]

Practical and administrative difficulties notwithstanding, the establishment of a new environmental court is not entirely unproblematic. A commonly held perception, often resisted by environmentalists, is that environmental crime is not "real" crime.[68] To transfer environmental offences from the magistrates' and crown courts would perpetuate the conception that environmental crime is different from other types of criminal conduct and therefore less serious and/or less worthy of society's indignation.

The second possible solution to the problems created by a magistracy which is arguably "environmentally myopic" (to paraphrase Lord Woolf LCJ)[69] is increased training of magistrates in an effort to enhance environmental awareness and expertise of the technical aspects of environmental law. This has occurred to a limited extent but again is not without difficulties. In particular, one could point to the apparent unfairness in allowing the environment agency, for example, to educate the magistracy on the importance of environmental protection without affording a similar opportunity to the regulated sector to emphasise the public benefits of the activities they engage in.[70]

The final solution to low fines is the creation of minimum or mandatory fines for environmental offences.[71] Providing minimum statutory penalties for certain environmental crimes would create difficulties. Such a course of action would essentially strip the judiciary and magistracy of their long-held sentencing discretion. Although a common feature of the English legal system in the eighteenth century, in modern times mandatory sentences are the exception rather than the rule.[72] Thus, increased use of mandatory fines would be at odds with the long-established tradition of prescribing maximum sentences only. Mandatory sentences in general are often utilised where the legislature intends to convey the clear message that certain proscribed conduct will not be tolerated. The most obvious contemporary example is the introduction of mandatory sentences for the possession of firearms following considerable public unease at the levels of gun-related crime in the UK.[73] The imperative to deter certain types of conduct supersedes traditional mitigating sentencing factors such as previous good record, guilty pleas etc. The major difficulty of using such sentences is the creation of an entirely blunt instrument, which clearly has the

66 R Macrory and M Woods, *Modernizing Environmental Justice: Regulation and the role of an environmental tribunal* (London: Centre for Law and the Environment, University College London 2003) available at: www.ucl.ac.uk/laws/environment/tribunals/. The authors' advocate the establishment of an environmental tribunal with a non-criminal jurisdiction dealing with, inter alia, appeals against administratively imposed notices or the refusal to grant a licence or permit.

67 *R* v *F Howe & Son (Engineers) Ltd* [1999] 2 All ER 249, at p. 254.

68 See n. 106 and accompanying text.

69 See Woolf, "Are the judiciary", n. 63 above.

70 See Parpworth et al., "Environmental offences", n. 61 above, at p. 564.

71 R Mushal, "Reflections upon American environmental enforcement experience as it may relate to post-Hampton developments in England and Wales" (2007) 19(2) *JEL* 201, at p. 215.

72 K Warner, "Mandatory sentencing and the role of the academic" (2007) 18 *Criminal Law Forum* 321, at p. 323.

73 See the Firearms Act 1968, s. 51A, inserted by the Criminal Justice Act 2003, s. 287.

potential to create considerable unfairness. Such unfairness is more pronounced where the minimum sentence prescribed is custodial; where one is dealing with mandatory fines in an environmental regulatory context, objections based on unfairness may be tempered by the imperative to protect the environment.[74] To mitigate such unfairness (perceived or real), Parliament may allow a court to avoid the minimum sentence where exceptional circumstances exist. The extent to which the sentence retains a mandatory quality is therefore contingent upon the breadth of the interpretation of "exceptional circumstances" proffered by the courts. Given the tradition of judicial discretion in sentencing, the courts may well promulgate an expansive interpretation of such a proviso thereby undermining the rationale of a mandatory fine.

The preceding section has illustrated that the possible solutions to the problem of low fines for environmental crime (assuming one accepts that fines for such offences are indeed lower than they ought to be) are not entirely unproblematic. Further, in practical terms, with the possible exception of increased training of the judiciary and magistracy, all would require primary legislation. Given such difficulties, one should consider alternatives to fines as a form of sanction.

Alternatives to fines

A possible solution to the externalities considered above is the use of custodial sentences even within the context of corporate regulatory breach. As Brickey has noted: "jail time is one cost of business that cannot be passed on to the consumer".[75] Further, Glasbeek advocates the tactic of prosecuting individuals within in a company:

> Every breach of a legal proscription requires the doing of an act by one or more persons. There is no reason that they should not be prosecuted as individuals. This might have the desired effect of deterring them and others.[76]

While incarceration has been considered an option in American academic literature,[77] the greater use of custodial sentences for corporate regulatory breach has never been advocated as a viable enforcement option in England and Wales. The use of custodial sentences in the context of corporate crime is not without precedent. The Environmental Protection Act 1990, section 157, and the Water Resources Act 1991, section 217, both provide that a director, manager, secretary or other similar officer may be prosecuted where a body corporate is guilty of an offence under the Acts and that offence is proved to have been committed with the consent or connivance of, or to be attributable to any neglect on the part of, any director, manager, secretary etc. of the body corporate.

74 Warner, "Mandatory sentencing", n. 72 above, at p. 330.

75 Brickley, "Environmental crime", n. 14 above, at p. 506.

76 H Glasbeek, "Why corporate deviance is not treated as a crime – the need to make "profits" a dirty word" (1984) 22 *Osgoode Hall Law Journal* 393. at p. 432.

77 M Minister, "Federal facilities and the deterrence failure of environmental laws: the case for criminal prosecution of federal employees" (1994) 18 *Harvard Environmental Law Review* 137, at p. 146; P Strom, "The United States Attorney's policy towards criminal enforcement of environmental laws" (1993) 2 *South Carolina Environmental Law Journal* 184, at p. 185; M Nolan and S Stahl, "The rules have changed, but the game remains the same: why the government has turned to criminal prosecution as a means of enforcing environmental laws" (1990) 7 *Cooley Law Review* 407, at p. 410.

Although an extremely rare occurrence, individual directors have been subjected to a custodial sentence.[78] However, given the need to prove consent, connivance etc. it would seem that the above provisions may only feasibly be utilised in the case of sole-traders or very small companies. Further, from the limited guidance provided by the courts, it appears that the imposition of a custodial sentence will only be considered where the case involves repeated breaches of an Act which are blatant and/or expose members of the public to hazardous substances.[79]

The idea which underpins the use of individual custodial sentences within the context of corporate crime is simple: if the upper echelons of a company are aware that if the organisation for which they are responsible engages in activities which constitute environmental crime, which in turn creates a significant possibility of imprisonment, then a real incentive is provided to put into place mechanisms to ensure full regulatory compliance. However, the increased use of incarceration for environmental crime is not without practical difficulties; the most obvious of which is an ever-burgeoning prison population.[80] Such practical problems are easy to overstate; it is not suggested that imprisonment for environmental crime ought to be become commonplace but rather should not be ruled out as a possible sanction for the most serious of offences. Even this modest change of emphasis would require the statutory language used in provisions such as section 157 of the Environmental Protection Act 1990 to be modified. The requirement to establish consent, connivance etc. is an onerous hurdle to overcome unless dealing with the smallest of companies or sole-traders. Further, the greater use of incarceration for environmental offences would be politically highly contentious. First, from the perspective of public opinion, it is highly questionable whether environmental crimes are subjected to the same level of moral indignation that one would associate with other crimes. "Tough on environmental crime, tough on the causes of environmental crime" is an unlikely political slogan. Secondly, those on the right of the political spectrum are generally vociferously critical of what they see as lenient sentences in general. A criminal justice policy which imprisoned environmental offenders and imposed non-custodial sentences on muggers etc. would be hugely controversial and perhaps politically untenable.

In order to complete the sanctioning picture, it is necessary to consider the final weapon in the regulator's armoury, namely revocation or suspension of a licence or permit. Many activities which have the potential to cause environmental damage may only be undertaken with some form of permit, usually but not exclusively, issued by the environment agency. Waste management and water treatment works are two obvious examples, in addition, a vast range of industrial activities and processes require an IPPC (Integrated Pollution Prevention and Control) permit. Given that the acquisition of a permit is a condition precedent of engaging in the relevant activity, the revocation or suspension of a permit is tantamount to corporate capital punishment on the basis that the company in question cannot continue (legally) with the activity in question. Thus, in terms of deterrence theory,

78 E.g. *R* v *Terence William Garrett* [1997] 1 Cr App R 109, where the defendant was sentenced to 12 months' imprisonment for offences committed contrary to the Environmental Protection Act 1990, s. 33. See also ENDS Report 273 (October 1997), p. 45. The defendant was sentenced to two months' imprisonment following warnings from the Environment Agency in relation to an unbunded oil container from which oil escaped and polluted a nearby watercourse.

79 *R* v *O'Brian and Enkel* (2000) Env LR 653; the case involved the dumping of some 2000 lorry tyres contrary to the Environmental Protection Act 1990, s. 33. On the facts, the Court of Appeal overturned an eight-month sentence on the basis that while the tyres were unsightly they posed no threat to public safety and were not a nuisance. As such the "custody threshold" was not passed.

80 See generally: www.hmprisonservice.gov.uk/resourcecentre/publicationsdocuments/index.asp?cat=85.

the revocation of a permit may be seen as a form of incapacitation.[81] Given the draconian nature of this type of enforcement action, the revocation or suspension of a licence is located at the apex of the regulatory pyramid conceptualised by Gunningham et al.[82] It seems that revocation or suspension of a licence is a seldom-utilised option of last resort, only considered by enforcement agencies where criminal prosecution has proven to be ineffectual.[83] As with the externalities associated with the imposition of fines or other financial penalties, an absence of competition within the market inhibits the extent to which revocation or suspension may be utilised as a sanction. For instance, where a sewage treatment company, for example, enjoys a near monopoly, the revocation of a licence or permit will have severe, even catastrophic, environmental consequences on the basis that no other company will be in a position to undertake the activities previously carried out by the miscreant firm. As such, the undesired consequences of suspension or revocation will often rule out such a course of action.

Within the context of deterrence, the main strength of utilising custodial sentences and the suspension/revocation of a licence is that such penalties are highly draconian in nature. However, this strength is also a weakness: in a climate where attitudes to environmental crime arguably remain ambivalent, such sanctions are too severe to be contemplated in all but the most serious of cases.

It is not submitted here that those who advocate larger fines or more severe sanctions are misguided. On the contrary, it may be safely assumed that recourse to stricter penalties so that the punishment administered by the courts is commensurate with the gravity of the particular offence would be a development welcomed by environmentalists and regulators alike. However, one should not overlook the importance of effective enforcement: it seems that the certainty of punishment appears to have a greater inhibitive effect on levels of crime than the severity of the punishment.[84] Indeed, the general consensus of academic opinion seems to indicate that, given the psychological features of deterrence, subjective perception is critical, a viewpoint succinctly expressed by Williams and Hawkins: "The more the individual *perceives* legal sanctions as certain, swift, and/or severe, the greater the *perceived* cost of crime and thus the probability of deterrence."[85] In sum, in order to comprehend fully the concept of deterrence, one cannot simply analyse legal sanctions per se. Of equal, if not greater, importance is enforcement style, in particular, whether the enforcement policy adopted is likely to augment the perception that apprehension is more probable.

Enforcement strategies: compliance versus deterrence

It is not the purpose of this article to consider in detail the nature of enforcement styles adopted by regulators, save to say that two discernable regulatory strategies have been identified, namely compliance and deterrence.[86] Rather than a strict dichotomy, one is dealing here with a spectrum of approaches so any particular enforcement policy may be located on a continuum with compliance and deterrence situated at either end. Compliance-based approaches are characterised by advice, persuasion and negotiation where criminal prosecution is used as a last resort. Deterrence-orientated strategies, as the name suggests,

81 Ogus and Abbot, "Sanctions", n. 36 above, at p. 294.

82 N Gunningham, P Grabosky and D Sinclair, *Smart Regulation: Designing environmental policy* (Oxford: Clarendon Press 1998), p. 397.

83 Ogus and Abbot, "Sanctions", n. 36 above, at pp. 288–99.

84 Jacob, "Deterrent effects", n. 32 above, at p. 63.

85 Williams and Hawkins, "Perceptual research", n. 27 above, at p. 547 (emphasis added).

86 See, generally, A Reiss, "Selecting strategies of social control over organisational life" in K Hawkins and J Thomas (eds), *Enforcing Regulation* (Boston: Kluwer-Nijhoff 1984).

rely heavily on stringent enforcement and punishment of offenders. A third "hybrid" option has been identified, where regulation is "responsive" in the sense that the style adopted by the enforcement agency will vary according to the perception of the regulated actor.[87] More recently, the concept of "really responsive" regulation has been promulgated whereby the regulator is not only responsive to the attitude of the regulatee but also to wider considerations such as, inter alia, the institutional environment and performance of the regulatory regime.[88] In the recent past, one could confidently state that the qualitative studies undertaken on regulatory enforcement indicated that enforcement bodies in England and Wales relied predominantly on the compliance approach as an enforcement style.[89] Whether compliance remains in the ascendancy is an open question, a question which requires empirical investigation. Baldwin, for example, refers to the "new punitive regulation" but acknowledges that aggressive rhetoric is not always matched by enforcement practice.[90] From one perspective, the arguments advanced in this article are unaffected by the question of which enforcement style is more prevalent. If compliance-orientated enforcement is still widely used, this article may be seen as advocating a subtle departure from such an approach. Alternatively, if a transition from compliance to more punitive methods of assessment is already underway, the contentions promulgated here may be used to support such developments.

The reasons which underpin compliance-based approaches are multifarious and will be considered under the following non-exhaustive headings: resource limitations, perceptions of the regulated community and the status of environmental crime.

Resource limitations

From a practical perspective, prosecution may be utilised as a tool of last resort because of resource limitations, that is enforcement bodies simply do not have the financial capability to prosecute every regulatory breach. The issue is succinctly summed up by Stigler: "There is one decisive reason why the society must forego 'complete' enforcement of the rule: enforcement is costly."[91] Those of a cynical persuasion may well view the Environment Agency's policy of "risk-based regulation"[92] euphemistically in light of funding constraints. As Grabosky has contended, "the question of efficiency looms even larger in the climate of fiscal austerity in which most governments operate today".[93] The predominance of strict liability for most environmental offences removes any necessity to prove fault. Nevertheless, the accepted wisdom is that criminal prosecution is a relatively costly and time-consuming option. As Mushal has contended:

> With respect to enforcement, generally, toughness and speed are inversely related to one another – and there is a third element in that relationship: cost. The toughest sanctions are criminal, but criminal prosecution is not a speedy process and it has high resource costs . . .[94]

87 I Ayres and J Braithwaite, *Responsive Regulation: Transcending the deregulation debate* (Oxford: Oxford Socio-legal Studies, OUP 1994).

88 R Baldwin and J Black, "Really responsive regulation" (2008) 71(1) *MLR* 59.

89 Hawkins, *Environment and Enforcement*, n. 1 above; B Hutter, *The Reasonable Arm of the Law? The law enforcement procedures of environmental health officers* (Oxford: Clarendon Press 1988).

90 R Baldwin, "The new punitive regulation" (2004) 67(3) *MLR* 351, at pp. 358–9.

91 Stigler, "Optimum enforcement", n. 38 above, at pp. 526–7.

92 Environment Agency, *A Guide to Modern Regulation*, available at: www.environment-agency.gov.uk.

93 P Grabosky, "Regulation by reward: on the use of incentives as regulatory instruments" (1995) 17(3) *Law and Policy* 257, at p. 258.

94 Mushal, "Reflections", n. 71 above, at pp. 206–7.

Indeed, the greater use of administratively imposed monetary penalties is seen as an attractive option because they avoid recourse to the criminal courts. Such logic would seem to dictate that deterrence-based strategies, which rely more heavily on criminal prosecutions, are therefore less economically efficient than compliance-orientated approaches. Is this a safe assumption? Although an issue which requires detailed empirical exploration, it could be contended that negotiation, persuasion and advice, which are the hallmarks of any compliance strategy, are more time-consuming and therefore resource-depleting activities than the instigation of a criminal prosecution. For example, Nelken has postulated that those agencies with fewer personnel are more likely to adopt an "insistent" strategy:

> The reason for the adoption of these different strategies is partly that those authorities with less manpower available are forced to use the brisker, less time-consuming "insistent" methods (negotiating takes time).[95]

Further, the resource implications of instigating criminal prosecutions may be mitigated by the award of costs to the prosecuting authority. It is not suggested here that deterrence-based strategies are more inexpensive than those based on compliance, but that the tendency to assume the opposite, without extensive empirical data on which to base one's assertion, is premature.

Perceptions of the regulated community

Kagan and Scholz have identified three categories of non-compliant actor, namely, the amoral calculator, the political citizen and the organisationally incompetent.[96] Compliance-orientated strategies are perhaps partially explicable on the basis that most members of the regulated community are considered by enforcement agencies to be essentially socially responsible, as Hawkins has claimed: "Pollution control staff regard most dischargers as basically, if reluctantly, law abiding."[97] As such, the commonly held perception is that most violations occur as a result of ignorance or inadvertence.[98] Thus, in terms of the tripartite analysis highlighted above, the presumption is that the regulated community is largely comprised of political citizens and, where breaches occur, they do so as a result of organisational incompetence. Very few organisations would therefore fall into the category of amoral calculator who deliberately violates the law on the basis that the gains/profits from such breaches will exceed the cost of apprehension. This perception of the corporation is clearly not universally accepted. Proponents of Marxist ideology would contend that the corporation, as an entity operating within a market system, is inherently criminogenic; the prime motivation of business is the creation of profit and this supersedes any other consideration, including legality.[99] Such a contention is a radical, perhaps even an extreme perspective. Nevertheless, the paradigm which holds that the regulated sector is largely comprised of responsible citizens seems at odds with research conducted in 1998 where 38 per cent of Environment Agency field officers considered profit-seeking to be a major cause of environmental breaches.[100] Of course, the regulated sector is keen to perpetuate a benign perception; many corporations go to considerable lengths to portray themselves as "green":

95 D Nelken, "Why punish?" (1990) 53(6) *MLR* 829, at p. 832.

96 R Kagan and J Scholz, "The criminology of the corporation and regulatory enforcement strategies" in Hawkins and Thomas, *Enforcing Regulation*, n. 1 above.

97 K Hawkins, "Bargain and bluff" (1983) 5(1) *Law and Policy Quarterly* 35, at p. 43.

98 J Rowan-Robinson and A Ross, "Enforcement of environmental regulation in Britain: strengthening the link" (1994) *JPL* 200, at p. 201.

99 See, generally, F Pearce, *Crimes of the Powerful* (London: Pluto Press 1976).

100 Cited in P de Prez, "Excuses, excuses: the ritual trivialisation of environmental prosecutions" (2000) 12(1) *JEL* 65, at pp. 68–9.

> In a nutshell, the corporate redefinition of the word "green" presented the public with a mild, less radical and de-politicised environmental vision along with *less drastic responses to . . . environmental issues.*[101]

In addition to perceptions of the regulated community, compliance-based enforcement is also influenced by what enforcement agencies consider to be their optimal role. In particular, regulators often refuse to perceive of their role as akin to that of the police but rather as primarily givers of guidance and advice. A corollary of this perception is that the power to instigate a criminal prosecution will only be utilised where other attempts to secure compliance have been exhausted. Pearce and Tombs have noted that those who reject the "regulator-as-policeman" strategy often invoke negative stereotypes of regulatory inspectorates, which became a feature of "anti-statist" rhetoric from the 1980s onwards.[102] These negative stereotypes, which see an enforcement agency operating in an inflexible, legalistic manner, are implicit in certain aspects of the Macrory report. When considering the potential use of administrative penalties, the report suggests that the revenue generated from such penalties should not be retained by the relevant enforcement agency, to avoid "any perverse financial incentives"[103] or the fostering of a "parking ticket mentality".[104] This mistrust of regulators is unsurprising. The overall focus of the contemporary regulatory reform debate is on how regulation may be modified so that administrative burdens on business are minimised. As such, it has been contended that compliance strategies "dovetail" with neo-liberal ideology.[105] It is not suggested that the regulated sector is largely comprised of malicious corporations, keen to increase profits at the expense of the environment, but rather it should not be automatically assumed that those subject to regulation are socially responsible. Of course, such a statement seems to do little more than subscribe to the concept of responsive regulation. However, it is submitted that regulators, as an initial starting point, could reverse their long-held presumption in favour of a more sceptical attitude toward those that they regulate.

Status of environmental crime

Writing in 1983, Hawkins contended that: "Though the law recognises no distinction, most people are unwilling to talk of crime when they discuss pollution."[106] In 1992, Harris optimistically claimed that the enactment of the Environmental Protection Act 1990 heralded a shift in thinking so that environmental crime was no longer seen in morally neutral terms.[107] The question here is: to what extent is environmental crime subjected to the same level of moral indignation that one associates with other forms of criminal conduct? In contemporary times, environmental issues have become increasingly

101 M Lynch and P Stretsky, "The meaning of green: contrasting criminological perspectives" (2003) 7 *Theoretical Criminology* 217, at p. 220 (emphasis added).

102 F Pearce and S Tombs, "Ideology, hegemony, and empiricism" (1990) 30 *British Journal of Criminology* 423, pp. 427–8.

103 The Macrory report, n. 3 above, at para. 3.45.

104 Ibid., at para.3.43.

105 R White, "Environmental issues and the criminological imagination" (2003) 7 *Theoretical Criminology* 483, pp. 486 and 497.

106 Hawkins, "Bargain and bluff", n. 97 above, at p. 37. See also Stigler, "Optimum enforcement", n. 38, at p. 526, who even questions the use of the word "criminal" to describe what he considers to be "trifling offences".

107 R Harris, "The Environmental Protection Act 1990 – penalising the polluter" (1992) *JPL* 515, at p. 516.

pertinent.[108] Nevertheless a considerable level of moral ambivalence is still evident when one considers environmental offences. The reasons for such ambivalence are no doubt numerous and complex.

The obvious question is, therefore: can the criminalisation of conduct which harms the environment be justified using traditional theories of criminalisation? Arguably the most widely accepted test of whether a particular type of conduct ought to be criminalised is the harm principle first promulgated by Mill[109] and later developed by Feinberg.[110] From an environmental perspective, the principle is usually expressed as an anthropocentric concept in terms of harm to other humans. It can clearly be determined that harm to the environment additionally causes, to a greater or lesser extent, harm to humans. One of the distinctive features of environmental crime is that often the offence will lack an immediate victim, or at least an immediate human victim. However, the harm principle is sufficiently malleable to incorporate less immediate forms of harm.[111] Indeed, one commentator has considered the issue of environmental harm in apocalyptic terms to the extent that humankind's very existence is in jeopardy.[112] Considered in this light the view that environmental crimes are fundamentally less serious than other crimes may be seen as puzzling, as Mushal has contended:

> Some judges . . . do not view them [environmental crimes] as being as serious as conventional crimes such as bank robberies and drug dealing . . . even though the consequences of environmental crimes may be far broader, more severe and longer lasting.[113]

As such, without considering each and every environmental offence and its potential harmful consequences, it is submitted that from a general theoretical standpoint, the designation of conduct which damages the environment as criminal may be justified on the basis of the harm principle. Given that the proscription of activities which cause environmental damage may be supported using the traditional test of criminalisation, one may question why environmental crime is considered inherently less serious than other forms of criminal behaviour.

The potential benefits of deterrence-oriented enforcement

The preceding sections of this article have sought to establish two propositions. First, that to concentrate simply on the severity of sanction is only ever likely to represent a partial solution to the problem of regulatory breach. Secondly, the assumptions on which compliance-orientated strategies are based can be questioned, if not totally undermined.

108 A survey carried out on behalf of the Department for Environment, Food and Rural Affairs discovered that, of the issues people think government should be dealing with, the environment (19%) was the fourth most commonly mentioned behind crime (49%), health (47%) and education (36%): 2007 Survey of Public Attitudes and Behaviours Toward the Environment, available at www.defra.gov.uk/environment/statistics/pubatt/download/pubattsum2007.pdf.

109 J S Mill, "On liberty" in S Collini (ed.), *Mill, On Liberty and Other Writings* (Cambridge: CUP 1989).

110 J Feinberg, *Harm to Others* (New York: OUP 1984); *Offense to Others* (New York: OUP 1985); *Harmless Wrongdoing* (New York: OUP 1990).

111 A von Hirsch, "Extending the harm principle: 'remote' harms and fair imputation" in A Simester and A Smith (eds), *Harm and Culpability* (Oxford: Clarendon Press 1996), pp. 259–76, has questioned the use of the harm principle as a justification for the criminalisation of remote harms and suggests "fair imputation" as a possible alternative.

112 C Cullinam, *Wild Law* (Claremont, SA: Siber Ink/Gaia Foundation 2002), at p. 15, states: "Human societies are savaging Earth. Right now the human societies that currently dominate our planet are precipitating what is being described as the sixth mass extinction."

113 Mushal, "Reflections", n. 71 above, at pp. 211–12.

The role of regulatory enforcement bodies such as the environment agency and local authorities is multifaceted and clearly includes the giving of advice and guidance. Further, one cannot deny the importance of maintaining a constructive relationship with the regulated sector.[114] However, an over-reliance on compliance-oriented enforcement strategies is problematic. First, compliance approaches have the potential to undermine transparency, the "power to be lenient [also] is the power to discriminate".[115] Second, reliance on compliance methods creates the danger of an overly cosy relationship between regulators and those regulated, i.e. regulatory capture. In a study of regulatory compliance within the context of factory legislation it has been contended that one effect of "under-enforcement" is that the relevant form of illegal activity becomes "conventionalised".[116] If the regulated sector is aware (and one can assume that it will be) that the relevant enforcement authority adopts a strategy based on co-operation, advice and persuasion, they are likely to "test the water" by engaging in activities which do not conform to the standards set in legislation, safe in the knowledge that they will be given a second chance should the violations be brought to the attention of an enforcement agency. The concept of "risk-based regulation"[117] is problematic in this context; if a regulatee is not considered to pose a high risk of non-compliance, any illegal activity may remain undetected for a considerable time. Proponents of compliance-orientated approaches sensibly argue for persuasion and negotiation against a backdrop of potentially tough enforcement via the criminal law, i.e. regulators "will be able to speak more softly when they are perceived as carrying big sticks".[118] However, where prosecution is a relative rarity, this threat has the potential to become utterly empty.

In terms of deterrence theory, it is axiomatic that the likelihood of apprehension is a crucial factor. However, the probability of apprehension is a broad concept with three distinct elements. First, there is the likelihood that a given regulatory transgression will actually be brought to the attention of an enforcement agency. Second, if an enforcement body is aware of an alleged breach, then a decision has to be taken whether or not to launch a prosecution. Finally, if a prosecution is instigated and a not-guilty plea is put forward, the additional likelihood of a conviction has to be considered. A compliance-orientated enforcement style, where recourse to prosecution is utilised as a last resort, clearly has the potential to undermine the deterrent effect at stage two of the above tripartite analysis. Those who are contemplating transgressing the law may well take into account the fact that even if an illegal activity is discovered, the probability of the relevant enforcement agency instigating a criminal prosecution is relatively small. One can safely assume a correlation between the likelihood of prosecution and the extent of any deterrent effect. Given that the general academic consensus seems to be that the perceived likelihood of apprehension has a greater influence on the deterrent effect than the severity of sanction,[119] a move towards a more deterrence-orientated style of enforcement has the potential to augment the

114 C Abbot, "The regulatory enforcement of pollution control laws: the Australian experience" (2005) 17(2) *JEL* 161, at p. 165.

115 *McCleskey* v *Kemp*, 481 US 279, 297, 312 (1987) (quoting K C Davis, *Discretionary Justice: A preliminary inquiry* (Baton Rouge: Louisiana State UP 1969), p. 170).

116 W Carson, "Some sociological aspects of strict liability" (1970) 33 *MLR* 396.

117 See, generally, J Black, "The emergence of risk based regulation and the new public risk management in the UK" (2005) *PL* 512. For a brief critique of risk-based regulation, see Baldwin and Black, "Really responsive regulation", n. 88 above, at pp. 66–7.

118 Ayres and Braithwaite, *Responsive Regulation*, n. 87 above, at p. 6.

119 For a comprehensive review of relevant literature, see A von Hirsch, A Bottoms, E Burney and P Wikström, *Criminal Deterrence and Sentence Severity: An analysis of recent research* (Oxford: Hart Publishing 1999).

perception that the probability of apprehension is higher, thereby bolstering the deterrent effect of the law.

It could be argued that deterrence-orientated strategies are likely to prove counter-productive on the basis that greater recourse to prosecution is likely to create an "organised culture of resistance".[120] Baldwin has cogently identified the possible shortcomings of deterrence or punitive-based enforcement;[121] most notably, it is contended that corporations may respond to "punitive risks" irrationally and may be poorly equipped to deal with or anticipate such punitive risks and the effects of sanctions.[122] Further, Baldwin contends that even where a corporation acts rationally (or at least along rational lines), it cannot be assumed that this will lead to compliance.[123] For example, the prospects of future punitive sanctions may induce corporations merely to reduce the impact of any sanction by, inter alia, taking out insurance.[124] In relation to the insurance point, one could question the extent to which it is desirable to allow insurance against the risk of fines;[125] in any event while a simplistic analysis would seem to suggest that the availability of insurance undermines any deterrent effect, the nature of the actuarial analysis is such that those corporations with internal structures prone to non-compliance will be required to pay higher premiums. Further, it is safe to assume that, after a corporation has invoked an indemnity clause on one occasion, future insurance premiums are likely to rise considerably.

The case against punitive forms of regulation articulated by Baldwin is underpinned by the view that corporations have difficulty in responding to punitive risks:

> It may be the case that sanctions, when imposed, do have an impact well beyond the quantum of a fine but such secondary impacts may stimulate extremely low levels of *ex ante* compliance-seeking behaviour because they are very poorly anticipated.[126]

In response to this assertion, one may point to the intuitive view that corporations (and indeed individuals) will often experience difficulty in accurately assessing new risks or increases in risk. Thus, if within a particular regulatory context, a shift of emphasis is made so that punitive risks become more prevalent due to an increase in the likelihood of prosecution, an initial period of confusion is unsurprising. However, assuming a sufficient level of transparency, once the new policy has "bedded in", a corporation may be able to anticipate and respond to any punitive risks in a more coherent and informed manner.

If one assumes that the majority of the regulated sector is basically law-abiding and that transgressions occur due to inadvertence, carelessness or ignorance, then a strategy focused on advice and guidance may be seen as particularly suitable. However, it is not submitted that enforcement bodies ought to adopt a "zero tolerance" approach, but rather that a subtle but noticeable shift towards deterrence could potentially have a greater preventative effect. Even if one accepts the view that most violations are non-deliberate, a move towards deterrence could still be justified. First, can it be contended that the ignorant or inadvertent company is free of culpability? Perhaps. The stock response would be that with the

120 Grabosky, "Regulation by reward", n. 93 above, at p. 263.

121 Baldwin, "New Punitive", n. 90 above.

122 Ibid., at p. 370.

123 Ibid.

124 Ibid., at p. 371. For a critique of the practice of insuring against fines in the context of marine pollution, see O Lomas, "The prosecution of marine oil pollution offences and the practice of insuring against fines" (1989) 1 *JEL* 48.

125 See the *dicta* of Rowlatt J in *R Leslie* v *Reliable Advertising etc. Agency Ltd* [1915] 1 KB 652, at pp. 658–9, and Denning J in *Askey* v *Golden Wine* [1948] 2 All ER 35, at p. 38.

126 Baldwin, "New punitive", n. 90 above, at p. 372.

ever-burgeoning body of environmental regulation to which business is subject, maintaining a high level of awareness of one's legal responsibilities is an onerous task. Such arguments are worthy of rejection for two reasons. First, such a contention comes close to violating the fundamental principle that ignorance is no defence. Second, compliance strategies developed in a pre-internet age, where information did not flow as freely as today. In a contemporary context, regulators go to considerable lengths to publicise new and existing legal obligations. The Environment Agency's NetRegs is the most obvious example.[127] In light of the availability of numerous avenues of information, can it really be contended that the ignorant or unintentional transgressor is free of blame? It is submitted that a move towards deterrence-based strategies would ultimately not only provide a greater retributive element, where the polluter is a recalcitrant offender, but would also reduce the likelihood of less culpable breaches. If many offences are committed simply as a result of human error or inadvertence, then an increase in prosecutions would help to publicise the extent of the regulated communities' legal obligations. Further, it is logical to assume that a desire to avoid a potentially punitive sanction would encourage the education and training of employees. Those who advocate the greater use of deterrence-based enforcement as a means of influencing the internal procedures of corporations may well gain encouragement from qualitative research conducted by Baldwin: "For many firms interviewed, the imposition of a first punitive sanction produced a sea-change in attitudes."[128] So much for specific deterrence – one may question whether the same holds true for general deterrence. The answer seems to be in the affirmative:

> Of the three-fifths of respondents who were aware of firms in their sector being punitively sanctioned, 57 per cent stated that this had impacted very strongly on their own management of punitive regulatory risks.[129]

It may be the case that many environmental offences are committed as a result of accident, mechanical malfunction, operator incompetence or even unforeseeable events. For example, in *Empress Car Company (Abertillery) Ltd* v *National Rivers Authority*,[130] the House of Lords confirmed the defendant's liability, notwithstanding that the constituents of the relevant offence were brought about by an unidentified vandal. The potential effectiveness of deterrence is based on the assumption that a rational corporation would respond by putting in place measures which reduce the likelihood of vandalism in the future. The obvious problem, notwithstanding that rationality cannot automatically be presumed, is that a company will be subject to limitations (financial or otherwise) in terms of reducing the risk of even inadvertent non-compliance. However, it is not suggested that deterrence-based enforcement is in any way a panacea; rather it is argued that deterrence is capable of performing an internal educative and managerial function. One should not rule out a particular approach merely on the ground that it will not prove effective in all cases.

Preceding sections of this article have attempted to analyse some problematic aspects of an approach which simply concentrates on the severity of sanction as a means of increasing the deterrent effect of regulatory law. It is contended that a move towards deterrence-oriented strategies would attenuate, to a greater or lesser extent, many of the difficulties highlighted. First, the oft-quoted reason for the problem of low fines for environmental offences is the inexperience of the magistracy and judiciary in dealing with environmental

127 NetRegs is an Environment Agency website which aims to provide detailed information to businesses in the UK on the extent of their environmental legal obligations. Companies may also subscribe to free NetRegs e-alerts which highlight new developments: www.netregs.gov.uk/netregs/.

128 Baldwin, "New punitive", n. 90 above, at p. 361.

129 Ibid.

130 [1998] Env LR 396.

crime. A strategy based on deterrence, where recourse to criminal prosecution is made more readily, would increase the number of cases heard by the magistrates, which would, in turn, facilitate awareness of environmental issues and the technical aspects of environmental offences. In effect, an increase in the number of cases heard by the courts would go some way towards breaking the "vicious circle" highlighted by Watson.[131] Thus, the overall level of fines could increase organically, without the need for increased training of the magistracy and judiciary and/or the imposition of mandatory fines.

Compliance strategies may be partially justified on the basis that a softer enforcement approach is appropriate on the grounds that environmental crimes are subject to moral ambivalence. Put another way, where an illegal act is the subject of moral indignation, it is extremely difficult to imagine an enforcement agency adopting a strategy of advice, persuasion and negotiation. For example, one can only imagine the public outrage and media hysteria if the police merely advised paedophiles how to avoid abusing children and the Crown Prosecution Service prosecuted only where all other avenues had failed to prevent the undesired conduct. This example is clearly an extreme one; one would never equate the moral culpability associated with paedophilia with crimes which damage the environment. Nevertheless, an amoral conception of environmental crime creates a climate where compliance-orientated approaches are able to gain ascendancy. The traditional argument is that one of the causes of the prevalence of compliance styles is a morally neutral conception of environmental crime. However, one could question whether this view reverses cause and effect; could it not be contended that compliance-orientated enforcement is a cause of the moral ambivalence associated with environmental crime? As Nelken had postulated:

> Part of the difficulty of resolving this problem is often said to result from the uncertainty about what the public feels concerning these offences. But it could well be argued that community perceptions of these offences is as much an effect as a cause of the method of enforcement used.[132]

In this context, the offence of driving under the influence of alcohol is a suitable analogy; as with many environmental crimes, the offence is inchoate in the sense that a crime is committed even if the standard of driving has not been noticeably impaired, i.e. the particular activity is proscribed because of the potential to cause serious harm notwithstanding that the harm has not actually occurred. More importantly for present purposes, driving while under the influence of alcohol is an activity which has undergone a transition from public indifference to moral indignation:

> Not only is it less common these days to hear the view expressed that drivers caught with excess alcohol have just been "unlucky", but the fact that majority opinion in this country is currently in favour of random breath testing also suggests that drink-driving may be condoned to a lesser extent that formerly.[133]

Thus, those who espouse the view that environmental crime ought to be treated more seriously may gain encouragement from the extent to which the law is capable of influencing (albeit slowly) public perception. A transition from compliance to deterrence-based enforcement strategies has the potential to modify progressively how the public perceive of environmental crime. Alternatively, if it is accepted that public perception is gradually changing so that environmental crime is decreasingly seen in morally neutral terms, greater use of deterrence-orientated enforcement could accelerate such change. The

131 Watson, "Environmental offences", n. 22 above, at p. 199.

132 Nelken, "Why punish?", n. 95 above, at pp. 831–2.

133 C Corbett and F Simon, "Police and public perceptions of the seriousness of traffic offences" (1991) 31 *British Journal of Criminology* 153, at p. 154.

moral status and public perception of environmental crime is of fundamental importance, not only from a theoretical standpoint but also a practical perspective. If policymakers are cognisant of changing attitudes to the environment, this will place considerable political constraints on any attempts to reduce funding of enforcement agencies and/or would increase considerably the lobbying power of those who argue in favour of the dedication of greater resources to enforcement activity.

One may predict with a reasonable degree of certainty that a move towards more deterrence-orientated enforcement will be resisted by the regulated sector; business will no doubt advance the familiar argument that most regulatory breaches occur through inadvertence and ignorance. Indeed, the use of more coercive methods may be seen as running counter to the trend of using more indirect means, such as economic instruments, in order to achieve environmental objectives. A laissez-faire conception of state intervention, borne out of antipathy toward bureaucracy and/or the allure of market instruments,[134] would seem to suggest that non-coercive means are preferable to coercive ones. The distinction between tools which encourage desirable conduct and tools which punish undesirable conduct is often difficult to draw.[135] Nevertheless the criminal law has a unique stigmatising effect and sends the clear message that certain types of conduct will not be tolerated by society.[136] Neiman argues strongly in favour of the use of more coercive methods of achieving policy objectives but accepts that "heavy-handedness" in government is only acceptable within a democracy.[137] It is submitted that democracy is a necessary but not sufficient condition for the greater use of deterrence-based enforcement strategies. In order to provide a clear justification for the use of more insistent enforcement a number of developments would be required. First, in the interests of fairness and balance, if one is to make greater use of the "stick" then a corresponding increase in the use of the "carrot" would seem appropriate. There is no logical reason why the use of incentives to achieve regulatory compliance and the increased use of prosecution and punishment should be seen as mutually exclusive: "a regulated entity may be simultaneously offered a carrot and threatened with a stick".[138] The use of incentives or rewards to achieve environmental objectives may take many forms (a comprehensive discussion of such policy instruments is beyond the scope of this article).[139] However, far from being contradictory, the use of such facilitation strategies may actually compliment and enhance deterrence-based enforcement strategies. For example, one very obvious objective of waste management is to ensure that waste is disposed of responsibly in a manner which minimises, as far a is possible, environmental damage.[140] In order to further this objective, a two-pronged approach is used: first, facilitate the appropriate disposal of waste via the operation of a network of civic amenity sites[141] and the provision of refuse collection services;[142] secondly, criminalise any unauthorised disposal of waste.[143] Thus, those who are prosecuted for

134 M Friedman, *Capitalism and Freedom* (Chicago: University of Chicago Press 1967).

135 Brigham and Brown, "Distinguishing penalties", n. 19 above.

136 H L A Hart, *The Concept of Law* 2nd edn (Oxford: Clarendon Law Series, OUP 1994), p. 39.

137 M Neiman, "The virtues of heavy-handedness in government" (1980) 2(1) *Law and Policy Quarterly* 11, at p. 17.

138 Grabosky, "Regulation by reward", n. 93 above, at p. 271.

139 See, generally, G Balch, "The stick, the carrot, and other strategies" (1980) 2(1) *Law and Policy Quarterly* 35.

140 See Art. 4 of the Waste Framework Directive (75/442/EEC).

141 Environmental Protection Act 1990, s. 51, places a duty on waste disposal authorities (commonly a local council or county council) to provide sites where waste may be disposed of.

142 Environmental Protection Act 1990, s. 45, places a duty on waste collection authorities (commonly a local authority) to provide waste collection services.

143 The crime of unauthorised depositing of waste, commonly known as fly-tipping, is an offence by virtue of the Environmental Protection Act 1990, s. 33.

fly-tipping can hardly complain that they had no choice, given the existence of mechanisms designed to facilitate compliance with the law. In this sense, the provision of strategies designed to encourage or incentivise compliance provide a clear normative basis for the speedier recourse to prosecution associated with deterrence-orientated enforcement strategies. In summary, the use of the stick may be more readily justified when the regulated entity has first been offered the carrot.

The increased use of facilitation strategies would provide a theoretical justification for the strict enforcement of the law when sections of the regulated sector fail to take advantage of such mechanisms and transgress the law. The greater use of deterrence-based enforcement would also require changes of a more practical nature. The accepted wisdom is that recourse to prosecution is more costly than more insistent methods; an earlier section of this article has questioned such wisdom on the basis that the assumption is made without reference to detailed empirical evidence. However, given that the likelihood of apprehension is a crucial element of the deterrence formula, additional resources would be required to increase surveillance activity etc. Furthermore, if a move towards deterrence is made, the increased use of facilitation strategies would obviously create additional costs for enforcement agencies. Numerous ways of providing additional resources to enforcement agencies exist; perhaps the most obvious is to increase the funding such bodies receive from central government. This could be achieved either through the diversion of existing funds or through increases in taxation – both options would be politically contentious. An alternative is to hypothecate the fines received from environmental offences. The Macrory report conspicuously rejected the hypothecation of the proceeds of administratively imposed monetary penalties on the basis that to allow enforcement agencies to retain the income generated from such penalties has the potential to create "perverse financial incentives".[144] However, the use of hypothecation is not without precedent in environmental law. For example, following amendments to the Noise Act 1996, local authorities may now retain the proceeds of fixed penalty notices imposed for the night noise offence created by the Act providing those receipts are used to fund "qualifying functions", i.e. are utilised to fund noise-related enforcement activity.[145] One may question, therefore, why local authorities are trusted not to adopt a "parking ticket mentality" in relation to the exercise of noise-related enforcement powers. Perhaps the answer lies in the fact that the Noise Act 1996 is only applicable to dwellings and licensed premises (public houses, night clubs etc.), which reduces considerably the likelihood that the issue of fixed penalty notices will place an onerous financial burden on business. However, the apprehensive attitude to hypothecation evident in the Macrory report cannot be solely explained by a concern not to interfere too greatly with economic interests. It is possible that such a development could lead to the misuse of prosecutorial discretion. An enforcement agency, cognisant of funding constraints, may well target those offenders who have the financial resources easily to pay fines and/or administrative monetary penalties; solvent offenders are not necessarily the most serious polluters and, as such, the hypothecation of penalty receipts may operate counter-productively, with the environment suffering as a result. Nevertheless, if policymakers are serious about reducing environmental regulatory breaches then perhaps they need to place more faith in enforcement agencies to exercise a degree of common sense. The introduction of hypothecation could be brought about by an enabling statute, granting the Secretary of State the power to enact secondary legislation. If, after a sufficiently long probationary period, it was demonstrated that enforcement agencies were abusing their prosecutorial discretion by concentrating

144 The Macrory report, n. 3 above, at para. 3.45.

145 Noise Act 1996, s. 9(4A).

predominately on "easy targets" then the ability to retain the proceeds of fines etc. could be easily revoked by a further statutory instrument.

At numerous points throughout this article phrases such as "shift of emphasis" and "a move towards deterrence-orientated approaches" have been used. This raises the obvious question of how such a development could be effected. As with any prosecutorial regime, the enforcement of environmental law operates within a highly discretionary system: there is scant legal control or oversight of the use of prosecutorial powers endowed to enforcement agencies.[146] It is not in any way suggested here that prosecutorial powers ought to be more extensively regulated or more freely challenged. Even if such a course was considered desirable, effectively regulating prosecutorial decision-making is extremely problematic.[147] The current government favours the use of voluntary codes of conduct as a means of exercising a degree of control over enforcement practice. The most obvious example is the *Enforcement Concordat*,[148] which has been adopted by over 96 per cent of local and central government bodies with enforcement functions. The principles of good enforcement policy contained in the *Enforcement Concordat* are clearly geared toward a compliance-orientated approach with little or no regard paid to the dangers of regulatory capture and the importance of the probability of apprehension. Perhaps an amended concordat could be the first stage in any transition from compliance to deterrence. In any event, the main purpose of this article is to consider the arguments which may be used in support of more deterrence-orientated enforcement, i.e. we are concerned with the question of whether a deterrence approach would bolster the preventative aspects of environmental criminal law. How any changes in enforcement style would be actuated is a separate question for another time.

Conclusion

The desirability of preventing environmental harm is beyond doubt: as such, the preventative principle is universally accepted as a sound basis for environmental laws. Further, it is generally acknowledged that the imposition of reactive criminal liability may operate preventatively via the concept of deterrence. This article has focused on how the deterrence effect and, by extension, the preventative effect of regulatory criminal law may be bolstered. A common response to the question of how one reduces regulatory breach is to increase the severity of sanction; indeed, academic literature is replete with assertions that the fines imposed for environmental crime are lamentably low. It is not suggested here that an overall increase in the severity of sanction, so that the punishment imposed is commensurate with the gravity of the particular offence, would be an unwelcome development; rather it is submitted that the severity of punishment may only ever represent a partial solution to non-compliance. A preoccupation with sanctioning may allow one to overlook the fundamental importance of the likelihood of apprehension; here the enforcement style adopted by enforcement agencies is critical. The assumptions on which compliance-orientated enforcement strategies are based may be questioned if not totally undermined, as such one may consider the potential benefits of deterrence-based

146 The Deregulation and Contracting Out Act 1994, s. 5, vested in a Minister of the Crown the power to improve enforcement procedures. The provision was repealed by the Regulatory Reform Act 2001, s. 12(1), which itself was repealed by the Regulatory Reform Act 2006, Sch. 1, para. 1.

147 See, generally, S Bibas, "Prosecutorial regulation versus prosecutorial accountability", *University of Pennsylvania Law School: Scholarship at Penn Law*, Paper 253 (5 December 2008): http://lsr.nellco.org/upenn/wps/papers/253; J Vorenberg, "Decent restraint of prosecutorial power"(1981) 94 *Harv L Rev* 1521.

148 *Enforcement Concordat: Good practice guide for England and Wales*, DTI, available at www.berr.gov.uk/files/file10150.pdf.

enforcement. Speedier recourse to criminal prosecution would help increase the number of environmental cases heard by the courts which would in turn publicise the extent of the regulated sectors' legal obligations. Further, the use of more insistent methods of enforcement has the potential gradually to modify the public perception of environmental crime. The increased use of deterrence-orientated enforcement could be more easily justified if a corresponding increase in the use of facilitation strategies is actuated. Finally, serious thought ought to be given to the introduction of hypothecated penalty receipts to mitigate any potential cost implications of the change of emphasis advocated by this article.

In essence, the theme of this article could be summarised in the following manner: mainstream criminological opinion, as influenced by empirical evidence, strongly suggests that in terms of the two integral deterrence variables, the likelihood of apprehension has a greater influence on deterrence than the severity of sanction. Thus, one could consider how the probability of apprehension may be actually increased, or, given that deterrence is a psychological phenomenon, how the perception that apprehension is more probable could be bolstered. It has been argued that, in terms of approaches to regulation, deterrence-orientated enforcement has the potential to increase the dissuasive aspects of environmental criminal law, which *a fortiori* could further the aims of the preventative principle.

NILQ 60(3): 305–24

Adverse possession and informal purchasers

UNA WOODS

Lecturer, School of Law, University of Limerick

In England and Wales the enactment of the Land Registration Act 2002 dramatically reduced the scope of the doctrine of adverse possession in relation to registered land. The new adverse possession regime was justified in the report preceding the 2002 Act on the basis that it was more compatible with principles of title registration and struck a more appropriate balance between the landowner and squatter.[1] Although the 2002 Act confers the registered owner of land with a power to veto most adverse possession applications, an application by a squatter who satisfies one of the three conditions set out in Schedule 6, paragraph 5, will succeed in spite of an objection by the registered owner. The Law Commission felt that in these situations, the balance of fairness lay with the squatter.[2] Two of the conditions were designed to preserve the traditional effect of the doctrine for applicants who went into possession pursuant to an informal transfer.[3] The Law Commission noted that when a dealing takes place "off the register", the applicant does not represent a "land thief" and it would be unjust to allow the registered owner to veto the applicant's registration.[4]

The Law Commission neglected to indicate whether informal transactions in relation to registered land were a widespread phenomenon in England and Wales or if the adverse possession procedure was frequently relied on in such circumstances to update the register.[5] Such information would have helped to contextualise the recognition of this exception to the veto system introduced by the 2002 Act. The purpose of this article is not, however, to second-guess the policy reasons behind the preferential treatment afforded to informal purchasers under the new adverse possession regime. Instead, it examines from a doctrinal perspective whether the possession of the informal purchasers envisaged by Schedule 6, paragraph 5, amounts to adverse possession. It illustrates that the point at which a right of

1 See *Land Registration for the Twenty-First Century: A conveyancing revolution*, Law Comm No 271 (2001), at para. 14.4.

2 Ibid., at para 14.36.

3 The third condition set out in Sch 6, para. 5(4), of the Land Registration Act 2002 facilitates the adjustment of boundaries between neighbours where one neighbour adversely possessed the other's land reasonably believing it to be his or her own throughout the limitation period.

4 See *Land Registration for the Twenty-First Century: A consultative document*, Law Comm No 254 (1998), at para. 10.14.

5 Note that a culture of informal land transactions and a high incidence of abandoned land forced New Zealand and certain Australian territories to re-introduce adverse possession for registered land, see M M Park, *The Effect of Adverse Possession on Part of a Registered Title Land Parcel* (PhD thesis, University of Melbourne, 2003), ch. 6 (available at http://eprints.unimelb.edu.au/).

action accrues against a purchaser in possession pursuant to an oral or a written contract for sale is far from clear. A convincing argument could be made that such purchasers are not in adverse possession. While informal purchasers may be entitled to rely on the doctrine of proprietary estoppel[6] or specific performance to have themselves registered as owners, the conditions set out in Schedule 6, paragraph 5, of the 2002 Act were designed to offer them the more expedient remedy of an adverse possession application. For this option to be of benefit, however, it is essential to eliminate any doubts about whether such informal purchasers are in adverse possession. The article concludes by recommending the introduction of certain legislative amendments which would clarify when such purchasers become entitled to avail of this more expedient remedy.

Schedule 6, paragraph 5(2), of the 2002 Act sets out the first exception to the new veto regime governing adverse possession of registered land. It provides that an applicant is entitled to be registered as the new owner of the estate if it would be unconscionable because of an equity by estoppel for the registered owner to seek to dispossess the applicant and the circumstances are such that the applicant ought to be registered as the owner. In the discussions of this condition in the report which preceded the enactment of the 2002 Act, the Law Commission gave an example of a purchaser who went into possession of land pursuant to an oral contract for sale which failed to comply with requirements set out in section 2 of the Law of Property (Miscellaneous Provisions) Act 1989.[7] The second exception is set out in paragraph 5(3) of Schedule 6 which requires the applicant to prove an entitlement to be registered as owner "for some other reason". This condition was designed by the Law Commission to cater for a purchaser who went into possession pursuant to an enforceable contract for sale.[8] The informal purchasers envisaged by the Law Commission had paid the entire purchase price but were never registered as owners as the necessary steps to complete the transaction had not been taken. The purchaser envisaged by the first condition may be entitled to an equity by estoppel, while the purchaser envisaged by the second condition holds an equitable interest in the land. The Law Commission clearly assumed that the possession of such purchasers amounts to adverse possession which, if maintained for 10 years, extinguishes the title of the registered owner.

However, the informal transactions just described fracture the ownership of land so that legal and equitable interests or equities become distinctly identifiable. Although adverse possession of land subject to fractional interests, such as the interests of co-owners or future owners, has always raised complications, the Law Commission did not discuss, in any detail, the controversy which has arisen over whether the possession of such purchasers can truly be described as adverse to the vendor or the implications of such an approach. This article begins with a discussion of the status of a purchaser in possession pursuant to an enforceable contract for sale who can, therefore, be described as the equitable owner. The position of a purchaser in possession pursuant to an oral contract for sale who possesses an equity by estoppel is examined in the second part of the article.

6 Although, as will be demonstrated in Part 2 of this article, informal purchasers may face considerable difficulty in proving that the circumstances give rise to an estoppel in the aftermath of the House of Lords' decision in *Yeoman's Row Management Ltd* v *Cobbe* [2008] UKHL 55.

7 Law Comm No 271, n. 1 above, at para 14.42.

8 Ibid., at para. 14.43. The Law Commission also gave an example of a claimant who is entitled to the land under the will or intestacy of the deceased registered owner.

Part 1 – Adverse possession by a purchaser entitled to equitable ownership

The Law Commission reiterates the wisdom accepted by many modern land law textbooks that while the squatter–buyer who has paid the purchase price is a beneficiary under a bare trust and can be in adverse possession, a buyer who has not paid the whole of the purchase price will not be in adverse possession as his or her possession is attributable to the contract.[9] The authorities typically cited for this proposition are *Bridges* v *Mees*[10] and *Hyde* v *Pearce*.[11] The difficulties with such an approach stem from the absence of any provision in the Limitation Act 1980 which explicitly sets out that the limitation period may run in favour of a purchaser and against a vendor who holds the property on constructive trust.[12]

A few rules can be deduced from the provisions of the 1980 Act which deal with adverse possession in the context of trust property. Although adverse possession by a stranger against a beneficiary is permitted,[13] adverse possession by a trustee[14] or by another beneficiary[15] in possession of the trust property against a beneficiary is prohibited. The courts have been forced to extrapolate from these provisions when deciding whether a purchaser can be in adverse possession against a vendor who holds the property on constructive trust. One other rule has influenced the reasoning of the courts on this issue and, although it no longer appears in the 1980 Act or the Limitation (Northern Ireland) Order 1989, it has been retained in the Irish Statute of Limitations 1957. It was originally set out in section 7 of the Real Property Limitation Act 1833 which provided that where a person was in possession of land as a tenant at will, the right of the owner to bring an action to recover such land would be deemed to have accrued at the determination of such tenancy or the expiration of one year after the commencement of the tenancy. Section 7 included a proviso which set out that a beneficiary would not be deemed to be a tenant at will to his trustee within the meaning of that section.

It is important to understand the reasons for the inclusion of section 7 in the 1833 Act. Before the enactment of the 1833 Act, the Limitation Act 1623 governed the limitation of actions and the courts supplemented it by developing technical rules to determine whether possession was adverse. These rules frequently made it difficult to identify when a right of action had accrued. For example, there had to be something in the nature of an "ouster" which would put the true owner on notice that time was running against him. Another rule treated the possession of a person with the consent of the owner as under a tenancy at will and incapable of amounting to adverse possession. In addition, once possession commenced lawfully it could never become adverse. *The First Report of Commissioners on Real Property* published in 1829 criticised some of these common law rules,[16] in particular the

9 Law Comm No. 271, n. 1 above, at para. 14.43. See C Harpum, S Bridge and M Dixon (eds), *Megarry and Wade: The Law of Real Property* 6th edn (London: Thomson Sweet & Maxwell 2000), at para. 21.040; K Gray and S F Gray, *Elements of Land Law* 4th edn (Oxford: OUP 2005), at para. 6.63.

10 [1957] Ch 475.

11 [1982] 1 WLR 560.

12 The same is true of the Limitation (Northern Ireland) Order 1989 and the Irish Statute of Limitations 1957.

13 S. 18(1) of the 1980 Act provides, subject to s. 21, the provisions of the Act shall apply to equitable interests in the land as they apply to legal estates.

14 S. 21(1) of the 1980 Act provides that no period of limitation shall apply to an action by a beneficiary under a trust to recover trust property from the trustee.

15 Sch. 1, para. 9, of the 1980 Act provides that where any land subject to a trust is in the possession of a person entitled to a beneficial interest in the land (not being a person solely or absolutely entitled to the land), no right of action shall be treated as accruing during that possession to the trustee or any other beneficiary.

16 They commented that certain rules were of questionable expediency and greatly impaired the healing tendency of the Statutes of Limitation, at p. 47.

rule which prevented possession which began rightfully from maturing into adverse possession. It concluded that a finding of adverse possession should be possible once the rightful estate of the party had been determined.[17] However, in the case of a tenancy, at will it was frequently difficult to determine whether or when the tenancy had been determined by the owner. Section 7 of the 1833 Act provided clarity on this issue by deeming the right of action to accrue one year after the tenancy commenced, unless it had already been determined. In *Ramnarace* v *Lutchman*[18] Lord Millet explained the rationale behind the introduction of section 7:

> It was the deliberate policy of the legislature that the title of owners who allowed others to remain in possession of their land for many years with their consent but without paying rent or acknowledging their title should eventually be extinguished.[19]

This policy was also reflected in section 8 of the 1833 Act which provided that where land is subject to an oral periodic tenancy, the right of action of the owner shall be deemed to have accrued at the expiry of the first period or, if rent was subsequently received, on the date rent was last received. A clear distinction was drawn between the possession of tenants under such informal arrangements and that of a tenant in possession pursuant to a lease for a fixed term. Section 3 of the 1833 Act provided that time only ran against the landlord entitled to the fee simple reversion when he became entitled to his estate in possession.[20]

The proviso to section 7 was deemed necessary as the tenancy at will played an important role in classifying possession which was not authorised by the legal ownership at a time when the common law and equity were administered by separate courts. At common law, a trustee was entitled to possession, but the possession of a beneficiary was frequently authorised by the trust deed or the trustee. Such possession was classified by the common law as being held pursuant to a tenancy at will. The proviso clarified that time did not begin to run against a trustee one year after the commencement of the beneficiary's possession. Some of the conflict in the caselaw revolves around whether the proviso only applied to express trusts, so that time would start to run against a vendor who held the property on constructive trust one year after the purchaser went into possession. The proviso was omitted from the corresponding section in the Limitation Act 1939[21] and in the 1980 Act the entire provision dealing with tenancies at will was repealed. The repercussions of these reforms for a purchaser–beneficiary in possession have yet to be fully teased out.

The early caselaw

One of the chief difficulties with the Law Commission's endorsement of the approach taken in *Bridges* v *Mees* is that it directly contradicts earlier caselaw on this issue, in particular *Drummond* v *Scant*[22] and *Warren* v *Murray*.[23] In Drummond, four brothers had entered into an agreement for a building lease for 99 years in relation to a plot of land next to the Thames. An Act of Parliament was passed to reclaim land from the Thames and vested it in the owners of the land on its banks in accordance with their respective interests. The case concerned the reclaimed land which the court was satisfied became vested in the owners

17 *The First Report of Commissioners on Real Property* (1829).

18 [2001] 1 WLR 1651.

19 Ibid., at para. 12.

20 Time could only run against the landlord during the currency of the lease if the rent was paid to the wrong landlord or if the landlord was also entitled to the lessee's interest in the property.

21 S. 9(1) of the Limitation Act 1939.

22 (1871) LR 6 QB 763.

23 [1894] 2 QB 648, CA.

subject to the equitable interest of the brothers pursuant to the Act. While leases were executed in favour of the brothers over the houses that were built, no lease was ever demanded in respect of the reclaimed land. After the 99 years had expired the owners claimed possession from the defendants who were the brothers' successors in title and argued that the owners' title had been extinguished by adverse possession.

The court noted that before the 1833 Act was passed there was no possibility that the possession of the brothers during the term could be considered adverse. Time could not begin to run against a fee simple reversioner during an equitable term anymore than it could during a legal term. The defendants argued that the 1833 Act did away with the doctrine of non-adverse possession and consequently, after the transitional period of five years, the title of the owners was barred. This argument presumably depended on the brothers being classified as tenants at will with the result that their possession became adverse one year after the tenancy commenced. The court rejected this argument stating that the legislature could not have intended time to run against the owners during the 99-year term when they could not have interfered with the possession of the defendants without risking an injunction or an order for damages for a breach of trust.[24] The court noted that it was bound by the decision in *Garrand* v *Tuck*[25] where it was held that although a beneficiary in possession is deemed to be a tenant at will to his trustee, the proviso to section 7 means that the trustee's estate is not destroyed by the mere lapse of time.[26] The final argument made by the defendants was that the proviso at the end of section 7 applied only to express trusts and had no application to the constructive trust which arises between a vendor and a purchaser. The defendants relied on *Doe d. Stanway* v *Rock*[27] in support of this point. The court was of the opinion that the Stanway case did not involve adverse possession by a contracting purchaser against a vendor, rather it involved a squatter in adverse possession against a contracting purchaser and a vendor and so it could not be cited as an authority on whether the proviso only applies to express trusts. However, the court reached the peculiar conclusion that, even if the proviso was limited in such a manner, the agreement between the brothers and the owners of the fee simple constituted an "actual direct trust", a term which it seems to use interchangeably for an express trust.[28]

The decision in *Warren* v *Murray*[29] includes a more detailed discussion of the law and therefore sheds slightly more light on the area. This dispute also involved an agreement for a building lease for 99 years. The builders went into possession and, although the covenant to build two houses was complied with, no lease was ever asked for or granted. The interest of the builders under the agreement was assigned to the plaintiff's father and he and the plaintiff maintained possession until the term expired and the defendants took possession of the houses. The plaintiff, who contended that the defendants' title had been extinguished

24 (1871) LR 6 QB 763, at p. 767.

25 (1849) 8 CB 231. A mortgage had been redeemed but not cancelled which results in the legal title being held by the mortgagee on trust for the mortgagor. This case involved a claim by a mortgagor in possession that time ran against the mortgagee in respect of the satisfied term.

26 The probable rationale behind the exception, as explained by the court in *Garrand* v *Tuck*, was to avoid the need for a trustee to take active steps to preserve his estate from being destroyed.

27 (1842) Car & M 549. The facts of this case are a little confusing – following an agreement to purchase, the purchaser was left into possession and subsequently agreed to assign his equitable interest to a sub-purchaser who paid the original vendor. The original purchaser remained in possession until he died when his widow and subsequently one of her children, the defendant, took over possession. The plaintiff acquired the legal interest of the vendor and the equitable interest of the sub-purchaser and the court ruled that his interest had been extinguished by the possession of the original purchaser who held as a tenant at will. As the proviso only applied to express trusts, time ran from the anniversary of the commencement of the tenancy.

28 (1871) LR 6 QB 763, at pp. 768–9.

29 [1894] 2 QB 648.

by virtue of the statute of limitations by the time they took possession, lost his case. Lord Esher MR applied *Drummond* v *Scant* and concluded that if the defendants were unable to recover the land in question during the term of the lease, the Statute of Limitations had not run against them. He explained:

> Looking at the agreement, it seems to me clear that although at common law they might say that the tenants were mere tenants at will, yet according to the law as a whole, including the equity doctrines applicable to this case, they (the trustees) could not exercise their will and re-enter, because, if the tenants had performed their obligations under the agreement, a court of equity would at once interfere by injunction to prevent their being dispossessed and would compel specific performance of the agreement.[30]

He was of the opinion that section 7 of the 1833 Act applied to "tenancies at will pure and simple, where there was no clog or difficulty such as arises out of an agreement like that in question here". Kay LJ added that the circumstances of the plaintiff were in any case covered by the proviso to section 7. He noted that, where a purchaser enters into an agreement for sale, a constructive trust arises regardless of whether the agreement is to purchase a fee simple or a leasehold term. He was of the opinion that the proviso encompassed such trusts which meant that a right of action did not automatically accrue to the vendor one year after the purchaser went into possession.[31]

In both *Warren* and *Drummond* the court seemed to assume that, if adverse possession could be proved, the purchaser would have extinguished the title of the fee simple owner. However, in both situations the purchaser was entitled to an equitable lease, which cannot exist without an equitable fee simple reversion. A right of action could not have accrued against a legal fee simple reversioner during the currency of the lease and if, as the 1833 Act sets out, its provisions apply to equitable interests as they apply to legal estates,[32] time could not run against the equitable fee simple reversioner until the determination of the equitable lease. The court in *Drummond* only referred to this point in passing[33] and it would seem that it was not argued at all in the *Warren* case. It is submitted that if adverse possession can take place in such circumstances it must operate solely to bar the legal title to the lease which remains vested in the vendor pursuant to the contract for sale. Time cannot run in relation to the fee simple estate until the legal or equitable lease has expired.

Warren and *Drummond* were discussed in two New Zealand cases reported at the beginning of the twentieth century.[34] The English cases were distinguished by the court in *Glenny* v *Rathbone* which was satisfied that a sub-purchaser was in adverse possession against the original vendor but, in *Ormond* v *Portas*, they were followed and the court ruled that the defendant, who was in possession pursuant to an agreement for a lease, was not in adverse possession against the owner. The judgments delivered in both New Zealand cases grappled with the distinction between the position of a vendor who has agreed to sell the fee simple and the vendor who has only agreed to the grant of a lease in relation to the property. In *Glenny* v *Rathbone*, Williams J stated that if this had been a case of a straightforward agreement to purchase a fee simple, section 7 would apply as the purchaser would be deemed to be a tenant at will of the vendor. However, he felt that the vendor would not have been a trustee within the meaning of the proviso to that section. The meaning of the term "trustee" in the context of the proviso should, he said, be limited to cases where the terms of the trust

30 *Warren* v *Murray* [1894] 2 QB 648, at pp. 652–3.

31 Ibid., at p. 658.

32 S. 24.

33 (1871) LR 6 QB 763, at p. 767.

34 *Glenny* v *Rathbone* (1900) 20 NZLR 1; *Ormond* v *Portas* [1922] NZLR 570.

contemplate the retention of the legal fee simple by the trustee during the suggested period of limitation. He noted that where a lessee enters into possession under an agreement for a lease for 99 years, he or she has an equitable lease for that period. The owner of the legal fee simple holds that land on trust to grant such a lease to the intending lessee throughout the agreed term. The lessee's possession is entirely consistent with the right of the lessor to retain the legal fee simple during the whole proposed term. He felt that *Drummond* and *Warren* could be explained on the basis that it was necessary for the trustee to retain the legal fee simple in order to carry out the trust. Therefore, the proviso applied to prevent time running under section 7. However, where an agreement to purchase the fee simple is involved and the purchaser has paid the purchase money and been let into possession, the continued retention by the vendor of the legal fee simple is directly antagonistic to the possession of the purchaser and so time can run in favour of the purchaser.[35] Williams J also discussed the *dicta* in *Warren* and *Drummond* which suggests that time cannot run against the owner if the owner would have been prevented from recovering possession of the land by the existence of some equitable remedy. He felt that these comments could not be taken in their widest sense as a true exposition of the law and were certainly not necessary upon the facts of those cases to arrive at the decision made.[36] In *Ormond* v *Portas*, the court noted that *Glenny* v *Rathbone* did not weaken the effect of the judgment in *Warren* as applied to the case of a lessee in possession under an agreement for lease.[37]

These New Zealand cases imply that although a purchaser in possession under a contract for a sale of the fee simple may bar the legal estate of the vendor, a purchaser under an agreement for a lease may not do so. As already mentioned, the Law Commission of England and Wales is of the opinion that a purchaser who is in possession who has not paid the entirety of the price will not be in adverse possession as such possession is attributable to the contract and not his or her absolute equitable interest. The Law Commission noted that, were it otherwise, the validity of agreements for leases as a substitute for legal leases would be undermined.[38] It is difficult to follow the reasoning of the Law Commission in this regard; perhaps it viewed the payment of rent over the term of the lease as making up the purchase price under the contract. However, such an approach would prevent an adverse possession application by a purchaser who had paid the purchase price and gone into possession pursuant to an agreement to purchase a registered leasehold estate. It is difficult to see why such an applicant should not benefit from Schedule 6, paragraph 5(3) of the 2002 Act, regardless of whether rent was paid during the term. The landlord has remedies for the non-payment of rent which he or she may avail of both before and after the statutory transfer of title from the registered lessee to the purchaser–squatter on proof of 10 years' adverse possession. However, where a registered freehold or leasehold owner enters into an agreement for a lease or a sub-lease, paragraph 5(3) does not apply as the lease has not as yet been granted or registered. Therefore, the old

35 *Glenny* v *Rathbone* (1900) 20 NZLR 1, at pp. 28–9. Williams J ruled that the plaintiff–sub-purchaser was not a tenant at will to the original vendor. There was no evidence that the original vendor knew that he had purchased the property or had given any recognition of his possession which could give rise to the inference that it was the intention of both parties to create such a tenancy. Therefore, s. 7 of the 1833 Act was inapplicable and it was necessary for the plaintiff to show that a right of action had accrued within the meaning of s. 2 (at pp. 30–1). He was satisfied that a right of action had accrued at law and the existence of some equitable remedy did not preclude a finding of adverse possession.

36 Ibid., at pp. 31–3. Even if the possessor did have a remedy in equity, Williams J noted that the owner had a hostile right at law. The object of the 1833 Act was to get rid of hostile rights. If an entire right can be barred (where there is no entitlement to an equitable remedy) why not a part of that right which is hostile to the rights of the person in possession?

37 [1922] NZLR 570, at p. 580.

38 See Law Comm No 271, n. 1 above, at para 14.43, n. 155.

rules which govern adverse possession of unregistered leasehold land set out in *Fairweather* v *St Marylebone Property Ltd*[39] will apply. It was held in that case that the squatter does not acquire the title of the lessee on the expiry of the limitation period. The original lessee remains liable on the covenants in the lease and, although he may no longer eject the squatter he may surrender the lease to the landlord who will become entitled to immediate possession. If the owner holds the land on a constructive trust to grant a lease to the purchaser and the legal title to the lease is extinguished, the possible ramifications of the *Fairweather* decision for the vendor and the purchaser become ridiculously complicated. The vendor clearly retains an equitable fee simple reversion and time cannot run in respect of this estate until the expiry of the term of the equitable lease. However, the extinguishment of the legal title to the lease must surely go hand in hand with the extinguishment of any rights which the vendor had pursuant to the contract. It is arguable that adverse possession in such circumstances operates to deprive the vendor of the right to enforce the covenants which the parties had agreed to include in the lease. In contrast, possession for the limitation period may confer the best of both worlds on the person who had agreed to purchase the lease: that person could claim the benefit of certain covenants under his or her equitable lease when it suited and retain the option of relying on the title acquired through adverse possession which seems to be free of such covenants.

An approach which precludes the running of time against a vendor who holds under a constructive trust to grant a lease is to be preferred, regardless of whether the purchaser has paid the entire purchase price. Unlike an agreement to purchase a fee simple or take an assignment of an existing lease, which only gives rise to a bare trust, an agreement to grant a lease gives rise to a constructive trust which envisages an active role being played by the vendor–landlord during the currency of the equitable lease. Such an approach accords well with the provisions of the legislation which prevent the extinguishment of a trustee's estate if a right of action of any person entitled to a beneficial interest in the land either has not accrued or has not been barred.[40] The equitable fee simple reversion implicit in an agreement to grant a lease should be viewed as preventing time running against the vendor during the term of the equitable lease. Such a purchaser who wishes to regularise his or her occupation should be forced to rely on the remedy of specific performance which will ensure that the vendor is guaranteed the protection of the covenants the parties agreed to include in the lease.

More recent caselaw

As already mentioned, *Bridges* v *Mees*[41] is the authority typically cited in support of the proposition that a purchaser under an incomplete contract for sale can extinguish the title of the vendor through adverse possession. The plaintiff and the defendant owned neighbouring houses and the plaintiff had entered into an oral contract to purchase a small piece of land at the rear of both houses from a company. The purchaser went into possession after paying the deposit and by 1937 he had paid the entire purchase price. In 1955 the defendant purchased the same piece of land from the liquidator of the company and proceeded to register his ownership in the Land Registry. The plaintiff sought a declaration that he was the beneficial owner of the land and a rectification of the register to reflect his overriding interests which arose due to his adverse possession and his actual occupation of the land.

39 [1963] AC 510.

40 See s. 18(3) of the 1980 Act, previously s. 7(3) of the 1833 Act.

41 [1957] 1 Ch 475.

Harman J was satisfied that the purchaser initially went into possession of the property with the vendor's permission and pursuant to a licence. However, when the vendor's lien on the property for the unpaid purchase monies disappeared, the character of the purchaser's possession changed. At that point, the vendor became a bare trustee and the purchaser became the sole beneficial owner. The vendor, as trustee, was prima facie entitled to resume possession and as he did not exercise that right for 12 years the plaintiff argued that it was extinguished and he was required to hold the legal estate on trust for the plaintiff.[42]

In discussing whether a purchaser was a person in whose favour the period of limitation could run, Harmon J referred to section 7(3) of the Limitation Act 1939, which provided:

> Where any land is held upon a trust . . . and the period prescribed by this Act has expired for bringing an action to recover the land by the trustees, the estate of the trustees shall not be extinguished if and so long as the right of action of any person entitled to a beneficial interest in the land . . . has not accrued or been barred by this Act, but if and when every such right of action has been so barred, the estate of the trustees shall be extinguished.

As no one other than the purchaser had a beneficial interest from 1937 onwards, Harmon J concluded that time could and did run in his favour and therefore by 1949 the trustee's title would have been extinguished (but for section 75 of the Land Registration Act 1925).[43] Harmon J examined the argument which had swayed the court in *Drummond* and *Warren*, that the vendor could never have brought an effective action to recover the land because he would have been met by the plea that the whole beneficial interest was vested in the purchaser. Harmon J ruled that the question cannot turn on whether the action would have succeeded or not. He cited *Re Cussons Ltd*[44] in support of this approach even though counsel for the plaintiff had acknowledged that this case had been criticised because of the court's failure to refer to the proviso in section 7 of the 1833 Act.[45] Harmon J noted that since the proviso was omitted from section 9 of the 1939 Act, time can now run in favour of a beneficiary. He quoted *Underhill on Trusts* who stated that a trustee, including a constructive trustee, can be divested of a legal estate by possession of a person entitled in equity in exactly the same way as if the beneficiary were a stranger.[46] He acknowledged that no precise authority was given by Underhill but he felt that this proposition was implicit in section 7(3) (outlined above) and section 7(5) which provided:

> Where any settled land or any land held on a trust for sale is in the possession of a person entitled to a beneficial interest in the land . . . not being a person solely and absolutely entitled thereto, no right of action shall be deemed . . . to accrue during such possession to any person in whom the land is vested as trustee, or to any other person entitled to a beneficial interest in the land . . .

Basically, Harmon J felt that it was implicit in both subsections that a person solely and absolutely entitled to the beneficial ownership could be in adverse possession against his or her trustee.

---

42 *Bridges* v *Mees* [1957] 1 Ch 475, at pp. 484–5.

43 Ibid.

44 (1904) 73 LJ Ch 296. In this case, partners who had incorporated neglected to transfer the disputed property into the name of the company. The court was satisfied that the partners held the property on a bare trust for the company and that it was possible that such trustees without duties may lose their trust estate at the end of 12 years if they allowed the beneficiary to remain in possession and did not interfere.

45 [1957] 1 Ch 475, at p. 482.

46 Ibid., at p. 486.

Harmon J noted that even if he was wrong about his interpretation of section 7(3) and (5), the plaintiff had another string to his bow. A plaintiff was entitled to a rectification of the register on the basis that his or her interest under the contract coupled with the plaintiff's actual occupation of the property rendered it an overriding interest which bound the defendant on registration.[47] It is submitted that the second reason which Harmon J gave for his decision is less open to challenge, particularly since the introduction of the Limitation Act 1980. In its 1977 Report, the Law Commission noted that time only ran against a licensor on the determination of the licence and agreed with those who had commented on this issue that the divergent treatment of tenancies at will and licences was artificial and unduly harsh on those categorised as lessors.[48] The Law Commission was convinced that the distinction between a tenancy at will and a gratuitous licence is, at best, tenuous. Accordingly, under the 1980 Act, time does not begin to run in favour of the occupier until the tenancy has actually been determined. Therefore, if the occupation of a purchaser continues to be classified as that of a tenant at will, the determination of the tenancy would seem to be a prerequisite to establishing adverse possession.

Jourdan describes Harmon J's decision in *Bridges* v *Mees* on the adverse possession point as questionable,[49] but continues:

> The question whether the decision was correct is, however, now largely academic. The decision was based on section 9(1) of the Limitation Act which provided that time ran against a landlord of a tenant at will from the first anniversary of the grant of the tenancy. That subsection was repealed with effect from 1 August 1980. Accordingly, if the same case fell to be decided now, a different result would be reached.[50]

The status of a purchaser in possession was considered more recently by the Court of Appeal in *Hyde* v *Pearce*.[51] This decision is frequently cited to support the proposition that a purchaser who has not paid the entire purchase price cannot be in adverse possession. The purchaser went into possession after making a successful bid at an auction and paying the deposit. The completion of the sale was delayed as a dispute arose over the level of the abatement in the purchase price which the purchaser was entitled to because a small portion of the property which he had contracted to buy had already been sold. The court concluded that although his initial possession was with the licence of the vendor, that licence was determined by a letter sent in 1958 which demanded that that the purchaser return the keys.[52] The court was satisfied that a right of action had accrued to the vendor at that time. However, Templeman LJ continued:

> . . . in the peculiar circumstances of this case, it seems to me that it is not sufficient to show that a right of action had accrued. Mr Hyde must show some further quality, namely adverse possession.[53]

This was an unusual approach as section 10 of the Limitation Act 1939 specifically provided that no right of action to recover land shall be treated as accruing unless someone was in adverse possession.[54] It is not clear why the court decided to determine these issues

47 *Bridges* v *Mees* [1957] 1 Ch 475, at pp. 486–7.

48 Law Reform Committee, *Twenty-First Report* (Final Report on Limitation of Actions) (September 1977), at para. 3.55.

49 S Jourdan, *Adverse Possession* (London: Butterworths Lexis Nexis 2003), at para. 28.29.

50 Ibid., at para. 28.30.

51 [1982] 1 WLR 560.

52 Ibid., at p. 569.

53 Ibid., at p. 570.

54 See also M Dockray, "What is adverse possession: Hyde and seek" (1983) 46 *MLR* 89, at p. 92.

separately. The court pointed to a number of factors which illustrated that the purchaser was not in adverse possession. It noted that Mr Hyde was not a squatter without a shadow of a claim of right – he was the equitable owner pursuant to the contract, subject to the vendor's lien for the price. Also, his possession was attributable to a subsisting contract for sale. He had never changed his status as a purchaser in possession pending completion by doing something which showed that he repudiated the contract. The court also argued that litigation by the vendors may not have resulted in the vendors obtaining possession. Finally, the court referred to the final letter which had been sent by the vendor's solicitor to Mr Hyde proposing that the dispute over the purchase price be referred for arbitration. The court argued that this letter was equivocal and did not make it clear whether the vendors were still requiring possession to be handed over to them.[55]

It is submitted that these were all issues that should have been considered in assessing whether a cause of action had accrued to the vendor. Either the purchaser was in possession pursuant to a licence, in which case no cause of action had accrued, or his licence had determined, in which case a cause of action had accrued. If you apply the reasoning in *Bridges* v *Mees*, the purchaser's licence would have automatically terminated if the entire purchase price had been paid. The repudiation of the contract would also have ended the licence and, in this particular case, the parties had agreed that the licence could be determined by a demand for the keys. If the letter demanding the keys terminated the licence, a right of action would have accrued to the vendor at that point and time must have run against him regardless of whether the contract continued to subsist or not. On balance, it seems that Mr Hyde should have succeeded in his adverse possession claim and his rights should, therefore, have bound Mr Pearce as an overriding interest when he bought the property from the vendors. Even if the court had ruled that Mr Hyde was not in adverse possession, he should have succeeded on the basis that his interest under the contract coupled with his actual occupation of the property amounted to an overriding interest. The court ruled that Mr Pearce was not bound by the contract as it had not been registered as an estate contract but these rules only apply to unregistered land.[56] It is submitted that *Hyde* v *Pearce* should not be interpreted to mean that time can only run against the vendor if the entire purchase price has been paid. A purchaser can also be in adverse possession if his or her licence to occupy the premises is validly terminated and, in such circumstances, his or her status as a purchaser under an incomplete contract is irrelevant.

McLean v *McErlean*[57] provides a useful illustration of the attitude of the Northern Irish courts to this issue although the removal of the tenancy at will provision in the Limitation (Northern Ireland) Order 1989 may cast doubt on its continuing relevance. The purchasers in this case were sand merchants who had entered into a contract to purchase land from the defendant's father with the sole purpose of extracting and selling sand for use in building work. They went into possession with the consent of the vendor and, by 1958, they had paid the entire purchase price. They extracted sand on the land until it ran out sometime in 1964 or 1965 and afterwards they occasionally used the land for washing sand. After the purchasers had contracted to buy the land, the vendor continued to use it for grazing and used a small amount of it for cropping. Neither of these activities interfered with the purchasers' use of the land. After the vendor died, his executor commenced an action in trespass and the purchasers applied for an order that they were entitled to be registered as owners of the land by virtue of their adverse possession. The court had to decide whether the purchasers had barred the legal title of the vendor, or the vendor had barred the

55 [1982] 1 WLR 560, at pp. 569–70.

56 Dockray has also commented on this in "What is adverse possession?", n. 54 above.

57 [1983] NI 258.

beneficial title of the purchasers. The Court of Appeal found in favour of the purchasers and Gibson J noted two differences between the Northern Irish Statute of Limitations 1958 and the English Limitation Act 1939. Firstly, the Northern Irish version of the tenancy at will provision included a proviso that a beneficiary shall not be deemed to be a tenant at will to his or her trustee for the purposes of that subsection.[58] However, this proviso was not included in the 1939 English version of the tenancy at will provision.[59] The second distinction is replicated in current legislation.[60] The definition of a "trustee" provided in the Northern Irish version is limited to an express trustee and does not include a person whose fiduciary relationship arises merely by construction or implication of the law.[61] In contrast, all references to trusts in the English version include constructive or implied trusts.[62] The Irish Statute of Limitations 1957 is identical to the Northern Irish legislation in these respects.[63] Gibson J concluded that, where the trust in question is constructive, a person entitled to the beneficial estate would begin to run a title after the first year. He stated that it was obvious why an express trustee should not have to take active steps to preserve his estate but noted that these considerations do not apply where the trust in question is constructive, for example where a purchaser has been let into possession and has paid the purchase price. The vendor is a bare trustee and his only duty is to convey and therefore there is no need to preserve his estate in order to allow him to perform the trust. The purchaser could and did in such circumstances commence to run a statutory title against the vendor. He also referred to section 29(4) of the 1958 Act (which corresponds to section 7 (5) of the Limitation Act 1939) to support his decision and implies that a right of action may accrue to a trustee if the beneficiary is solely and absolutely entitled.

Finally, it is worth mentioning briefly the recent Irish High Court decision in *Moley* v *Fee*[64] which concerned an agreement to sell two sites for mobile homes. The transaction was never completed and Laffoy J noted that, if the purchasers had paid the full purchase price, the vendor would be deemed by law to be a bare constructive trustee. In such circumstances, if the vendor had remained in possession of the land he would have barred the beneficial interest of the purchasers and if, on the other hand, the purchasers had gone into possession, the outstanding legal estate of the vendor would have been barred. On the facts, Laffoy J was satisfied that the vendor had never received the purchase price from the purchasers. She also noted that although the vendor had continued to exercise acts of ownership over the disputed plot during the period which followed the negotiation of the deal, the purchasers had not engaged in acts of possession sufficient to prove adverse possession during the same period.

58 S. 21 of the Statute of Limitations 1958.

59 S. 9 of the Limitation Act 1939.

60 See s. 2(3) of the Limitation (Northern Ireland) Order 1989 and s. 38 of the Limitation Act 1980.

61 S. 74(2) of the 1958 Act.

62 S. 31 of the Limitation Act 1939.

63 See s. 2(2)(a) of the Statute of Limitations 1957.

64 [2007] IEHC 143.

Is a purchaser in possession or an occupant negotiating a purchase, a licensee or a tenant at will?

As has already been mentioned, in England and Northern Ireland since the enactment of the 1980s legislation dealing with the limitation of actions, a right of action will only accrue on the determination of a licence or a tenancy at will. The distinction between licences and tenancies at will as regards the accrual of a right of action continues to be maintained by the Irish Statute of Limitation 1957. Therefore, in Ireland the success of the purchaser's adverse possession claim may be dependent on whether the purchaser is classified as a tenant at will. In *Bridges* v *Mees* and *McLean* v *McErlean*, the purchaser was classified as a tenant at will,[65] while, in *Hyde* v *Pearce*, he was classified as a licensee.

Traditionally, exclusive possession was treated as the sole distinguishing feature of a tenancy at will. However, for the last 50 years the courts have been prepared to recognise an arrangement that confers exclusive possession on the occupant as a licence if satisfied that this was the intention of the parties. The courts will more readily infer a licence if, in the words of Denning LJ in *Facchini* v *Bryson* "there has been something in the circumstances, such as a family arrangement, and act of friendship or generosity . . . to negative any intention to create a tenancy".[66] For example, in *Heslop* v *Burns*,[67] the court described the arrangement between Mr and Mrs Burns, who had occupied a house for 16 years without paying rent, and Mr Timms, its owner, as akin to a family arrangement. He was a good friend of the Burns, visited them frequently, had become a godfather for one of their daughters and paid for her education. There was no evidence to infer a tenancy at will and, therefore, while the licence continued, a right of action did not accrue to the owner.[68] Scarman LJ noted that the courts will be less and less inclined to infer a tenancy at will from an exclusive occupation of indefinite duration due to the emergence of the licence to occupy into prominence as a possible mode of land-holding.[69] He added:

> It may be that the tenancy at will can now serve only one legal purpose and that is to protect the interests of an occupier during a period of transition. If one looks to the classic cases in which tenancies at will continue to be inferred,

65 In *Glenny* v *Rathbone* (1900) 20 NZLR 1 and *Ormond* v *Portas* [1922] NZLR 570, the court also classified a purchaser under a contract for sale or an agreement for the grant of a lease as a tenant at will, although it was satisfied the proviso to s. 7 of the 1833 Act would prevent time running against the vendor where the agreement was for the grant of a lease.

66 [1952] 1 TLR 1386, at pp. 1389–90.

67 [1974] 3 ALL ER 406.

68 A similar approach had been taken in *Cobb* v *Lane* [1952] 1 All ER 1199 and in *Hughes* v *Griffin* [1969] 1 WLR 23.

69 Much of the caselaw which discusses the distinction between a tenancy at will and a licence involves claims by occupiers that they were entitled to certain statutory protections available to tenants. Frequently, the occupier went into possession pending the negotiation of a lease or a sale or held over possession pending the negotiation of a renewal of a lease. A tenant at will does not qualify for the security of tenure afforded to business tenants pursuant to the Landlord and Tenant Act 1954, see *Wheeler* v *Mercer* [1973] 1 All ER 829. However, it is dangerously easy for such an arrangement to become a periodic tenancy on the payment of rent and such a finding would bring the tenancy within the protection of the 1954 Act. A residential tenancy at will may come within the Rent Restrictions Acts provided that the payments made by the occupant do indeed represent rent and not the payment of the purchase price by instalments: see *Dunthorne and Shore* v *Wiggins and Another* [1943] 1 All ER 577; *Francis Jackson Developments Ltd* v *Stamp* [1943] 2 ALL ER 601. There is an increasing tendency on the parts of the courts to classify occupation pending the negotiation of a formal agreement, particularly those in occupation for commercial purposes, as being pursuant to a licence: see *Baikie* v *Fullerton-Smith and Another* [1961] NZLR 901; *Isaac* v *Hotel De Paris Ltd* [1960] 1 All ER 348. This classification will have no impact on the rights of a business occupier as neither a licensee or a tenant at will is entitled to protection pursuant to the 1954 Act, and, since the enactment of the Housing Act 1988, residential tenants are unlikely to benefit from a reduced rent or increased security of tenure.

> namely, the case of someone who goes into possession prior to a contract of purchase, or of someone who, with the consent of the landlord, holds over after the expiry of his lease, one sees that in each there is a transitional period during which negotiations are being conducted touching the estate or interest in the land that has to be protected, and the tenancy at will is an apt legal mechanism to protect the occupier during such a period of transition; he is there and can keep out trespassers; he is there with the consent of the landlord and can keep out the landlord as long as that consent is maintained.[70]

Ramnarace v *Lutchman*,[71] which involved an appeal from Trinidad and Tobago, shows the implications of classifying such transitional occupation as a tenancy at will. The appellant had entered into possession of the land with the consent of its owners, her uncle and aunt in 1974. They had told her she could live on the land until she could afford to buy it and she built a house and lived there with her family without paying any rent. Her uncle died in 1977 and her aunt died in 1988 and the respondent had periodically challenged her right to live on the land. She claimed to have acquired a possessory title by virtue of section 8 of the Real Property Limitation Ordinance 1940 which provided that time began to run against a tenant at will one year after the commencement of the tenancy. The Court of Appeal found that she was in possession pursuant to a licence which had terminated either in 1985 by the service of a notice to quit or in 1988 on the death of her aunt and therefore she had not succeeded in extinguishing the respondent's title. The Privy Council disagreed with this classification and noted that although the appellant was allowed into occupation as part of a family arrangement and at least in part as an act of generosity, the intention was that she would purchase the land when she could afford it – this was one of the classic circumstances in which a tenancy at will arose. She therefore succeeded in her claim for adverse possession.

In *Bellew* v *Bellew*,[72] the Irish Supreme Court demonstrated a willingness to regard even the occupation of a person in negotiations with the owner as being pursuant to a licence. However, in the circumstances, the classification did not prevent time running as the court was satisfied that the licence had ended. The plaintiff held a life estate in the farmlands surrounding Barmeath Castle, where he lived with his father, his wife and his children. He began an affair and moved to England to live with the woman, neglecting to make any provision for the maintenance of his wife and family. The plaintiff permitted the father to continue farming the land while negotiations were taking place in relation to an agreement for a lease and the provision to be made for plaintiff's family. These negotiations broke down in 1963 and the father remained in occupation and farmed the land as if it was his own. In 1978 the plaintiff sought a declaration that the lands were vested in him as tenant for life but the Supreme Court held that this estate had been extinguished by the father's adverse possession. Griffin J, with whom Hederman J agreed, was of the opinion that the father originally went into possession pursuant to a licence but that the licence terminated when the negotiations broke down and from that point onwards he was in adverse possession. O'Higgins CJ preferred to classify his occupation as a tenancy at will as he went into occupation pending negotiations for a long-term lease. Therefore, a right of action accrued one year after he commenced occupation and the title of the plaintiff would have been extinguished by 1974. He noted that the intricacies of running a large farm with a danger of trespass and the possibility of agistment and other contracts required that the person running such a farm would have some legal interest therein.

70 [1974] 3 All ER 406, at p. 416.

71 [2001] 1 WLR 1651.

72 [1982] IR 447.

It is difficult to see what the tenancy at will can achieve in such circumstances that a licence cannot. Previously, when exclusive possession was the sole determinant of whether an occupant held under a tenancy or a licence, an action in trespass could only be maintained by a tenant. Nowadays, a licensee may maintain such an action. A licensee may also be entitled to contractual notice or reasonable notice to vacate the premises and, therefore, the licensee may even have more security of tenure than a tenant at will. In addition, it is difficult to see how an agistment contract made by a tenant at will would be more secure than one made by a licensee.

It would appear that the tenancy at will no longer plays a distinctive role in property law and these arrangements could easily be absorbed by the licence. It is increasingly difficult to justify the distinction between such tenancies and licences as regards the accrual of a right of action and it is submitted that Ireland should follow the lead of England and Northern Ireland and delete the tenancy at will provision from the Statute of Limitations 1957. This appears even more pressing in light of recent recommendations for the abolition of the tenancy at will.[73]

However, the abolition of the tenancy at will provision does not alleviate the confusion over whether a purchaser under an incomplete contract for sale can be in adverse possession. It is important to bear in mind the consequences of recognising adverse possession in such circumstances. If the vendor's title is extinguished, presumably he or she no longer has any rights pursuant to the contract and so will no longer be entitled to the benefit of easements which were to be reserved or restrictive covenants which were to be imposed on the purchaser in the deed. By neglecting to formalise the transfer, he or she will have lost the rights that would have been conferred by it. The remaining difficulty is identifying the point at which the purchaser's possession becomes adverse. This issue could be cleared up by inserting a "deeming provision" – where a purchaser is allowed into possession before the completion of the transaction, a right of action shall be deemed to accrue to the vendor on the determination of the licence to occupy or the payment of the entire purchase price, whichever occurs earlier. However, the definition of a "purchaser" for the purposes of benefiting from the deeming provision should exclude a purchaser under an enforceable agreement for the grant of a lease. As mentioned earlier, such a purchaser should be forced to rely on the remedy of specific performance so that the landlord is not deprived of the benefit of the agreed covenants.

Part 2 – Adverse possession by a purchaser entitled to an equity by estoppel

The Land Registration Act 2002 provides that where it would be unconscionable because of an equity by estoppel for the registered owner to seek to dispossess the applicant and that applicant is entitled to be registered as owner, an application for registration on the basis of adverse possession will succeed in spite of an objection by the registered owner.[74] The adjudicator has jurisdiction to consider the elements of the proprietary estoppel claim and is expressly authorised to award a lesser remedy if it would be unconscionable for the registered owner to seek to dispossess the applicant but the circumstances are not such that the applicant ought to be registered as owner.[75]

Since the enactment of section 2 of the Law of Property (Miscellaneous Provisions) Act 1989, contracts for the sale of land must be made in writing and incorporate all the

73 See Law Reform Commission, *Consultation Paper on General Law of Landlord and Tenant Law* (LRC CP28-2003), at para. 1.24. A "tenancy" is defined as not including a tenancy at will in s. 3 of the draft Landlord and Tenant Bill included in the *Report on the Law of Landlord and Tenant* (LRC 85-2007).

74 Sch. 6, para. 5(2).

75 S. 110(4) of the 2002 Act.

terms expressly agreed in one document which has been signed by both parties. Although the doctrine of part performance was not specifically abolished, the fact that an oral contract is no longer valid renders the doctrine defunct as there is no contractual obligation which can be partly performed. The Law Commission, when recommending the reforms introduced by section 2, noted that the present law provided sufficient alternative remedies to deal with the hard cases which may arise where one party unconscionably seeks to take advantage of a failure to comply with the statutory formalities.[76] In particular, the Law Commission envisaged that the doctrine of proprietary estoppel would step into the breach and provide a remedy for certain purchasers who had entered into informal contracts for sale.[77] This doctrine applies where a person represented to the claimant that he or she had rights in the land and the claimant acted to his or her detriment in reliance on this representation in circumstances where it would be unconscionable to allow the representor to insist on the strict legal position. Until such an equity has been established to the satisfaction of the court, the claimant has only an inchoate right; once it has been established the court has a broad discretion as to how to satisfy the equity. It may order the landowner to convey the freehold or some other right to the claimant,[78] it may order the payment of compensation,[79] or simply make an order restraining the owner from enforcing his or her strict legal rights.[80]

The application of this doctrine to invalid contracts for sale has proved very controversial and points up a failure on the part of the Law Commission and Parliament, when proposing and enacting section 2 of the 1989 Act, "to consider either adequately or in sufficient detail which potentially hard cases can be effectively dealt with by existing doctrines".[81] The main sticking point is whether reliance on the doctrine permits a claimant to bypass the formalities set out in section 2 of the 1989 Act. Some of the caselaw[82] requires the claimant to prove what Dixon refers to as a "double assurance",[83] for example, a representation by the defendant that he is a man of his word or that the agreement is binding in honour. This approach enables the court to describe the doctrine in terms of remedying the unconscionability caused by this assurance rather than facilitating the enforcement of an invalid contract.[84] Another development is the increasing tendency on the part of the courts to restrict the remedy of proprietary estoppel in cases involving invalid contracts to situations which also give rise to a constructive trust.[85] This limitation is seen as necessary to square the enforcement of the contract with the failure to comply with the formalities imposed by section 2 of the 1989 Act, as section 2(5) provides that nothing in this section affects the creation or operation of resulting, implied or constructive trusts. Some have criticised this approach as unnecessary, as a remedy based on the doctrine of proprietary estoppel is an independent cause of action which does not seek to enforce

76 *Formalities for Contracts for Sale etc of Land*, Law Comm No 164 (1987), at para. 4.13.

77 Ibid., at part 5.

78 See *Pascoe* v *Turner* [1979] 1 WLR 431; *Crabb* v *Arun DC* [1976] Ch 179.

79 See *Dodsworth* v *Dodsworth* (1973) 228 EG 1115.80

80 See *Maharaj* v *Chand* [1986] AC 898.

81 C Davis, "Estoppel: an adequate substitute for part performance?" (1993) 13 *OJLS* 99.

82 See *Lloyd* v *Dugdale* [2002] 2 P & CR 13 and Etherton J's decision in *Cobbe* v *Yeoman's Row Management Ltd* [2005] EWHC 266.

83 M Dixon, "Proprietary estoppel and formalities in land law and the Land Registration Act 2002: a theory of unconscionability" in E Cooke (ed.), *Modern Studies in Property Law* vol. 3 (Oxford: Hart Publishing 2003); M Dixon, "Invalid contracts, estoppel and constructive trusts" (2005) *Conv* 247.

84 See *Cobbe* v *Yeoman's Row Management Ltd* [2005] EWHC 266, at p. 165.

85 See *Yaxley* v *Gotts* [2002] Ch 162; *Kinane* v *Kackie-Conteh* [2005] EWCA Civ 45.

the invalid contract but rather seeks to satisfy the equity.[86] McFarlane, in an article published before the delivery of the House of Lords' decision in *Yeoman's Row Management Ltd* v *Cobbe*,[87] noted that the courts have reacted to this limitation of the doctrine by taking a very lenient approach to the establishment of a constructive trust.[88]

The recent House of Lords' decision in *Yeoman's Row Management Ltd* v *Cobbe*[89] creates the most serious impediments for informal purchasers seeking a remedy under the rubric of propriety estoppel. An oral agreement in principle had been reached that the defendant would sell the claimant a property and, in return, the claimant undertook to obtain planning permission and invested considerable time and money in doing so. When planning permission had been obtained, the defendant sought to withdraw from the agreement relying on non-compliance with section 2 of the 1989 Act. Although Etherton J and the Court of Appeal awarded a remedy pursuant to the doctrine of proprietary estoppel, the House of Lords felt that they had stretched the boundaries of the doctrine too far[90] and chose instead to provide a remedy in restitution. The House of Lords ruled that, where an essential element of the doctrine of proprietary estoppel could not be proved by the claimant, a finding of unconscionable behaviour on the part of the defendant was not sufficient to warrant reliance on the doctrine.[91] If proof of the essential elements of the doctrine were not required, Lord Scott claimed "proprietary estoppel will lose contact with its roots and risk becoming unprincipled and therefore unpredictable".[92] The crucial difficulty for the claimant was that the oral agreement was incomplete which meant that he could not prove "an expectation of a certain interest in land".[93] According to Lord Scott, the claimant's expectation, which was encouraged by the defendant, was that upon the grant of planning permission there would be a successful negotiation of the outstanding terms of the contract for the sale of the property to him. At that point, a formal contract, which would include the already agreed core terms as well as additional new terms, would be prepared and entered into. He concluded that an expectation dependent upon the conclusion of a successful negotiation is not an expectation of a sufficiently certain interest in land.[94]

Even if a complete agreement for the acquisition of an interest in land had been reached, Lord Scott stated *obiter* that the doctrine cannot be relied on to render enforceable an agreement that statute has declared to be void.[95] Unlike constructive trusts, proprietary estoppel is not expressly exempted from the provisions of section 2.[96] Lord Scott's *obiter* comments make the doctrine's future role in the world of informal agreements highly precarious, as two recent decisions of the Chancery Division of the High Court

86 M Dixon, "Invalid contracts, estoppel and constructive trusts" (2005) *Conv* 247, at pp. 252–5; B McFarlane, "Proprietary estoppel and failed contractual negotiations"(2005) *Conv* 501, at pp. 516–21.

87 [2008] UKHL 55.

88 See McFarlane, "Proprietary estoppel", n. 86 above.

89 [2008] UKHL 55.

90 Ibid., at para. 85 (per Lord Walker).

91 Ibid., at paras 17 and 28.

92 Ibid., at para. 28.

93 Ibid., at para. 18–19. Lord Scott referred to the *dicta* of Oliver J in *Taylors Fashions Ltd* v *Liverpool Victoria Trustees Co Ltd* [1982] QB 133, at p. 144, and Lord Kingsdown in *Ramsden* v *Dyson* (1866) LR 1 HL 129, at p. 170, setting out this requirement.

94 Ibid., at para. 18.

95 Lord Walker did not think it necessary or appropriate to comment on this issue. See ibid., at para. 93.

96 He was satisfied that the circumstances did not give rise to a constructive trust, as, unlike other joint venture cases where a constructive trust was imposed, the defendant owned the property in question before they embarked on the joint venture. See ibid., at paras 30–6.

demonstrate. In *Herbert* v *Doyle*,[97] the court distinguished the *Cobbe* case on the basis that the terms of the agreement being considered were complete. It ruled that, where all the other requirements of the doctrine of proprietary estoppel are satisfied, a claim will not fail because it consists of an agreement which falls foul of section 2. The proper means of giving effect to the estoppel is to recognise or impose a constructive trust so that the remedy falls within the exception set out in section 2(5).[98] In contrast, the court in *Hutchinson* v *B and DF Limited*[99] endorsed Lord Scott's *obiter* view and rejected a proprietary estoppel claim where a complete oral agreement for the grant of a lease had been reached.

In the *Cobbe* case, Lord Scott felt that it was unacceptable to excuse parties to an oral contract in relation to land from the statutory formalities for such contracts, which they must have been aware of. On the other hand, Lord Walker's exposition of the law in the *Cobbe* case indicates tolerance of the highest levels of naivety amongst claimants who are on the receiving end of gifts or testamentary dispositions which fail to comply with the relevant statutory formalities.[100] It is submitted that a remedy pursuant to the doctrine of proprietary estoppel should be available to informal purchasers, as well as informal donees or successors, and although the commercial nature of the transaction may be an appropriate factor to take into account in considering whether the elements of the doctrine have been satisfied, it is an inadequate basis for the decision to grant or withhold such a remedy. McFarlane, in his 2005 article, described how "the law recognises the practical reality that expectations can be generated in the absence of finalised bargains and responds to the consequent need to provide redress to those who rely on such expectations".[101] In the aftermath of the *Cobbe* case, the potential withdrawal of this remedy from all informal purchasers gives cause for grave concern and highlights the need to insert an express exception for proprietary estoppel in section 2 of the 1989 Act.

The Law Commission, in its deliberations of when an adverse possession application should succeed in spite of an objection by the registered owner, assumes that a purchaser who paid the purchase price and went into possession pursuant to an oral contract for sale or a lease would succeed in obtaining a remedy under the doctrine of proprietary estoppel.[102] Recent caselaw highlights the danger of such an assumption and how easily the preferential treatment afforded by the Land Registration Act 2002 to such an informal purchaser who can also prove adverse possession for 10 years could become meaningless. Such an informal purchaser could be left in a legal limbo, unable to regularise his or her position through proprietary estoppel or adverse possession and in constant danger of being evicted.

Even if it is assumed that a purchaser under an invalid contract for sale would be entitled to be registered as owner on the basis of his or her equity by estoppel, the Law Commission also neglected to discuss, in any detail, the difficulties which such a purchaser may encounter in proving adverse possession. However, it briefly acknowledged that in many cases where an equity arises by proprietary estoppel, the possession of the party

97 [2008] EWHC 1950 (Ch).

98 Ibid., at paras 12–15.

99 [2008] EWHC 2286 (Ch), at paras 64–70.

100 See [2008] UKHL 55, at para. 68.

101 See McFarlane, "Proprietary estoppel", n. 86 above, at p. 515.

102 In Ireland and Northern Ireland this purchaser would be able to seek an order for specific performance of his or her contract for sale as the doctrine of part performance, which renders an oral contract enforceable, is still in force. However, if a concluded contract for sale had not been reached, the Irish courts may award a remedy on the basis of proprietary estoppel, see *An Cumann Peile Boitheimeach Teorenta* v *Albion Properties Ltd & Others* [2008] IEHC 447.

asserting it will not have been adverse, because he or she will have been in possession of the land with the consent of the registered proprietor.[103] Where the vendor represents to the purchaser that the purchaser may go into possession in spite of the absence of formalities, it may be difficult for the purchaser to argue subsequently that his or her possession was adverse.[104] The Law Commission envisaged that the condition set out in Schedule 6, paragraph 5(2), of the 2002 Act would cater for a purchaser who had paid the purchase price, gone into possession and treated the property as their own. Such actions may easily be construed as detrimental reliance, thus rendering the recovery of possession by the vendor improbable. The payment of the purchase price easily marks the moment at which the purchaser's possession becomes attributable to his or her equity and not the owner's permission. However, the Limitation Act 1980 fails to provide any guidance on whether time can run in favour of a person with an equity by estoppel. As was the case with a purchaser under an enforceable contract for sale, it is difficult to ascertain whether a right of action has accrued to the vendor and, if so, when that right accrued.

Cullen v *Cullen*[105] is the main authority for the proposition that a person, entitled to an equity by estoppel on the basis of a representation made by the owner, may in time extinguish the title of that owner through adverse possession. In that case, a son placed a portable house on his father's land on the basis of an assurance given by his father to his mother that he did not mind. When the father tried to eject the son, Kenny J ruled that he could not go back on his assurance. Kenny J made his decision on the basis of the doctrine of promissory estoppel as he was of the opinion that the doctrine of proprietary estoppel only applied in the case of a mistaken belief about the ownership of the land.[106] As a result, Kenny J thought that he did not have jurisdiction to order the father to transfer the site to the son, as promissory estoppel can only be used as a shield, not a sword. He noted, however, that once 12 years had passed, the son would be able to bring a successful adverse possession application to the Land Registry. It is submitted that undue reliance may have been attached to Kenny J's *obiter* comment. Kenny J clearly wished to make a positive order which would confer rights on the son in relation to the site but he was not sufficiently comfortable with the principles of proprietary estoppel to avail of the extended jurisdiction which it permits. He must have felt compelled to suggest a method by which the title could be regularised in the future. Although he stated that time began to run when the son commenced building, he failed to clarify whether this was the point at which his possession ceased to be consensual and became attributable to his inchoate equity.

According to Walstead:

> Until the representor attempts to resile from the assumption engendered by him, the representee has merely a permissive licence and will be unable to claim to be in adverse possession . . . Once the representor retracts the permissive licence a

103 Law Comm No 254, n. 4 above, at para. 10.50, n. 162.

104 A claimant who can establish an estoppel pursuant to the mistake limb of the doctrine of proprietary estoppel will find it much easier to prove adverse possession as the circumstances are not complicated by a representation which could be construed as permission by the owner. The estoppel exception to the veto system introduced by the 2002 Act may allow an applicant who made an initial mistake in relation to the position of his or her boundaries but cannot prove good faith over the 10 years required by Sch 6, para. 5(4), to be registered as owner. The applicant must, however, be in a position to prove that his or her neighbour acquiesced in the mistake. In contrast, an applicant who believed he or she was the owner of the boundary land throughout the 10-year period need not prove acquiescence on the part of a neighbour.

105 [1962] IR 268.

106 Ibid., at pp. 291–2. J Mee discusses the development of the doctrine of proprietary estoppel in Ireland in "Lost in the big house: where stands Irish law on equitable estoppel" (1998) 23 *Irish Jurist* (ns) 187.

> right of action may accrue to him. At that time the representee may be said to be in adverse possession as his licence has been replaced by an inchoate equity.[107]

Where the owner of registered freehold or leasehold land enters into an oral contract to sell that estate and the purchaser enters into possession in reliance on a representation made by the owner, it seems fair to assume that the purchaser's licence would be implicitly terminated once the entire purchase price was paid. At that point, the purchaser should also be in a position to demonstrate the *animus possidendi* essential in proving adverse possession. In such circumstances, the adjudicator would probably view the purchaser as entitled to be registered as the owner of the freehold or leasehold estate. However, if the registered owner entered into an agreement to grant a lease or a sublease of the land, clearly the circumstances would preclude the registration of the purchaser as the proprietor of the entire registered estate. The complications, discussed earlier, engendered by a finding of adverse possession against a purchaser who holds pursuant to an enforceable agreement for a lease are avoided if the contract is unenforceable and the circumstances give rise to an equity by estoppel. This is because the adjudicator is specifically endowed with the discretion to order an appropriate remedy, which would clearly be an order for the grant of the agreed lease.

Conclusion

It is clear that the reforms to the law on adverse possession introduced by the 2002 Act leave the informal purchaser in possession, who was intended to receive preferential treatment, in a precarious position. It is far from clear whether such a purchaser will be able to prove that his or her possession was adverse. It is submitted that the reforms should have been accompanied by legislative clarification on when time begins to run against the vendor. Earlier, it was argued that a right of action should be deemed to accrue on the determination of the vendor's licence or the payment of the purchase price, whichever occurs earlier. It was also pointed out that the definition of a "purchaser" for the purposes of such a deeming provision should exclude a person who has entered into an enforceable agreement for the grant of a lease. It would not be necessary to distinguish between contracts which comply with section 2 of the Law of Property (Miscellaneous Provisions) Act 1989 and those which do not in deeming when a right of action accrues. However, a purchaser in possession pursuant to an oral contract for sale who wishes to make an adverse possession application faces the additional hurdle of establishing the elements required by the doctrine of proprietary estoppel. In the aftermath of the *Cobbe* case, such a purchaser looks increasingly unlikely to succeed. To restore this remedy to purchasers it would be necessary to insert a statutory exception for proprietary estoppel into section 2 of the 1989 Act. To conclude, although the Law Commission's intention was to make the remedy of adverse possession available to informal purchasers, this intention is likely to be frustrated without the introduction of the legislative clarifications outlined in this article.

107 M Walstead, "Proprietary estoppel and the acquisition of possessory title" (1991) *Conv* 280, at p. 283.

NILQ 60(3): 325–42

A consensus on the reform of the House of Lords?

MARK RYAN*

Coventry University

Introduction

> . . . there is the potential to reach a degree of cross-party consensus that will lead to the completion of Lords reform. The free votes in the Commons in March gave us a clear direction of travel on an issue that has dogged the country for decades. We now have a chance finally to finish the job.[1] (Jack Straw MP, Secretary of State for Justice and Lord Chancellor, 19 July 2007)

The above Formal Statement on Lords reform followed a series of free parliamentary votes in March 2007[2] in which both Houses debated and voted on the future composition of a fully reformed second chamber. The House of Commons voted to retain the second chamber and remove the remaining hereditary peers. In addition, it simultaneously endorsed both a fully elected and an 80 per cent elected House, whilst rejecting all other options.[3] In contrast, the House of Lords rejected all options except for a wholly appointed House. These parliamentary votes led eventually to the publication in July 2008 of a further government White Paper[4] on the House of Lords aimed at completing the reform process. The government's proposals in this paper are based on the votes cast in the

* BA, MA, PGCE, Barrister (non-practising), Senior Lecturer in Law at Coventry University. This article is, in essence, the paper which was presented to the Society of Legal Scholars Conference held at the London School of Economics on 18 September 2008. I am very grateful to an esteemed constitutional academic who, following this conference, made some invaluable and constructive comments in respect of this paper.

1 Hansard HC Debs, 19 Jul 2007, vol. 463, cols 450–1.

2 In March 2007, both the House of Commons and the House of Lords were provided with the following options to vote on: fully appointed, 20 per cent elected, 40 per cent elected, 50 per cent elected, 60 per cent elected, 80 per cent elected and fully elected. In addition, the House of Commons also voted on the issue of bicameralism as well as whether to remove the remaining hereditary peers. These options were the same as those voted on by both Houses in February 2003, except that in 2003 the House of Commons did not vote on the issue of the hereditary peers. For an analysis of the 2003 debates and votes see, M Ryan "Parliament and the Joint Committee on House of Lords reform" (2003) 37 *The Law Teacher* 310 and I Mclean, A Spirling and M Russell, "None of the above: the UK House of Commons votes on reforming the House of Lords, February 2003" (2003) 74 *Political Quarterly* 298.

3 It should be noted that this paper draws upon, and extrapolates from, the information provided by the following paper of how each MP voted in March 2007 together with any voting patterns identified: R Cracknell, *Commons Divisions on House of Lords Reform: March 2007* (London: House of Commons Library 2007).

4 The Governance of Britain, *An Elected Second Chamber: Further reform of the House of Lords*, Cm 7438 (Norwich: The Stationery Office 2008).

House of Commons; although it has been made clear that it was not "a final blueprint for reform",[5] but intended to generate further debate on Lords reform. The expectation was that, following consultation, government Ministers would consider how to take these issues forward before a commitment was made in the manifesto for the next general election so that reform could be implemented thereafter. The purpose of this article is to consider the 2007 parliamentary debates and votes and examine whether there really is any meaningful consensus on how to complete the reform of the House of Lords. It is important to note that although these were free votes, all three of the main political parties (and so their MPs) had a specific 2005 manifesto commitment on the reform of the House of Lords.

At the outset, it is essential that three points are made concerning the limited scope of this article. Firstly, as the article involves an examination of the parliamentary debates and votes on the future composition of a fully reformed upper House, it does not, therefore, focus on the functions that such a chamber *should* perform. Although it is commonly argued that in reforming the House of Lords, composition should necessarily be determined by function, the fact remains that the votes in both Houses of Parliament were confined strictly to the issue of composition. In any event, it is possible to contend that the principal functions of the second chamber, viz deliberative, legislative revision, scrutiny and examination of the government of the day, are already fairly well settled. Secondly, although the government has referred specifically to seeking a broad consensus between the political parties in completing the reform of the Lords, this article does not seek to question the hypothesis that a cross-party political approach is necessarily the correct one to adopt. Thirdly, the arguments concerning elected/appointed members are analysed through the prism and specific perspective of members of both Houses of Parliament as they were expressed during the various parliamentary debates.

Bicameralism

It is clear that there is a general consensus within both Houses in favour of bicameralism. All three major political parties predicated their 2005 manifesto commitments on retaining a second chamber. Moreover, in the two separate votes in 2003 and 2007, the House of Commons has successively, and decisively, dismissed the option of abolition (390 votes to 172[6] and 416 to 163,[7] respectively). Such was the support for retaining a second chamber in 2007 that Jack Straw was moved to comment that the possibility of unicameralism had been "buried" by this overwhelming vote.[8] In addition, although in the House of Lords there were no motions for unicameralism in either 2003 or 2007, it is obvious that its view on this matter is hardly in doubt. In fact, it is interesting to reflect that during the course of the 2007 debates, the issue of unicameralism merited little discussion in either House.

The above notwithstanding, there are a few points which are worth noting. Firstly, it is significant that a sizeable minority of MPs (172 in 2003 and 163 in 2007) were prepared to countenance a single-chamber Parliament. In particular, these figures were composed largely of Labour MPs (158 in 2003[9] and 153 in 2007[10]) with the latter vote including one

5 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at p. 3. For a very general overview of the 2008 White Paper, see M Ryan, "The house that Jack built" (2008) 158 *NLJ* 1197.

6 Hansard HC Debs, 4 Feb 2003, vol. 399, col. 221.

7 Hansard HC Debs, 7 Mar 2007, vol. 457, col. 1601. These figures include, rather confusingly, an MP who voted both Aye and No.

8 See n. 1 above, at col. 459.

9 *The House of Lords: Reform*, Cm 7027 (Norwich: The Stationery Office 2007), p. 17.

10 See Cracknell, *Commons Divisions*, n. 3 above, at p. 3.

member of the June 2008 Cabinet.[11] Although in 2007 all six SNP MPs also voted for abolition, there was virtually no other support elsewhere. Secondly, the pattern of the votes cast overall is of interest as, in both 2003 and 2007, a number of MPs not only voted for abolition, but then also preceded to vote for other options in the context of a bicameral system. In 2003, the Joint Committee on House of Lords Reform pointed out that of the 172 MPs who had voted for unicameralism, 160 thereafter had voted for one or more of the other bicameral options.[12] Similarly, some research reveals that, in 2007, out of the 163 MPs who voted for abolition, only 11 of them (all Labour) voted No, or declined to vote at all, in respect of the other options.[13] The remaining (abolitionist) MPs, therefore, proceeded to vote, and have an impact on, one or more of the other options for a bicameral legislature (see below).

In any event, it is clear that there is a general consensus that unicameralism is not a realistic option and that any constitutional settlement, therefore, will inevitably be based upon a two-chamber Parliament. In fact, the 2008 government White Paper did not even mention the option of unicameralism in passing. In 2003 the late Robin Cook MP, the then Leader of the House and ardent supporter of a reformed second chamber, cautioned that, if the argument over reforming the House of Lords continued to stagnate, more and more of the public could end up losing patience with the issue altogether and support abolition after all.[14]

The remaining hereditary peers

A second issue on which there is broad agreement is the (ultimate) removal of the remaining hereditary peers. The government's 2007 White Paper made it clear that in the Cross-Party Group talks there was agreement "that the special arrangements"[15] made for the retention of these peers in 1999 should be brought to an end. Moreover, none of the three main political parties' manifestos at the last general election stated that the hereditary peers should remain. In fact, each of the last three Labour Party manifestos has expressly called for their expulsion.[16] In 2007, the House of Commons voted overwhelmingly by 391 votes to 111 to endorse the following motion stating: "That this House is of the opinion that the remaining retained places for peers whose membership is based on the hereditary principle should be removed."[17] Indeed, the majority for this (280) was slightly bigger than for the votes cast in favour of retaining bicameralism (253). The breakdown of how the parties voted is significant as they split on this motion. In brief, 60 Liberal Democrat MPs (all those who voted) and a large majority of Labour MPs (305) voted in favour of this motion. In contrast, the majority of Conservative MPs (110), together with one Independent MP, voted against.[18] Theresa May MP, the then Conservative Shadow Leader of the House, argued that although she supported the eventual removal of the hereditary peers, "they must be replaced by elected Members".[19] As a result, she tabled an amendment to the motion to the effect that the hereditaries be removed "once elected

11 Hazel Blears MP (Secretary of State for Communities and Local Government). In addition, a Minister attending Cabinet meetings, Caroline Flint MP (Minister for Housing), also voted for abolition.

12 Joint Committee on House of Lords Reform, Second Report of Session 2002–03, *House of Lords Reform: Second Report*, HL Paper 97, HC 668 (Norwich: The Stationery Office 2003), at para. 13.

13 These figures exclude any votes that may have been cast for the removal of the hereditary peers.

14 See n. 6 above, at col. 161.

15 See *House of Lords*, n. 9 above, at p. 8.

16 2005 *Manifesto*, "Britain *forward* not back" (Labour Party 2005), p. 110; 2001 *Manifesto*, "Ambitions for Britain" (Labour Party 2001), p. 35; 1997 *Manifesto*, "Because Britain deserves better" (Labour Party 1997), p. 32.

17 See n. 7 above, at cols 1632–5.

18 See Cracknell, *Commons Divisions*, n. 3 above, at p. 4.

19 Hansard HC Debs, 6 Mar 2007, vol. 457, col. 1414.

members have taken their places in a reformed House of Lords".[20] It is worth remembering that in the event of the hereditary peers being expelled, this would necessarily have a disproportionate impact on the strength of the Conservatives in the second chamber as they presently represent the largest group of hereditary peers. This amendment, however, was defeated by 329 votes to 241. The overwhelming majority of votes defeating it were cast by Labour MPs (309) against a combination of Conservative (173) and Liberal Democrat (62) votes.[21] Interestingly, Liberal Democrat MPs voted firstly for the amendment and then, following its rejection, for the original substantive motion to remove the hereditaries without any preconditions.

Although the House of Lords did not vote on the above motion, in July 2007 the Constitution Unit rather helpfully published the results of a survey into the views of peers on the issue of the remaining hereditaries. They found that out of 373 peers surveyed, 259 peers agreed (153 of them strongly) that there should no longer be automatic rights for hereditary peers to sit in the House, as opposed to 65 who disagreed (20 of them strongly).[22] These figures would appear consistent with the comment of the then Lord Chancellor, Lord Falconer of Thoroton, who said, during the debates in the House of Lords in March 2007, that there was a general recognition (though not universally shared), "that the remaining retained places for hereditary Peers should cease".[23]

Following these votes, the government, in the July 2007 Green Paper (*The Governance of Britain*), stated that "in line with the wishes of the House of Commons" it was committed to removing the anomaly of the remaining hereditary peers as part of a package of reforms.[24] This implied that the government was therefore intent on removing the remaining hereditaries in the context of a wholly or largely elected House, and not before then as a separate, free-standing reform. This explained the government's dismissive attitude towards the recent Steel[25] and Avebury[26] Private Members' Bills. Both proposed to abandon hereditary by-elections forthwith with the intention that the hereditary peers would gradually wither on the vine. The hereditary peer Earl Ferrers, however, preferred to describe the process as a "quiet, gentle strangulation, getting rid of them one by one until they no longer exist".[27] The justification for these Bills was that, in the absence of any immediate prospect of full reform being agreed, modest measures – such as the abolition of by-elections – could nevertheless be made in the interim in order to improve the House (i.e. running repairs). It is pertinent to note that the net effect of the abandonment of these by-elections would have been, ultimately, to achieve the 2005 Labour Party manifesto commitment to remove the remaining hereditary peers;[28] albeit that, rather than being expelled in one tranche, they would simply not be replaced when they died.

20 See n. 7 above, at cols 1627–32.

21 See Cracknell, *Commons Divisions*, n. 3 above, at p. 4.

22 UCL: The Constitution Unit, "Press notice": 19 July 2007.

23 Hansard HL Debs, 13 Mar 2007, vol. 690, col. 722.

24 *The Governance of Britain*, Cm 7170 (Norwich: The Stationery Office 2007), para. 138.

25 House of Lords Bill 2007 (HL Bill 3). Lord Steel of Aikwood introduced a very similar Bill again in December 2008 (HL Bill 4).

26 House of Lords (Amendment) Bill 2008 (HL Bill 22) – introduced by Lord Avebury.

27 Hansard HL Debs, 30 Nov 2007, vol. 696, col. 1428.

28 "Britain *forward* not back", n. 16 above. Interestingly, the Steel and Avebury Bills would also have achieved the objective of the government's 2003 Consultation Paper which, inter alia, sought to remove the hereditary peers in the absence of a consensus on Lords reform following the inconclusive results of the votes in February 2003: *Constitutional Reform: Next steps for the House of Lords*, CP 14/03 (London: DCA 2003). Although a Bill to implement this paper was foreshadowed in the 2003 Queen's Speech, it never materialised. On this paper see M Ryan, "Reforming the House of Lords: a 2004 update" (2004) 38 *The Law Teacher* 255.

Although the intention to remove the hereditary peers was repeated in the 2008 White Paper,[29] there is still debate over the timing of their exit. The government proposed that during the transitional phase to a reformed House, all hereditary by-elections would terminate (in effect achieving the aim of both the Steel and Avebury Bills above). According to Jack Straw, this process would take place following the passage of relevant legislation and during the transition to a fully reformed House.[30] In relation to the hereditary peers who do not die out, the White Paper put forward three possible options for removing them.[31] The first and second options would involve removing these surviving hereditary peers once the third group of new members arrived. The difference between these two options is that, in the first, the life peers would remain in the chamber until they died out whereas, in the second, they would be removed in tandem with the hereditaries. The third option would involve removing all life and hereditary peers in three tranches. The first option, in particular, looks far from being uncontentious as Nick Herbert MP, the then Conservative Shadow Secretary of State for Justice, argued that it would be both inequitable and invidious to remove the remaining hereditary peers whilst "the 400 life peers created under Labour" remained in the House.[32]

One thing is clear, it is inevitable that the remaining hereditary peers will be removed in any agreed long-term reform package. There is cross-party consensus on this in the Commons as, even though Conservative MPs (with Liberal Democrat support) wanted their removal specifically tied to a commitment to replace them with elected members, they clearly do not envisage them forming part of a completely reformed House. Recent research revealed that in the House of Lords a significant majority of peers surveyed also supported the principle that there should be no automatic rights for the hereditaries to sit in the chamber. Finally, support for their removal was urged recently in December 2007 by the House of Commons Public Administration Select Committee, albeit that this should take place in the context of, and form part of, an interim House of Lords Reform Bill.[33]

The fully elected option

The fully elected option was approved by the House of Commons by 337 votes to 224.[34] This majority of 113 was described by Jack Straw as "a very significant majority".[35] The importance of this is that the House of Commons has now approved a wholly elected House and the majority for it does appear to be a seemingly convincing one. Most of those that voted in favour were Labour MPs (210 out of the total of 337 Ayes)[36] including a majority of the June 2008 Cabinet (12 Ayes to 5 Noes). These Labour members were supplemented by a minority of 57 Conservative MPs and all the Liberal Democrats who voted (59)[37] which included virtually the whole of the June 2008 Liberal Democrat Shadow Cabinet. This is in line with the Liberal Democrat's previous votes in 2003 in which they

29 See Governance of Britain, *An Elected Second Chamber*, n. 4 above, at p. 7.

30 Hansard HC Debs, 14 Jul 2008, vol. 479, col. 23.

31 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at pp. 76–81.

32 See n. 30 above, at col. 25.

33 House of Commons Public Administration Select Committee, Second Report of Session 2007–08, *Propriety and Peerages*, HC 153 (Norwich: The Stationery Office 2007), at para. 176.

34 See n. 7 above, at col. 1623.

35 See n. 30 above, at col. 29.

36 See Cracknell, *Commons Divisions*, n. 3 above, at p. 4.

37 Ibid.

supported a fully elected House by a majority of 34.[38] In 2007, Liberal Democrat peers supported this option by a majority of 33 (41 to 8).[39]

There are, however, a host of issues raised in relation to the House of Commons' vote to approve a wholly elected House. Firstly, although 337 MPs voted for it, they represent only just over half of the overall total of the chamber. Furthermore, in 2003, the House of Commons rejected this option by 17 votes, with a majority of both Labour and Conservative MPs voting against it.[40] In the 2007 votes, a majority of 69 Conservative MPs opposed this option (126 to 57)[41] and this represented a more significant rejection of it than in 2003 (79 to 59).[42] In fact, both Labour and Conservative MPs were divided on the fully elected option with a sizeable block of just under 100 Labour MPs voting against it (including Jack Straw, under whose aegis the 2008 White Paper had been produced). Although in 2007 a majority of Labour MPs approved the wholly elected House, they were not so supportive in 2003 when a majority voted against it by 40 votes (197 to 157).[43]

Secondly, the fact is that the House of Lords has twice voted to dismiss this option by the substantial majorities of 204 in 2007[44] and 223 in 2003.[45] In fact, in 2007 the only group of peers in the Lords to support this option were the Liberal Democrats. In contrast, Labour peers rejected it by a majority of 23 votes, the Conservatives 125 and the independent Crossbench peers by 85.[46] In short, the House of Lords has consistently made it very clear, for whatever reason – self-serving or otherwise – that it opposes a wholly elected House. Indeed, in 2007 only a mere 122 peers[47] supported a wholly elected House and in 2003 it was even less with 106.[48] There is clearly no inter-House consensus on the fully elected option.

Thirdly, no previous report or White Paper in the past decade has recommended a wholly elected second chamber. This includes the 2000 Royal Commission, which specifically stated that it could not recommend a wholly directly elected House.[49] Similarly, the 2001 government White Paper did not recommend this option and described two wholly directly elected Houses as "a recipe for gridlock".[50] Neither the 2002 House of Commons Public Administration Select Committee[51] nor the 2005 *Breaking the deadlock* report (produced by a cross-party group of parliamentarians) recommended a wholly elected House.[52] Even the government's 2007 White Paper proposed a mixed House containing both elected and appointed members.[53]

38 See *House of Lords*, n. 9 above.

39 C Clarke, *House of Lords Reform Since 1997: A chronology* (London: House of Lords Library 2007), p. 53.

40 See *House of Lords*, n. 9 above.

41 Above Cracknell, *Commons Divisions*, n. 3 above, at p. 4.

42 See *House of Lords*, n. 9 above.

43 Ibid.

44 Hansard HL Debs, 14 Mar 2007, vol. 690, col. 756.

45 Hansard HL Debs, 4 Feb 2003, vol. 644, col. 120.

46 Clarke, *A chronology*, n. 39 above.

47 See n. 44 above.

48 See n. 45 above.

49 Royal Commission on the Reform of the House of Lords, *A House for the Future*, Cm 4534 (Norwich: The Stationery Office 2000), at p. 113.

50 *The House of Lords – Completing the reform*, Cm 5291 (Norwich: The Stationery Office 2001), p. 18.

51 House of Commons Public Administration Select Committee, Fifth Report of Session 2001–02, *The Second Chamber: Continuing the reform*, HC 494-I (Norwich: The Stationery Office 2002), at p. 25.

52 K Clarke, R Cook, P Tyler, T Wright, G Young, *Reforming the House of Lords: Breaking the deadlock* (London: The Constitution Unit 2005), p. 19.

53 *House of Lords*, n. 9 above, at p. 6.

In addition, the fully elected option appears to be at odds with the manifestos of both the Conservative and Liberal Democrats for the 2005 general election. The Conservative Party manifesto referred to seeking "cross-party consensus for a *substantially* elected House of Lords"[54] and, as noted above, the majority of their members voted against the wholly elected option both in the Lords and the Commons (in line with their manifesto). The Liberal Democrat manifesto made reference to replacing the House of Lords "with a *predominantly* elected second chamber".[55] Yet all of their MPs that voted (together with most of their counterparts in the Lords), not only voted for the 80 per cent elected option (see below), but also supported the wholly elected option as well, seemingly at variance with their 2005 manifesto. Interestingly enough, this support for a wholly elected chamber would, however, appear to be consistent with their 2001 manifesto commitment for a "directly elected Senate".[56] The 2005 Labour Party manifesto was much less specific than either of the two parties above. It stated that a reformed upper House "must be effective, legitimate and more representative",[57] without specifying any more details about its proposed composition. Instead, it promised a free vote on composition – hence the debates and votes in March 2007. Their previous manifesto commitment in 2001,[58] however, had sought to implement "in the most effective way possible" the recommendations of the 2000 Royal Commission which had recommended a hybrid (and largely appointed) House.[59]

Fourthly, quite apart from the somewhat mixed messages that the pattern of voting indicates, there is a suggestion that, in any event, the votes in the House of Commons were marred by tactical voting. Dr Meg Russell has argued that the 2007 votes cannot be taken at their face value as

> the vote for a wholly elected house was clearly influenced by tactical voting. Anybody who's followed this debate over recent years would be able to spot some very unlikely names going through the division lobby in favour of an all elected house. This was clearly a spoiling tactic because the 80% elected option, which had already passed, was seen as dangerous.[60]

Similarly, Lord Higgins cast doubt on these votes saying that it was well known that the vote in favour of the wholly elected House "was a result of tactical voting by those who were actually in favour of a wholly appointed House".[61] More recently in July 2008 in the Liaison Committee, Sir Patrick Cormack MP suggested to the Prime Minister that what had given the fully elected option "a good majority" were the tactical votes of Labour MPs (who did not want any elected members at all) voting for the fully elected option in order "to throw a spanner in the works".[62] On the unveiling of the 2008 White Paper, Sir Patrick Cormack repeated that the vote for the fully elected option "was caused by a tactical switch by a number of Members, led by the Hon. Member for Tyne Bridge (Mr Clelland), who is nodding vigorously".[63] The Labour MP, David Clelland, had made it clear during the 2007

54 "Are you thinking what we're thinking? It's time for action" (Conservative Party 2005), p. 21 (emphasis added).

55 "The real alternative" (Liberal Democrats 2005), p. 35 (emphasis added).

56 "2001 Liberal Democrat General Election manifesto: freedom, justice, honesty" (Liberal Democrats 2001), section on reforming politics and the constitution.

57 "Britain *forward* not back", n. 16 above.

58 "Ambitions for Britain", n. 16 above.

59 Royal Commission, *A House for the Future*, n. 49 above , at p. 8.

60 M Russell, *Lords Reform: Principles and prospects* (London: The Constitution Unit, 13 Nov 2007), at p. 7.

61 Hansard HL Debs, 20 Jul 2007, vol. 694, col. 497.

62 House of Commons Liaison Committee, *The Prime Minister, Oral Evidence*, HC 192-ii (Norwich: The Stationery Office, July 2008), Ev 31, Q 155.

63 See n. 30 above, at cols 28–9.

debates that he intended to vote for the fully appointed option "which calls for a reformed, but *non-elected* second Chamber"[64] and yet he proceeded to vote for both the wholly appointed and fully elected options.

It is clear that a sizeable number of MPs voted for both the wholly appointed and wholly elected options. These are options which are diametrically opposed to one another. Richard Cracknell has identified that of the 196 MPs (115 Labour, 80 Conservative and an Ulster Unionist) who voted for a wholly appointed House, 72 of them thereafter proceeded to vote for a fully elected House as well.[65] Some research into these figures indicates that 61 of these 72 MPs were Labour whilst 11 were Conservative. According to Lord Naseby, as a result of a number of MPs supporting both options, the legitimacy of the majority for the fully elected vote was therefore "highly questionable".[66] It is also pertinent to note that 42 (all Labour) of these 72 MPs also voted for abolition. In other words they voted for the somewhat unusual combination of abolition, a fully appointed and a fully elected House. What these options did at least have in common was the fact that none of them were hybrid. Furthermore, 111 out of the 163 MPs who voted for abolition also voted for the wholly elected option. Some analysis of these figures reveals some interesting information. For example, if we disregard all the votes (Ayes and Noes) cast by those supporting abolition in respect of the wholly elected option, we find that the overall majority for the wholly elected House shrinks from 113 (337 to 224) to 46 (revised figures of 226 votes to 180).[67]

It is noteworthy that the 2008 White Paper made no reference at all to any suggestion that tactical voting may have taken place or to any of the patterns of multiple voting. As a result, the paper proceeded on the basis that the votes delivered in March 2007 are to be taken at their face value. Further, in unveiling the 2008 White Paper in July 2008, Jack Straw stated that:

> Those of us who take seriously the way we vote have to be bound by the consequences of our votes. We cannot have a situation whereby Members vote in one Lobby and then say that they actually meant to vote in the other Lobby; indeed, that would be the road to complete disaster.[68]

If there was any tactical voting, however, it can be said that it has not exactly paid off as the government's 2008 White Paper confined the only two possible options for Lords reform to a wholly or 80 per cent elected House.

In terms of tactical voting, one way the issue could of course be resolved would be to take the step of having a further and separate vote (Aye or No) to decide definitively whether the House of Commons really is in favour of a wholly elected option. This would deal conclusively with any suggestion that the 2007 votes did not represent genuine preferences. If the House of Commons is truly supportive of the wholly elected option, it would appear to be a relatively straightforward matter to have it endorsed again. The obvious fear, of course, is that after failing in 2003 to secure a majority for any of the options, and then finally managing to muster it in 2007, there is inevitably a concern not to unravel this position. After all, where would Lords reform be if a further set of votes rejected the wholly elected option, thereby casting doubt on the veracity of the March 2007 votes?

64 See n. 7 above, at col. 1553 (emphasis added).

65 See Cracknell, *Commons Divisions*, n. 3 above, at pp. 3 and 5.

66 See n. 23 above, at col. 690.

67 It should be noted that some abolitionists did not vote on the fully elected option.

68 See n. 30 above, at col. 29.

The 80 per cent elected option

The other option the House of Commons voted for was the 80 per cent elected/20 per cent appointed chamber. This was approved by 305 votes to 267 – a majority of 38 described by Jack Straw as "quite a substantial majority".[69] This is the option which the House of Commons came very close to endorsing in 2003, falling short by a mere handful of votes (281 to 284 votes).[70] In 2007, 62 Liberal Democrats MPs (with one MP not voting) supported this option,[71] having previously supported it in 2003 by 47 votes to 3.[72] It also enjoyed the support of the Liberal Democrat peers in 2007 by a majority of 28 (38 votes to 10).[73] The 80 per cent elected House is also the option which is consistent with both the 2005 Conservative and Liberal Democrat manifestos (see above). Indeed, the present leaders of both of the political parties above voted for this option, together with the Lord Chancellor, Jack Straw. It is also the option that the current Prime Minister voted for. The 80 per cent elected House is, of course, a hybrid option (majority elected/minority appointed) and as such, it is consistent with both the reports of the 2002 Public Administration Select Committee which recommended a 60 per cent elected House[74] and very similar to the 2005 *Breaking the deadlock* report which proposed a 70 per cent elected chamber.[75]

There are, however, difficulties with the 80 per cent elected option. Firstly, it has not gone unnoticed that the figure of 305 MPs represents less than half of the total membership of the House of Commons. Lord Faulkner of Worcester has noted that although much had been made of the votes in the Commons "they are a long way short of demonstrating that there is a consensus, even in the other place".[76] Secondly, a majority of both Labour and Conservative MPs (5 and 18 respectively)[77] rejected this option, even though it appears consistent with the latter's 2005 manifesto commitment for a substantially elected chamber. Both sets of these MPs also rejected this option in 2003 by majorities of 44 and 2 respectively.[78] Thirdly, the House of Lords has twice rejected the 80 per cent elected option by majorities of 222 in 2007[79] and 245 in 2003.[80] The breakdown of peers in the 2007 votes indicated clear opposition to it with most Conservative, Labour and Crossbench peers rejecting it by majorities of 105, 65 and 73 votes, respectively.[81] It is also worth noting that research undertaken by the Constitution Unit indicated that of those peers surveyed, only 32.2 per cent wanted at least some elected members. Not surprisingly, as a group, only the Liberal Democrats (71 per cent of them) supported this principle.[82] As with the fully elected chamber, there is clearly no inter-House consensus on the 80 per cent elected option.

---

69 See n. 30 above, at col. 29.
70 See n. 6 above, at col. 234.
71 See Cracknell, *Commons Divisions*, n. 3 above, at p. 4.
72 See *House of Lords*, n. 9 above.
73 See Clarke, *A chronology*, n. 39 above.
74 See *The Second Chamber*, n. 51 above, at para. 96.
75 See Clarke et al., *Breaking the deadlock*, n. 52 above, at p. 20.
76 See n. 23 above, at col. 705.
77 See Cracknell, *Commons Divisions*, n. 3 above, at p. 4.
78 See *House of Lords*, n. 9 above.
79 See n. 44 above, at col. 752.
80 See n. 45 above, at col. 127.
81 See Clarke, *A chronology*, n. 39 above.
82 See UCL, "Press notice", n. 22 above.

In addition, Dr Meg Russell has raised concern about taking the vote for this option at face value as she has argued that

> even the 80% elected option looks pretty wobbly. The most obvious difficulty is that the Conservatives supported this position, but at the same time said that they were opposed to proportional elections. They haven't said what their alternative is, but if it's not proportional many in the pro-election camp would oppose it – quite rightly in my view. Once this factor is taken into account, the majority of 38 for an 80% elected house doesn't seem so decisive.[83]

Indeed, around a quarter of all the votes in favour of this option (80)[84] were cast by Conservative MPs and the 2008 White Paper indicated that the Conservative Party favours the (non-proportional) first-past-the-post system for elections to any reformed second chamber. This is in contrast to the Liberal Democrats (all of whose MPs that voted supported the 80 per cent elected option), who advocate either of the proportional systems of single transferable vote or open list.[85] In terms of the pattern of votes, it is significant to point out that unlike the fully elected option above which involved a block of MPs voting for both the fully appointed and wholly elected options, in contrast, only a handful of MPs (5) voted for both the fully appointed and 80 per cent elected options.

Finally, although the principle of hybridity has been recommended by all reports in the past decade, it must be remembered that both the 2001 White Paper[86] and the 2000 Royal Commission recommended a mixed House which was largely appointed, rather than being mainly elected. In fact, the latter specifically stated that it could not recommend a largely directly elected House.[87]

A fully or largely elected House?

Although in 2007 the House of Commons finally mustered the support to endorse the options of a wholly and largely elected House, the obvious problem is that *two* options were approved. Not only that, but these two options are contradictory (one supports the hybrid principle, whilst the other does not). The choice between these options is not a question of degree, but one of kind and begs the question: which option should prevail?

The 2008 White Paper, rather unhelpfully, made no recommendation as to which option to endorse. In fact, it did not even indicate a preference and its narrative went out of its way to point out that it was not implying support for one option over the other.[88] Lord Tyler, the then Liberal Democrat Spokesperson for Constitutional Affairs, noted that it was curious that given that the paper was very specific in its detail on many points, the government was "havering" between the two options. He added "No one is asking them to be absolutely determinate about that but at least some preference could be indicated. Surely they could now reveal that."[89] The then Parliamentary Under-Secretary of State, Ministry of Justice, Lord Hunt of Kings Heath, explained that this was deliberate given that the House of Commons had voted for both options.[90] As a result, the 2008 White Paper indicated no preference and the two options will have to be debated and decided in due course.

83 See Russell, *Lords Reform*, n. 60 above.

84 See Cracknell, *Commons Divisions*, n. 3 above, at p. 4.

85 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at p. 15.

86 See *Completing the Reform*, n. 50 above, at para. 11.

87 See Royal Commission, *A House for the Future*, n. 49 above.

88 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at p. 76.

89 Hansard HL Debs, 14 Jul 2008, vol. 703, col. 994.

90 Ibid., at col. 996.

Arguments in favour of a fully elected House

One obvious point in favour of the fully elected option is that it secured the largest majority of votes. In fact, its majority of 113 was 75 more than for the 80 per cent elected option. Indeed, if we aggregate the votes cast by MPs in both 2003 and 2007, we find that with an overall combined majority of 35 in favour of it, the 80 per cent elected option compares unfavourably to the aggregate majority of 96 for the fully elected option.

One advantage of adopting the wholly elected option is that it simplifies the constitutional issues surrounding a reformed second chamber by automatically excluding certain issues from further, possibly intractable, debate. For example, the issue of whether religious representation should be retained in the House (and the wider point of whether other religious representatives should also be included) is decided automatically in the negative. Similarly, under a wholly elected system, retired Supreme Court Justices would not be offered a seat (a proposal which would in any event compromise the separation of powers). In addition, as there would be no appointed members, the fully elected option also does away with the issues surrounding the establishment of an Appointments Commission, viz its constitution, composition, terms, powers, remit, etc. The 2007 White Paper stated that it was generally agreed that any appointed element would be overseen and made by an independent statutory commission[91] and previous reports have also advocated this. A year later, however, the 2008 White Paper conceded that opinion was divided on this issue with the government preferring a statutory Appointments Commission, whilst the Conservative Party favoured a non-statutory version.[92] In any event, there would be debate, inevitably, over the composition and remit of any Appointments Commission.[93]

One major advantage of the fully elected option is defined in the negative. In other words, its virtue is that it is *not* a hybrid system, as the principle of hybridity is considered highly controversial in some quarters. For some, far from combining the advantages of election (legitimacy and constitutional confidence) with appointment (expertise, independence, no overall control by one party), a mixed House would actually be the worst of both worlds. According to an ex-Lord Chancellor, Lord Irvine of Lairg, a hybrid House was "neither fish nor fowl".[94] Sir Gerald Kaufman MP dismissed this principle rather more peremptorily when he asserted that there were essentially three options for reform: abolition, wholly elected or wholly appointed, with all the rest being "gibberish".[95]

The constitutional consequence of a mixed House is that two classes of member would be created, viz the 80 per cent elected members (claiming, arguably, a greater constitutional importance and legitimacy owing to their election) and the remaining 20 per cent appointed members. According to Ben Chapman MP, there was a danger that this would result in a two-tier system with the two types of member challenging each other.[96] Further, Baroness Symons of Vernham Dean posited that the first time a vote turned on the votes cast by the minority appointed members, "a constitutional crisis" would ensue with the majority elected

91 See *House of Lords*, n. 9 above, at p. 40. It should be noted that the House of Commons Public Administration Select Committee recently recommended putting the present Appointments Commission on a statutory basis (see *Propriety and Peerages*, n. 33 above, at para. 135). In addition, the Constitution Unit's 2007 survey indicated that 89.5 per cent of peers supported the Appointments Commission being placed on a statutory basis (see UCL, "Press notice", n. 22 above).

92 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at p. 52.

93 Part 1 of the Steel Bill (n. 25 above) provides an interesting prototype of a statutory Appointments Commission.

94 Hansard HL Debs, 12 Mar 2007, vol. 690, col. 477.

95 See n. 7 above, at col. 1533.

96 See n. 19 above, at col. 1438.

members refusing to tolerate being overruled by the unelected members.[97] This point notwithstanding, one of the advantages which has been levelled in favour of the low proportion of 20 per cent appointees, is that the chances of a vote being decisively influenced by these appointed members is minimised. In any case, these problems would not arise in a chamber which was fully elected with just one single class of elected member.

Other objections are that, in the long term, a hybrid chamber is an inherently unstable constitutional settlement. Although other countries have mixed chambers, they are governed by a written codified constitution – interestingly of the other two countries that lack such a document (New Zealand and Israel), both have unicameral systems. The position of a mixed House can be likened to devolution, with devolution being seen as a *rolling process* rather than a static event (for example, see the Government of Wales Act 2006). Alan Williams MP has argued that, instead of being a solution, hybridity was merely a holding position which was stalling the inevitable in which the elected members, being thwarted by their appointed counterparts, would want more elected members, and that this process would continue: "Once we start down this road, we will eventually arrive at a fully elected House of Lords."[98] In other words, the expectation is that an 80 per cent elected House would inexorably lead to a wholly elected one.

The fully elected option would also avoid any debate inherent in any hybrid system over the proportional balance between elected and appointed members. It is well recognised that even if the principle of a largely elected House is accepted, there is certainly no paradigm percentage/proportion of members. In short, although the House of Commons supported the model of 80 per cent, why is this inherently preferable to a 70 per cent or even a 75 per cent elected House? Further, as the appointed members would comprise only one-fifth of the House, they would be in a clear minority which has led some to question whether such a low proportion of appointed members would be left feeling marginalised and "somehow surplus or supernumerary – aside from the main action".[99] In addition, in 2002 the Joint Committee on House of Lords Reform noted that in order for the appointed component to provide independent and expert elements, they would need "a sufficiently wide base" to provide opinions on a variety of subjects. The committee, however, questioned whether this could be realised satisfactorily in the context of an 80 per cent elected House (of reduced size).[100]

Arguments in favour of an 80 per cent elected House

Although the 80 per cent elected option received fewer votes than the fully elected option, in 2003, this was the option which came closest to being endorsed by the Commons, falling short by 4 votes only. Indeed, it is fascinating to note that earlier research has revealed that this option would have been carried if 4 MPs (who actually supported the option) had not mistakenly voted against it.[101] The hybrid option is also consistent with both the current Conservative and Liberal Democrat manifestos and it is also broadly similar to the 70 per cent elected recommendation made in the 2005 Cross-Party Group publication *Breaking the deadlock*. In fact, all reports in the past decade have favoured a hybrid House. Nick Herbert has posited that, given that the House of Lords voted for a wholly appointed House, the

97 See n. 23 above, at col. 572.

98 See n. 19 above, at col. 1427.

99 Lord Harries of Pentregarth, see n. 23 above, at col. 598.

100 Joint Committee on House of Lords Reform, First Report of Session 2002–03, *House of Lords Reform: First report*, HL Paper 17, HC 171 (Norwich: The Stationery Office 2002), at p. 24.

101 See Mclean et al., "None of the above", n. 2 above, at p. 305.

best hope of consensus would be to retain a minority appointed element.[102] In his reply, Jack Straw agreed that in his personal view "an appointed minority is the best type of consensus".[103] This is curious given that the House of Lords rejected the 80 per cent elected option by a larger margin than it did for the fully elected House (222 and 204 votes, respectively). It is also of interest to note that a significant majority of those favouring abolition also opposed the 80 per cent elected option and so, if we disregard all the votes (Ayes and Noes) cast by those MPs supporting unicameralism, the overall majority for the 80 per cent elected option would increase from 38 (305 votes to 267) to 110 (revised figures of 264 votes to 154).[104]

According to the *Breaking the deadlock* report, "A mixed chamber allows the strengths of both the elected and appointed models to be combined."[105] Further, one of the authors of this report, Dr Tony Wright MP, has argued that

> If we design the second Chamber properly, we can get two good things. We can get a mixed House that gives us enough election to give us enough legitimacy, and we can get enough appointment to give us enough independence and expertise.[106]

A hybrid chamber would also be consistent with international legislatures. Dr Meg Russell has pointed out that it is not uncommon for upper chambers to have a mixture of members with the commonest combination being "a predominantly elected chamber with a small number of appointed or ex-officio members".[107] According to the 2007 White Paper, the current House of Lords is already "hybrid" to some extent, as it contains different categories of members, viz bishops and hereditary and life peers.[108] In relation to this notion of being a hybrid House, Lord Cunningham of Felling has commented that the introduction of a majority of elected members "is a fundamental difference from everything that has gone before".[109]

As noted above, although the fully elected option would dispense with certain issues associated with reforming the second chamber, the problem is that it would raise others. For example, it would make it difficult to guarantee the recently recognised constitutional principle that no one political party should have a majority of seats in the House. Elections are of course unpredictable, with the electorate determining which parties enjoy representation. At least in a mixed House with a fifth of it being appointed the possibility of one party dominating it would be reduced. Another complication associated with the wholly elected option is the loss of expertise, which of course is invariably appointed, rather than being elected through the ballot box. The 20 per cent appointed option would enable appropriate experts to be specifically selected and then appointed (however, note the concern expressed above by the 2002 Joint Committee).

A further problem with a wholly elected House is that it would, in practice, rule out the possibility of having any meaningful independent representation. Although it is of course possible for independent members to be elected, the reality is that elected members tend to be affiliated to a particular political party and its party machine. The evidence for this is clear

102 See n. 30 above, at col. 25.

103 Ibid., at col. 26.

104 It should be noted that a handful of abolitionists did not vote on the 80 per cent elected option.

105 See Clarke et al., *Breaking the deadlock*, n. 52 above.

106 See n. 7 above, at col. 1547.

107 M Russell, *Second Chambers Overseas: A summary* (London: The Constitution Unit 1999), p. 6.

108 See *House of Lords*, n. 9 above, at p. 32.

109 See n. 94 above, at col. 494.

to see in the House of Commons where very few independent MPs are ever elected. Moreover, it could be argued that putative independent members may not wish to participate in the process and machinations of an election. A fourth problem of a fully elected chamber is the question of it being representative and reflective of society at large (the same charge could, of course, also arguably be levelled at a largely elected House, albeit with less force). Indeed, in terms of sex and race, MPs in the wholly elected House of Commons are hardly truly representative of society.[110] As pointed out by the 2007 White Paper, a concern is that, without "strict rules" about who could stand as a candidate being in place, it would be very difficult to ensure that a wholly elected upper chamber was sufficiently representative of the racial, gender and religious mix of the nation.[111] In short, whereas elections may confer constitutional legitimacy, they do not guarantee social representativeness. At least the 20 per cent appointed element could go some way to offsetting the problems of a lack of diversity typically associated with elected members.

Finally, one constitutional problem frequently attached to a wholly elected House is its potential threat to the primacy of the Commons (this primacy being a principle which the 2008 White Paper asserted as "acknowledged as beyond debate").[112] As elections confer constitutional legitimacy and authority, a wholly elected House would therefore necessarily be more constitutionally legitimate, aggressive and assertive than the present chamber. This issue, however, is clearly less acute with the 80 per cent elected option where not all members would be elected. According to the *Breaking the deadlock* report, a hybrid option "helps ensure that whilst the chamber gains legitimacy, it can never challenge the primacy of the fully elected House of Commons".[113] It should be remembered that the report recommended a 70 per cent elected House.

Consensus

In February 2007, in a written answer to a question concerning consensus in relation to Lords reform, the then Lord Chancellor, Lord Falconer, stated that "Consensus does not mean unanimity on all the issues but, as has been evident through the cross-party discussion, the government are seeking general agreement on key areas".[114] Jack Straw stated subsequently that, as reform of the Lords was a central aspect of the British Constitution, it was right that there was "as much all-party agreement as possible", however, he accepted "that there may well not be total agreement".[115] As history has demonstrated, universal agreement on the composition of a reformed second chamber is simply impossible. Indeed, Kenneth Clarke MP, speaking in the context of the 2007 votes, hoped that "we do not all defeat each other in our anxiety to ensure the perfect reform".[116]

Lord Howe of Aberavon has insisted it was crucial that consensus on Lords reform must not only be between political parties, but also be between the two chambers. He added that this consensus should involve giving as much weight to the views of the Lords as to those of the Commons. In fact, Lord Howe went further and suggested that the scales should be weighted in favour of the House of Lords, partly on the basis that the House of Commons did not understand how the second chamber operated and so "its judgment

110 For example, in August 2008 there were only 126 female MPs in the House of Commons.

111 See *House of Lords*, n. 9 above, at p. 30.

112 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at p. 4.

113 See Clarke et al., *Breaking the deadlock*, n. 52 above.

114 Hansard HL Debs, 22 Feb 2007, vol. 689, col. WA271.

115 See n. 1 above, at col. 450.

116 See n. 19 above, at col. 1430.

cannot be given much weight".[117] Unfortunately, the votes in 2007 reveal a constitutional chasm between the two chambers as the House of Lords voted only in favour of a wholly appointed House (by the commanding majority of 240 votes) and rejected decisively all other options. Further, this vote was made in the full knowledge of the earlier vote in the House of Commons, in the previous week. In July 2007, Jack Straw repeated that the government's intention was to proceed in line with the wishes of the lower (and primary) House.[118] In other words, in respect of composition, consensus meant consensus only within the House of Commons and this is reflected in only two possible reform options being set out by the subsequent 2008 White Paper. Indeed, it is significant to note that the Convenor of the Crossbench Peers and a member of the Cross-Party Group took the view that, as the basis of the group's talks (which helped shape the report) had essentially ignored the views of the House of Lords, it was inappropriate for the term consensus to be used in the White Paper.[119]

The stark differences of view between the two Houses on Lords reform leads inevitably to the probability of inter-chamber conflict if legislation is ever placed before Parliament to implement either a fully or largely elected chamber. Would the government be forced to have to resort to using the Parliament Acts to force such a measure through the legislature? Would this be constitutional? According to the Salisbury–Addison/government Bill Convention, in constitutional theory, resistance in the Lords should be lessened if such a Bill was preceded by a clear manifesto commitment. That said, however, we would certainly be in virgin constitutional territory with proposals for an elected second chamber (whether largely or wholly) and, therefore, the reaction of the Lords may be difficult to assess. Nevertheless, it has not gone unnoticed that it would be somewhat ironic for peers to accept the principle of primacy in general, but then thwart the view of the Commons on Lords reform itself. Lord McNally, the Leader of the Liberal Democrats in the Lords, has warned peers

> I urge noble Lords to listen carefully. There is a noise coming down that Corridor. It is a noise which this House has heard before. It heard it in 1832; it heard it in 1911. It is the thunder of reform, and this House ignores it at its peril.[120]

It is, of course, a peculiarity of our constitutional system that major constitutional changes do not require a convoluted procedure involving specially weighted parliamentary majorities, the consent of the upper chamber or even endorsement via a referendum.

In terms of inter-party consensus, on what issues is there general agreement? The 2008 White Paper stated that "there is already widespread consensus" concerning the constitutional role of the upper House (i.e. to act as an investigatory and scrutinising chamber, as well as holding the executive to account). Similarly, the principles of the primacy of the Commons and the right of the government to ensure its business proceeds through Parliament are also beyond doubt.[121] The paper added that, since the recommendations of the Royal Commission in 2000, there was also a widespread consensus that any elected members "should normally serve a single, non-renewable term of 12–15 years" elected in thirds and serving three electoral cycles.[122] The White Paper conceded, however, that there was division on a number of aspects associated with elections. In fact,

117 See n. 94 above, at cols 487–8.

118 See n. 1 above, at col. 449.

119 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at p. 9.

120 See n. 94 above, at col. 463.

121 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at p. 4.

122 Ibid., at p. 15.

one major criticism which can be levelled at the 2003 and 2007 parliamentary votes is that they did not particularise what "election" meant (i.e. did this connote direct or indirect elections?) – this is important as, for some, indirect elections are little more than a de facto system of appointment. The 2008 White Paper argued that there was a strong consensus within the Cross-Party Group and the government proposed that elections to a reformed House should involve direct elections.[123]

Not surprisingly, the most divisive electoral issue is the electoral system to be employed (indeed, there was no consensus within the Cross-Party Group on this). This is certainly not an inconsequential secondary constitutional matter as it will determine, necessarily, the composition and party balance of any reformed chamber. The 2008 White Paper put forward four different systems for discussion ranging from the first-past-the-post (favoured by the Conservatives) to the alternative vote, single transferable vote and list systems. The Liberal Democrats support the last two systems (albeit only the latter in form of the open list variant).[124] For its part, in its 2007 White Paper, the government proposed "a partially open regional list system".[125] It is also arguable that the electoral system for a reformed second chamber cannot be determined in isolation without reference to the electoral arrangements in the House of Commons, and whether the latter will be subject to reform in the near or long-term future. In short, from a constitutional perspective it would be particularly difficult to justify the use of two identical electoral systems. Similarly, there is no consensus between the main political parties as to which existing election the proposed electoral cycles should be tied (the government and Conservatives favour a general election, whilst the Liberal Democrats prefer the elections to the devolved institutions).[126] Nor is there agreement over the size of constituencies, with the Conservatives advocating smaller, more recognisable ones than the government prefers.[127] Even in the context of a hybrid House, as noted earlier, the two main political parties disagree over whether any Appointments Commission should be statutory or not. Finally, the size of the chamber itself, which is necessarily linked to the electoral system, is also up for discussion.[128]

One interesting aspect which should be mentioned is that the government in both the 2007 and 2008 White Papers made various references to the views (consensual or not) of the Cross-Party Group of parliamentarians. This group, however, was essentially comprised of a *coterie* of frontbench members and therefore lacked backbench involvement. The significance of this is that it is debatable as to whether the viewpoints/consensus forged by this Cross-Party Group (arguably with a frontbench perspective) will necessarily be translated onto the floor of both chambers when considered by all members in due course.

In terms of powers, in March 2007 Jack Straw indicated that there was general agreement "that the current powers of this place in relation to the powers of the Lords or any second Chamber should remain the same".[129] Over a year later, the government, in its follow-up report *Governance of Britain: One year on*, stated that cross-party talks had reached consensus that there should be no reduction in the present powers of the House of Lords.[130] The view that the reformed chamber should not have additional powers, however, is clearly not universal as it could be argued that a more democratic House should

123 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at p. 24.

124 Ibid., at ch. 4.

125 See *House of Lords*, n. 9 above, at p. 39.

126 Governance of Britain, *An Elected Second Chamber*, n. 4 above, at pp. 19–20.

127 Ibid., at p. 20.

128 Ibid., at pp. 21–22.

129 See n. 19 above, at col. 1399.

130 *Governance of Britain: One year on* (London: Ministry of Justice 2008), at p. 16.

arguably (and logically) enjoy a corresponding increase in its powers. Furthermore, even though in January 2007 the constitutional conventions which regulate the relationship between the two Houses (as identified by the 2006 Joint Committee on Conventions),[131] were approved by resolutions in both chambers,[132] these relate specifically to the present (unelected) House of Lords. As a result, there will, inevitably, be some debate over whether these conventions could adapt, or even survive, in the context of a fully reformed, elected second chamber.

Conclusion

In March 2007, the two Houses of Parliament voted on the future composition of a reformed second chamber. What is clear is that the principle of bicameralism is generally agreed and no political party proposes to retain the hereditary peers in the context of a fully reformed House. That said, although the House of Commons approved two options, a majority of Conservative MPs together with a sizeable block of Labour MPs voted against the fully elected option. Moreover, no report in the past decade has even recommended a wholly elected House and the vote in any event has been blighted by a claim that it has been undermined by tactical voting. In relation to the 80 per cent elected option, this was voted against by a majority of both Labour and Conservative MPs. Furthermore, the votes cast in favour of both these options approximated to around only half of the total membership of the House of Commons. In contrast, the House of Lords has very clearly indicated its view by endorsing a wholly appointed House and rejecting, by overwhelming majorities, both of the options endorsed by the Commons. As evidenced by their recent manifestos, although all three major political parties are now committed to reforming the Lords, there is no uniform agreement exactly on how this should be achieved.

Finally, it is important to point out that at, the time of this article going to print, the Prime Minister, in June 2009, issued an extraordinarily wide-ranging Formal Statement on Constitutional Renewal. He indicated that it was the intention of the government to "set out proposals for debate and reform" on various areas of the constitution, viz the rights of the citizen, devolution, the electoral system and public engagement in the political system. In relation to the House of Lords he made the following statement:

> we will move forward with reform of the House of Lords. The Government's White Paper, published last July, for which there is backing from other parties, committed us to an 80 or 100 per cent elected House of Lords, so we must now take the next steps as we complete this reform. The Government will come forward with published proposals for the final stage of House of Lords reform before the summer Adjournment, including the next steps we can take to resolve the position of the remaining hereditary peers and other outstanding issues.[133]

To put this statement into context, it should be remembered that the government's intention in July 2007 was that, following the earlier votes in March, a further White Paper would be published – informed by cross-party talks – so that a reform package could be produced and placed before the electorate at the next general election. The hope was that the other main political parties would include this commitment in their respective manifestos (although how this could be guaranteed is not clear).[134] One year later, Jack

131 Joint Committee on Conventions, First Report of Session 2005–06, Vol. 1, *Conventions of the UK Parliament*, HL Paper 265-I, HC 1212-I, (Norwich: The Stationery Office 2006).

132 House of Lords: Hansard HL Debs, 16 Jan 2007, vol. 688, cols 573ff; House of Commons: Hansard HC Debs, 17 Jan 2007, vol. 455, cols 808ff.

133 Hansard HC Debs, 10 Jun 2009, vol. 493, cols 797–8.

134 See n. 1 above, at col. 450.

Straw, in unveiling the promised 2008 White Paper, pointed out that, although the document represented "a significant step on the road to reform", it was not, however, the definitive blueprint. He also made it plain that it had "never been the intention to legislate in this Parliament".[135] More recently in March 2009, Michael Wills MP, the Minister of State, stated that the government was in the process of considering the responses it received to the White Paper and that

> More detailed plans for comprehensive reform will be developed and put to the electorate as a manifesto commitment at the next general election. Comprehensive legislation would then be possible in the next Parliament.[136]

In July 2009 the government published the Constitutional Reform and Governance Bill which, inter alia, proposed to allow members to resign from the House of Lords, remove those convicted of a serious criminal offence and end by-elections for hereditary peers. In September 2009, Jack Straw stated at the Labour Party Conference that legislative proposals for a new (elected) second chamber would be published shortly. Notwithstanding these recent developments, nevertheless it seems fair to argue that the final completion of the reform of the House of Lords is still some way off. After all, this is a constitutional issue which has bedevilled parliamentarians for almost a century and, in any event, any proposal for an elected or largely elected chamber is bound to encounter at least some resistance in the House of Lords itself. In conclusion, it seems inevitable that when we mark the centennial anniversary of the 1911 Parliament Act (which was only ever supposed to be temporary legislation), the second chamber will still be in the partly reformed state that it is in today.

135 See n. 30 above, at col. 21.

136 Hansard HC Debs, 17 Mar 2009, vol. 489, col. 1118W.

NILQ 60(3): 343–59

Justice without mercy

SEAN COYLE*

University of Exeter

The ground that I wish to explore in this essay has been covered before,[1] but there is (I think) some value in examining it further. Despite the previous coverage, it remains true that the jurisprudential thought of the modern era has maintained a steady focus on the idea of justice, but has paid much less attention to an important concept, that of *mercy*. An examination of indexes of the major texts of "the new liberalism" reveals many entries for "justice", but one will seek in vain for references to "mercy".[2] The neglect of mercy is not inexplicable, for it is famously associated with a number of paradoxes. For example, the idea of mercy depends upon its being conceived as a virtue that is in some way distinct from, and irreducible to, justice. Mercy, it is believed, moderates the operation of justice because it lies apart from the realm of law and justice, belonging instead to the domain of love, or compassion. Mercy *transcends* justice; but such transcendence would seem to involve a departure from justice (and therefore injustice). Were mercy always coeval with the requirements of justice, it would lose its identity as a separate virtue. Thus, mercy is either reducible to justice or, in undermining justice, ceases to be comprehensible as a virtue.[3]

Given this apparently paradoxical character, it is scarcely surprising that the major theories of justice largely ignore the idea of mercy. For if justice represents the *highest* ideal at which the enlightened polity should aim, then the perfection of society would appear to demand the constant refinement and realisation of that ideal rather than its abandonment in specific situations. Any imperfections arising from the operation of justice are interpreted as shortcomings in our perception of the ideal, and not indications of the limits of any such ideal. This attitude, which might be called "the idealisation of politics", is unfortunate. The sublimation even of liberal forms of social order has tended to reinforce feelings of dogmatic certainty, and to erode the very pluralism that liberal institutions notionally seek

* I am very grateful to Dr Fiona Smith for her valuable comments on an earlier draft.

1 See e.g. J Murphy and J Hampton, *Forgiveness and Mercy* (Cambridge: CUP 1988); R Harrison, "The equality of mercy" in H Gross and R Harrison (eds), *Jurisprudence: Cambridge Essays* (Oxford: Oxford University Press 1992), p. 107; N E Simmonds, "Judgment and mercy" (1993) 13 *OJLS* 52; J Tasioulas, "Mercy", CIII, *Proceedings of the Aristotelian Society* (2003), p. 101.

2 I derive the term "the new liberalism" from John Gray, *Enlightenment's Wake* (London: Routledge 1995), p. 29. Gray's argument addresses the hostility of the Rawlsian/Dworkinian tradition of liberal jurisprudence towards the value of tolerance, rather than of mercy, thus itself exhibiting the same general trend.

3 I summarise here Jeffrie Murphy's argument in "Mercy and legal justice", in Murphy and Hampton, *Forgiveness and Mercy*, n. 1 above.

to defend. It is in this sense an oddity that liberal thought should place a theory of justice at its centre. More importantly, a society that has forgotten mercy in its zeal for justice exhibits some of the least appealing characteristics of human nature. The most corrosive effects of the idealisation of politics have received deepest exploration not in philosophy, however, but in literature. Nor is this an accident, for the tendency towards idealisation and abstraction impedes exactly that humane understanding of which it is the function of literature to give expression. Morality is encountered, in literature as in life, as an active sensibility rather than a framework of general propositions; and it is only through a heightened awareness of this sensibility that we may come to appreciate (as Trollope did) the dangers inherent in the ambition: "Let justice be done though the heavens may fall."[4]

My argument will demonstrate the centrality of mercy to an understanding of political society and of the place of the juridical realm within it. In doing so I begin with an assumption: that "paradox" is a state which does not exist in reality. Paradox is rather the result of the application of a set of premises to an established body of ideas about reality. Because, generally speaking, the premises are not alighted upon wholly independently of the existing ideas but are in some way suggested by them, the correct response to a paradox is not to attempt to overcome the contradiction implied by those ideas whilst retaining as much of their integrity as possible, but rather to challenge the entire picture suggested by them, and find another way of looking at the world. That is the method adopted here. The initial discussion will focus not upon the character of mercy, but on the nature of our understanding of society. I argue that our understanding of the social world crucially depends upon the articulation of mercy as a value, and that unconsciousness of this dimension of value is one of the most damaging effects of the idealistic view of politics. Mercy, I argue, is first and foremost a *religious* idea, and that any political analogue of mercy must retain certain features of that intellectual inheritance if it is to make a genuine contribution to understanding. The argument concludes by suggesting that central features of the present "analytical" approach to jurisprudence must be given up if a clear understanding of either law or justice is to be achieved.

1 Law, justice and society

The predominant jurisprudential theories of the Western philosophical tradition are distinguished by their search for insight into the nature of law by reference to its place within a wider understanding of society. This intellectual strategy is in marked contrast to much of the jurisprudence of the present day, in which intellectual effort is directed towards the analysis of general features of law which are thought to obtain irrespective of the broader character of society. In his "Postscript" to *The Concept of Law*, for instance, Hart observes that his account of law is general "in the sense that it is not tied to any particular legal system or legal culture, but seeks to give an explanatory and clarifying account of law as a complex social and political institution".[5] When we encounter a writer such as Hobbes, on the other hand, we find an account of law that is derived from a particular understanding of the nature of human forms of association. Similarly, the Thomist natural lawyers sought an understanding of law by reference to a theological conception of the relationship between God and man, in which it is God's intentions regarding the condition and direction of human life toward a complex and collective goal, that informs our knowledge of legal concepts. These connections between jurisprudential understanding and theories of the "human condition" continued to be debated well into the nineteenth century; but the

4 This phrase appears a number of times in Trollope's works, but can be seen for instance in chs 61 and 62 of *The Last Chronicle of Barset* new edn (Harmondsworth: Penguin Classics 2002).

5 See H L A Hart, *The Concept of Law* 2nd edn (Oxford: Clarendon Press 1994), p. 240.

decline of this mode of thinking had in fact taken root much earlier. In Kant (for example), a new framework of thought is already fully manifest: his *Groundwork of the Metaphysics of Morals* is chiefly famous for its attempt to ground an understanding of human obligation in pure reason that is *severed* from all connections with metaphysical supposition and sociological observation:

> I do not . . . need any penetrating acuteness to see what I have to do in order that my volition be morally good. Inexperienced in the course of the world, incapable of being prepared for whatever might come to pass in it, I ask myself only: can you also will that your maxim become a universal law?[6]

The extent of Kant's detachment of the juridical realm from all concrete associations is made evident in his assertion that "Even the Holy One of the Gospel must first be compared with our ideal of moral perfection before He is cognized as such."[7] In the light of such treatment, Kant's proximity to modern modes of jurisprudential theorising is in fact much greater than his immersion within the older tradition of reflection on the nature of law. A common body of assumptions underpin both Kant's approach to legal questions and ours. These assumptions amount to the abandonment of the classical idea that morality is comprehended by reflection upon the actual conditions of historical forms of association, and instead perceive morality as a series of law-like rules finding an ultimate source in the will of the "rational agent". Laws are thus thought to embody a deep expression of autonomy. In this way, legal understandings are regarded as pertaining to the vertical relationship between the individual and the state, and to the complex matrix of horizontal relationships between individuals within the state.[8] By placing a conception of the "agent" at the centre, therefore, the Kantian inheritance forces upon us a certain structure in our theorising: one that is fundamentally removed from the older tradition of inquiry in which attention was focused on the law's ability to realise a human good that was believed to be attainable only in the presence of certain associative conditions. We, instead, see the law as an instrument for the projection of ideals that, in deriving from agency, represent insights finding an ultimate source outside history, in a transcendent dimension of reasons.[9]

I make these observations not in order to precipitate a debate about them, but so as to emphasise the fact that, despite its universalising ambitions, modern jurisprudence is wedded to a vision of human society that is neither the necessary nor the only understanding of society. In its most abstract terms, this understanding of human society might be seen to embody the following assumptions.[10] We confront the world, it may be supposed, as a set of conditions. These conditions, whilst of course being sufficient in all essential ways of supporting and sustaining life, are nevertheless imperfect when contemplated from the perspective of the ideal or preferred mode of life. Some of these conditions may be interpreted as "natural" or "given" states whereas others are brought about through human efforts. It is thus the function of human agency to seek to alter these existing worldly conditions in various ways. In undertaking such efforts, the overriding goal

6 I Kant, "Groundwork of the metaphysics of morals" in M J Gregor (ed.), *The Cambridge Edition of the Works of Immanuel Kant* (Cambridge: CUP 1996), p. 403.

7 Ibid., at p. 408. For some interesting discussion of the relationship between these propositions, see N Simmonds, *Law as a Moral Idea* (Oxford: OUP 2007), pp. 150–6.

8 For an excellent discussion of these assumptions, see S F C Milsom, "The nature of Blackstone's achievement" (1981) 1 *OJLS* 1.

9 I discuss these issues at length in my *From Positivism to Idealism* (Aldershot: Ashgate 2007), ch. 2.

10 Any attempt to state the general foundation of a body of theories will of course invite a chorus of objections that "this isn't what *my* theory is committed to". I nevertheless venture to hope that the reader will see in these assumptions a tolerably accurate, if inevitably somewhat blunt, depiction of the underpinnings of modern jurisprudential approaches.

must be the production of *improved* conditions rather than worsened ones, and some set of standards is therefore needed to guide the understanding of what represents improvement to the human lot: that is, some practical notion of what the "good" or preferred mode of existence consists in. All such systems of understanding (whatever be their vision of "the good" or the means of reaching it) deserve to be called "moralities". Since it is the function of morality to seek to *change* the world, it is then thought that moral insight (or whatever passes for it) cannot derive from an understanding of worldly conditions, but must come from elsewhere. Thus, it is often supposed that morality consists of general principles or ideals which together form an independent outlook upon the world of mundane and variable fact.

Because the institution of the ideal form of association requires the contemplation of the world, and of human action, from the perspective of general rules, a close and permanent relationship exists between justice (as a component of the ideal life) and law (as the form in which justice is both articulated and realised). Justice, as an idea, is thus simultaneously abstract and autonomous vis-à-vis the worldly conditions upon which it sits in judgment, and yet intelligible only within the context of collective social conditions. The attempt to implement justice in the world is always, then, the attempt to bring about the just *society*. This creates an interesting problem. For justice, being a creature of law and of society, must always be associated with the suppression of human freedom: the meaning of "society" and of a "form of association" involving precisely the avoidance of a state of total freedom, or anarchy. But where freedom is *itself* conceived to be a value, its suppression will come to be seen as an instance of injustice. Justice is therefore associated both with the creation of social conditions *and* their limitation. It is accordingly inevitable to the character of justice that the effort to realise within the dynamic forces of human society a tolerable justice, must end simply with the continual alternation between intolerable anarchy and intolerable tyranny.[11] (I refer here to the co-presence of anarchic and tyrannical elements *within* a form of association, rather than as names of specific kinds of association.)

It is perhaps not difficult to discern the influence of natural law thinking in this vision of the just society. The natural law tradition, in one sense, represented the belief that if human existence belongs to a rationally ordered cosmos in which it has an ultimate purpose and direction, nevertheless the full and final expression of that purpose lies outside the circles of the world. Divine law is the law that is natural and proper to the human condition; but it is always human effort and human interpretations which form the actual rules by which societies are governed.[12] So long as the *lex divina* was present within the world only in a reduced and attenuated form, human beings must confront a world that is chaotic and random as well as cosmically meaningful.[13] The secular moralities of the present day occupy a similarly ambiguous position in the modern world-view. In forming an autonomous standpoint for reflection, morality is presented as both the source of meaning of the human world *and* its judge. This is an interesting intellectual position which depicts in an especially direct way the failure of modern philosophy to resolve the tension of historical immanence and transcendence: morality embodies the absolute meaning of

11 I borrow this image (though not the sentiment expressed) from Reinhold Niebuhr. See R Niebuhr, "The Christian witness in the social and national order" in R McAfee Brown (ed.), *The Essential Reinhold Niebuhr: Selected essays and addresses* (New Haven: Yale UP 1986), p. 99.

12 See e.g. H Grotius, *De Iure Belli ac Pacis* (London: Kluwer Law International 1952), II.2.vi.91.

13 The theme of "the frailty of human reason" in interpreting the divine law was present in various forms throughout the canon of natural law writing in the seventeenth and early eighteenth centuries, and perhaps especially acute in Locke: see J Locke, *Essays on the Laws of Nature*, W Von Leyden (ed.), (Oxford: Clarendon Press 1954).

associative human relationships and not simply their present meaning, and in this it is transcendent; but it is also the *goal* of human society, and thus considered to be capable of historical implementation. It is the crucial failure of the jurisprudential and political thought of the present that it employs the idea of transcendence so as to absolutise rather than to criticise these partial achievements of history.[14]

This absolutising tendency is both symptomatic of, and offers encouragement to, an assumption concerning the historical significance of human efforts and the operative range wherein they can meet with success. "Law" and "justice" represent accomplishments that, in being rational, express the highest potentialities of the human spirit, bringing order where before was chaos, and meaning to life where there had been mere existence.[15] In this there is an important bridge between the tradition of natural law thinking and the secular political theories that came after. For whether or not such efforts are interpreted as having their source in the divine will, such divinity is imaginable only in virtue of its rationality. This, indeed, eventually paved the way for the removal of theological presuppositions from political thought: as Kant was to observe, "Even the Holy One of the Gospel must first be compared with our ideal of moral perfection before He is cognized as such"; and this ideal is one "which reason frames a priori and connects inseparably with the concept of a free will . . ."[16] The detachment of theology from politics resulted in a belief in rationality as the ultimate instrument of order and significance in the world. Thus, law and justice, as the supreme incarnation of rational principles, came to imply a view of human societies as the ultimate centres of order in the world. It then became possible to view the value-systems of these societies as the final arbiters of human good and evil. The upward potentialities of these efforts seemed limitless. Such sentiments, as Matthew Kramer has observed, lie at the heart of a liberal vision of politics:

> [T]he liberal philosophers had to introduce a new tone of public discourse to match their substantive outlook. Just as liberalism had overall been a salutary response to what had preceded it, so its characteristic tone would improve upon the tenor of discourse that had prevailed under the *ancien regime*. Christianity had been pessimistic, murderously intolerant, fanatical, and dogmatic; liberalism would hence be optimistic, generously open-minded, cool-headed, and responsive to rational persuasion. Truth would be tied no longer to sacred writings and divine revelations, but would be seen henceforward as the product of close analysis and wide-ranging debate.[17]

The most direct expression of this optimism, arguably, is to be encountered in Rawls's mighty book, *A Theory of Justice*.[18] It will be recalled that Rawls sought to demonstrate, by this work, that the idea of justice could be elucidated on the basis of uncontentious propositions that are accepted as a rational starting point for further reasoning. The book (and the method) were to have a profound re-orientating effect upon political philosophy that is partly explained by its advance over the intellectual environment into which it emerged. This environment was one that was dominated by utilitarianism, which depicted the modern society as a realm of conflicting preferences, the goal of politics being to maximize overall satisfaction of those preferences. The notion of justice was marginal to

14 See further, Niebuhr, "Optimism, pessimism and religious faith" in McAfee Brown, *Essential*, n. 11 above, at p. 6.

15 Perhaps the clearest expression of this view is to be found in Hobbes's characterisation of the transition from the state of nature to that of civil society: see T Hobbes, *Leviathan*, R Tuck (ed.) (Cambridge: CUP 1996), at ch. 13.

16 Kant, "Groundwork", n. 6 above, at p. 408.

17 M H Kramer, "The rule of (mis)recognition in the Hart of jurisprudence" (1988) 8 *OJLS* 401.

18 J Rawls, *A Theory of Justice* revised edn (Cambridge Mass: Belknap Press 1999).

such concerns, the focus of politics lying not upon the value or soundness of the preferences but rather upon their strength. Where justice was discussed, therefore, it came to be viewed as an issue of the practical and provisional balance to be achieved between randomly colliding syndicates advancing interests that must find some way to coexist. Rawls's theory challenged this picture by describing an "original position" wherein rational individuals seek to articulate, from behind a "veil of ignorance", a set of principles for the structuring and administration of their society. By grounding his argument in an "original position", Rawls produced a powerful vision of justice that was appropriate to conditions of pluralism in that an understanding of it *precedes* contentious understandings of "the good".

The intricate details of the Rawlsian theory are not relevant to the present discussion; for it is the general direction of the theory that has cast most influence over modern political thinking. It is (as Raymond Geuss has observed) remarkable that such a complex theory, consisting of more than 500 pages of dense and sustained argument, should culminate in a vision of the just and well-ordered society that is so striking in its resemblance to present constitutional arrangements of that most self-conscious example of liberal democracy, the United States. Moreover,

> [i]t strains credulity to the breaking point to believe that 'free and rational agents' (with no further qualifications), even if they were conducting a discussion from behind an artificial veil of ignorance . . . would light on precisely *these* arrangements.[19]

Despite the serious redistributive aims of *A Theory of Justice*, the points of convergence between the theoretical model and actual features of the liberal democratic society are sufficiently profound as to encourage the belief that such enlightened polities embody Rawlsian moral concerns. The "meaning" of this form of human association (and, by extension, all human association) is thus believed to be supplied by the precepts of the Rawlsian theory.

One obvious antidote to the belief that meaning is given to human affairs via the autonomous standpoint of theory has received little emphasis in modern jurisprudential argument: that the present condition of the liberal society owes much more to the social movements and culture of "permissiveness" which arose in the 1960s than to developments within the rarified world of academic philosophy.[20] Had more attention been paid to this aspect of social progress, the efforts of philosophy would have reflected to a greater extent the realisation that important dimensions of social meaning arise out of historical actions that are independent of philosophical concerns. The present point I wish to explore is, however, different. It concerns the consequences that result from those tendencies within modern philosophy which operate to deify certain features of the present social order.

I have described the attitude implicit within the Rawlsian theory of justice (and other, similarly conceived theories) as "the idealisation of politics". Social arrangements and institutions are taken as implying certain ideals, and the imperfection of present arrangements is treated as an indication that some refinement (and in some cases abandonment) of the established ideals is required, the better to reflect our most sophisticated and appealing conceptions of justice. Conceiving the goal of politics to be the identification and realisation of an ideal of justice, such secular philosophies ground the

19 R Geuss, "Liberalism and its discontents", in *Outside Ethics* (Princeton NJ: Princeton UP 2005), p. 22.

20 A *Theory of Justice* was first published in 1971 (though early versions of some of the book's arguments had appeared in article form in academic journals from the 1950s). It is perhaps worth emphasising that I focus on Rawls here as the most *direct* embodiment of the trends I wish to examine; the philosophical methodologies (reflective equilibrium etc.) integral to those trends have seen widespread use throughout political philosophy since the publication of Rawls's book.

belief that the social grouping (or form of association) is the ultimate source of meaning in human affairs. Rawls (for example) considers that rational agents in the "original position" will agree that issues of justice arise only in a context of scarce resources and conflicting preferences: if the allocation of resources to each individual carried no implications for the others, no-one would care who received what, and no question of justice would arise. Thus, the idea of justice makes sense only within the context of a shared form of association where men live in permanent proximity to one another. All values therefore make sense in terms of a mode of association, making the social grouping the ultimate source of meaning in human affairs, and the final centres of order in an otherwise random and chaotic world.

Where a general theory of justice is the ultimate source of value and meaning in this way, there is indeed no room for mercy in the administration of human affairs. The history of human effort can be seen as the progressive attempt to replace conditions of chaos with conditions of order and stability. The moralities which guide such efforts, inevitably, are moralities of *rules*, for which law supplies the archetype. Since mercy seeks specific *departures* from the rules, it will seem that mercy is on the side of the chaotic elements of human history against which the general part of human effort is set. The value of mercy, in standing in opposition to that of justice, remains unintelligible within the structures of meaning which ground the understanding of the human social condition.[21] It is not difficult to trace the logic of this outlook. The meaning of history (as a realm of human choice and action) is progress: the gradual removal of ignorance and immaturity, and the realisation of ultimate meaning. Because this ultimate meaning, and therefore the possibilities of order, are embodied in a *particular* form of association (the desired, or just society) then the first moral duty of humanity is to seek the universal implementation of this social form: that is, the elimination of "outlaw states" and the conversion of all regimes to that of the ideal.[22]

These dangerous and corrosive effects of the idealisation of politics are too familiar to the international politics of the day to require exploration here. Instead, I wish to explore an alternative conception of the world, which is absent from current political consciousness, but which (I believe) ought to inform its basis and pursuit. This alternative conception, I shall now argue, makes better sense of the relationship between justice and mercy.

2 Mercy and society

Philosophy, in an abstract sense, can be understood as the attempt to find meaning in the world. The wisest thinkers realise that this meaning is not to be discovered through the analysis of social forces, but requires a metaphysical perspective which relates the significance of those forces to an ultimate source of meaning that lies beyond them. In spite of the general hostility felt towards metaphysics within modern analytic philosophy, the mainstays of that tradition – the distinction of fact and value, the depiction of morality as an abstract and autonomous perspective on the world, emphasis on voluntarism etc. – enshrine just such a metaphysical position whereby ultimate meaning within human affairs (that which ought to be) is held to transcend the meaning of present conditions (that which is). The history of attempts to locate this ultimate source of meaning is instructive.

21 Nietzsche's view of the character of mercy is further demonstration of this: mercy, for Nietzsche, was the prerogative of the powerful sovereign to be employed as an emblem of power. Forbearance toward challenges to the sovereign authority demonstrated the security of that power, and its lack of diminishment in the face of challenges that lack the significance of a *necessary* response. See F Nietzsche, "The wanderer and his shadow", in *Human, All Too Human,* R J Hollingdale (trans.) (Cambridge: CUP 1996), s. 33. This line of thought is inherited from Seneca: see *Seneca: Moral and political essays*, J F Procopé and J M Cooper (eds), (Cambridge: CUP 1995), Essay 2: "On mercy", p. 134.

22 For the notion of "outlaw states" see J Rawls, *The Law of Peoples* (Cambridge Mass: Harvard UP 1999).

If there is meaning in the world, it might be supposed that it can derive only from one (or perhaps both) of two sources: either from human history (as a realm of freedom and action), or from nature (as a domain of forces that are in play independently of human action). The ultimate meaning of the human condition might then be thought to lie within the relationship between the two. A tradition of thought has existed from the earliest times which attributed to nature a moral significance. Natural events (the rich harvest, the poor harvest, the storm at sea) came to be interpreted as judgments upon the sinful condition of mankind, which could be influenced by prayer. This two-fold attitude toward the external world (of fearful obedience and thankful piety) generated two distinct but related interpretations of the relationship between humanity and the world, both of which are present in the biblical account of Genesis: on the one hand was the assumption that God had created the earth as a home for mankind, rich in resources and appropriate to the purposes of human flourishing; on the other, the earth was viewed as a hostile environment with no particular sympathy to human aims, to be conquered and tamed by human agency.

Both strands of thought can be traced throughout the history of legal thought, but are perhaps most clearly to be observed in the diverse canon of natural law philosophy in the seventeenth and eighteenth centuries. The seventeenth century proved to be an especially rich period for the development of juristic thought: the disintegration of religious unity in Europe, coupled with the gradual shift from old jurisdictional notions of "kingdoms" to something more closely resembling the modern "state", combined to produce a new understanding of human society, and thus of the place of the human being within the world. These new jurisdictional orders were no longer to be thought of as projections of the divine will, manifested in the claims of the ruling dynasties, but rather as independent zones of power and interest.[23] No grand plan could therefore be discovered in the relations between states whereby a final peace would emerge according to God's law; nor did continual warfare signal the painful birth-pangs of the new order of peace and harmony amongst nations, but simply the inevitable posture to be assumed as between such independent sources of absolute power. Being in competition with one another, these independent jurisdictions could no longer be thought to exist in order to realise or secure a common good, but had to be conceived simply as alternative domains of power operating to preserve their independence vis-à-vis each other. As jurists strove to understand the moral relations between these independent entities, it inevitably came to seem that relations between individuals within states must be treated in the same way: lacking an idea of the common good, the moral basis of the state could not concern the promotion of conditions conducive to the realisation of this preferred existence, but must instead consist in the protection and preservation of spheres of personal autonomy wherein the individual remains free of the will of others.[24]

These currents of thought, which served to place individuality at the heart of political understanding, were capable of development in various ways. Two such understandings were to be of particular importance for the future direction of jurisprudential thought concerning the relationship between law and society. The first was that of Grotius, for whom the purpose of law was the systematic protection of entitlements governing the moral relations between individuals who pursue independent and potentially conflicting goals. Such entitlements were thought to derive from a basic right of self-preservation inherent in the notion of a "human being", and expressed in the idea of the *suum* (or that

23 Such developments in thought did not, of course, occur overnight. Their discovery rather had the character of a gradual and deepening awareness of the implications of these new theoretical assumptions.

24 For a deeper account of this connection, see R Tuck, *The Rights of War and Peace: Political thought and the international order from Grotius to Kant* (Oxford: OUP 2001).

which properly belongs to a person).[25] This was essentially an eschatological notion, in that the existence of the *suum* was inferred on the basis of a theological view of the world as created by God so that man may survive and flourish; (only in such terms could there be a *right* to the means of survival, *suum ius*). Here was an image of the world as a domain of order and purpose in which the existence of man is subsumed within a wider cosmos that is ultimately related to and expressive of God's intentions. Viewed in such terms, the world is a domain of non-overlapping entitlements for which positive law is required simply for their clarification and enforcement. On the other hand was the view of Hobbes, for whom the basic premise of self-preservation implied not a harmonious realm of compossible entitlements, but rather a lawless world in which human interaction naturally takes the form of a war of each against all. In such a world, none but the most primitive set of assumptions could exist to guide human endeavour toward the attainment of peace; law could therefore only emerge as fabricated response to these basic conditions of the human predicament. In this way, Hobbes rejected the idea of a rationally ordered cosmos (like that of Aristotle) structured by compossible domains of *ius*, and instead represented the world beyond the boundaries of human society as a hostile environment from which escape, at almost any price, is necessary.[26]

Present within these variant pictures of the world were two distinct views of the character, not only of law, but of all human value-systems; and the tension between them has in large measure shaped all subsequent thought about the nature of morality and law. We are accustomed to addressing this tension from a number of related standpoints: from the perspective of pluralism vs absolutism; moral objectivity vs moral subjectivity; ethical relativism vs ethical realism; and so forth. But there is also a neglected, eschatological dimension to the tension which (as I shall argue) is of central importance for jurisprudential thinking on the relationship between justice and mercy. For it forces us to confront two distinct understandings of historical reality: one for which the world of human experience is interpreted pantheistically as domain in which all things that come to pass do so for a reason and have significance relative to an ultimate purpose; the other for which concrete reality is a corrupted realm of chaos and crude matter from which the spirit must detach itself. This latter perspective is informed by a variety of dualism which has itself taken numerous forms in the history of religious thought. It is present in the Hebrew division between the imperfection of the existing age and that of the perfect age to come; and it featured too, this time as a dualism of material and spiritual interests, in early Christian notions of the religious person's renunciation of material and earthly connections.[27]

Neither view of existence is an acceptable one, for both make the mistake of supposing that historical reality admits of purely rationalist explanation. Put another way, both perspectives on historical reality hold the ultimate meaning of existence to fall *within* the world. The unacceptability of either perspective is easily demonstrated: a pantheistic interpretation effectively sanctifies history, for every event and process is a contribution to the ultimate meaning of things. But though we do not know the meaning of the Holocaust (for example), we rob it of its tragedy if we believe its presence in history to be ultimately redeeming. A world in which every senseless act is related to a higher (if mysterious)

25 See Grotius, *De Iure*, n. 12 above, at I.1.iii.18.

26 See Hobbes, *Leviathan*, n. 15 above, chs 13–22. For detailed discussion see R Harrison, *Hobbes, Locke and Confusion's Masterpiece* (Cambridge: CUP 2002). Specific dimensions of Hobbes's treatment of *ius* and *lex* are analysed in my "Thomas Hobbes and the intellectual origins of legal positivism', (2002) XVI *Canadian J of Law and Jurisprudence* 243–70.

27 For an insightful discussion, see M Oakeshott, "Religion and the world" in T Fuller (ed.), *Religion, Politics and the Moral Life* (New Haven: Yale UP, 1993), p. 28.

purpose is one that must fail to comprehend the nature of human evil. Yet dualism represents no advance over this view, for a world in which events are wholly unrelated to transcendent values which make sense of them is intolerable. To assume that the Holocaust has no meaning beyond the brute facts of its occurrence, to accept that all human laws and endeavours are reflective of nothing but temporary meanings and base desires, is to condemn the world completely as a home for the spirit and to render all motivation finally otiose.[28]

I have raised these points because they reveal much about the assumptions that structure modern thinking on justice. For the modern juristic thinking shares with these perspectives the belief that rational historical explanations are possible, and this is to locate all structures of meaning within the realm of possible experience. Here, the transcendent context wherein the present meaning of human affairs is related to its absolute meaning is that of human history. Hence the final balance of judgment in all things falls within the scope of what is rationally intelligible (insofar as history itself is rationally intelligible). Thus, we are left with the implication that *either* the ideal form of the just society is (in principle) a realisable goal, *or* that there is nothing more to the idea of justice than can be discovered within the actual structures of meaning by which present society is ordered. My concern here is not with the differences between these positions, but with their essential commonalities. The crucial presupposition which unites both cases is that human communities represent the ultimate centres of order in the world. Thus, such structures represent the only means by which evil can be redressed; for in constituting the ultimate possibilities of order, human societies erect the final limits within which the forces of evil and disorder are contained. Historical progress is then equated with the eventual suppression of the randomising effects of human effort (and of nature), and the abolishment of unchecked evil. In this way, history is taken to represent the transition from barbarism and ignorance toward the highest forms of civility.[29]

Understood in this way, the only possible response to evil and disorder is the imposition of justice. The disorder is suppressed because a scheme of justice includes certain distributive goals in terms of which material goods, powers and liberties are apportioned so as to produce a rationally defensible outcome.[30] Similarly, the response to evil is achieved by way of judgment, and a fair and organised system of punishment. As justice is a *social* virtue, there can be no room for mercy in either context. For mercy in its purest sense represents the remittance of the consequences of evil by modification of the response to it: evil *demands* a response (if it is to he held in check), but the exercise of mercy is the decision not to exact the whole response upon the wrongdoer, but to reserve some of the suffering to oneself. Mercy is therefore always and exclusively the prerogative of the victim of evil; it is not a virtue that can be exercised on the part of anyone else. It is, therefore, incapable of being exercised by the organs of the state, or by any collective institution: the intelligibility of justice, as a social virtue, rests upon the demand ". . . that the state act on a single, coherent set of principles even when its citizens are divided about what the right principles of justice and fairness really are".[31] Any attempt by organs of the state to exercise mercy *on the part of* the victim is then a readjustment of this single set of principles, not the simultaneous application of two distinct systems of value.

28 I owe the basis of this argument to Neibuhr, "Christian witness", n. 11 above, at p. 15.

29 See e.g. F Fukuyama, *The End of History and the Last Man* (Harmondsworth: Penguin 1993).

30 See e.g. R Dworkin, *Law's Empire* (London: Fontana 1986), p. 165.

31 Ibid., at p. 166.

This aspect of mercy, and its relationship to society, has been explored by Ross Harrison.[32] Harrison's analysis is informed by a slightly different perspective, in that it concerns the close relationship that (he argues) exists between mercy and autonomy (the source of the merciful impulse being vital in falling within the choice of the person who bears the risks of his or her leniency). The state, in being responsible for its citizens, must rather make its decisions on the basis of their content, for which the appropriate criteria must be rationality and justice: mercy is not open to the state for it embodies precisely the denial of that keystone of formal justice, that like cases must be treated alike.[33] The details of this position are not of direct concern to the present argument; but what *is* of concern is one of the assumptions on which it rests. This is manifested most clearly in the *responses* to Harrison's claims. One such appears in an important essay by John Tasioulas, upon which I shall very briefly focus.[34] For Tasioulas, Harrison's understanding of the character of mercy belongs to a long-established sceptical tradition which challenges the *rationality* of mercy.[35] "The obvious problem with this contrast between individuals and organs of the state", he says, "is that mercy is inherently other-regarding, impinging heavily on the interests of those liable to punishment." In belonging to this matrix of interests, mercy would then seem to belong to the same area of ethical thought as justice. Thus, "Harrison's understanding of mercy as rationally ungoverned leniency leaves it mysterious what value it realizes, unless capricious deviations from justice are implausibly accorded value." Hence also, "he dresses mercy in the irrationalist garb favoured by its detractors, not its supporters".[36]

The assumption that I wish to tease out is that if mercy is a *rational* virtue, then it must be understood (as is justice) by reference to its ability to transform the structure of relationships that hold *within* a system of interests. If the role of mercy is to be explained in this context, then (by the usual meaning of "explanation") there must be a certain consistency in mercy's treatment of specific cases, and therefore a degree of abstraction in the criteria which govern its exercise. The burden of my remaining argument will not be to question this inference, but to undermine it at the root. I shall call into question the very idea that mercy belongs to this system of interests at all; and hence I will show that the comprehensibility of mercy transcends the narrow idea of "rationality" associated with this view.

3 The character of mercy

The preceding discussion suggested an alternative framework in which to contemplate the idea of human society. This, I argued, might be viewed as an attempt to locate the source of meaning in the world, and to expound that meaning. In terms of this framework, the modern outlook on politics can be loosely identified with the belief that human societies represent the possibility of meaning in a world that is otherwise chaotic and random. At the same time, this outlook manifests awareness of a possible (and indeed actual) gap between the *present* meaning of human social arrangements, and the *absolute* meaning of those arrangements. This absolute meaning (it is thought) cannot be found in the world, for it is yet to be fully realised by any worldly conditions; it must instead belong to a transcendent horizon of morality by reference to which existing conditions (or that which is) are

32 Harrison, "Equality", n. 1 above, at p. 117.

33 Ibid., at pp. 108–09. I have necessarily compressed Harrison's argument here.

34 Tasioulas, "Mercy", n. 1 above. Again, it is not my intention to explore this argument in detail, but simply to highlight an assumption. I hope therefore that the reader can excuse the very short treatment of a rich and complex argument.

35 Ibid., at p. 104.

36 Ibid., at pp. 104 and 106.

compared to a set of ideal conditions (or that which ought to be). But since this transcendent horizon is thought to be the product of human reflection (the original position, the act of interpretation, the "reasoned conviction" etc.), the absolute meaning of the human condition is thought nevertheless to be represented by a form of collective social organisation. Morality, in short, is assumed to concern, not the situation of human existence within a wider cosmos, but instead the much narrower realm in which human effort can manipulate and vary aspects of the social situation. It is, as I have argued, difficult to see how mercy can inform this process.

Suppose, instead, however, that human societies are not, in the above sense, the ultimate source of meaning in the world. Any attempt to uncover a perspective of absolute meaning betrays a religious impulse. The location of such meaning in ideal social arrangements, belief in the triumph of "secular society" over dogmatic superstition, and so on does not create a vision of human existence that is free from religious belief; it is simply the manifestation of a secular religion in which "law" has replaced "God" as the supreme mover against evil and disorder.[37] An understanding of mercy, then, is not one that must free itself of all religious association in order to make itself relevant to modern understanding, but must instead involve elucidation of the character of mercy as a religious idea. I therefore propose an eschatological vision of human society (and of the human condition) that is significantly divorced from that which informs modern jurisprudential reflection; and I shall argue that only within this alternative vision can mercy have its proper meaning. (I do not claim any originality for this vision, which is familiar within much Christian theology and, with certain important adaptations, that of other religious world-views.)

In order properly to understand the human condition, it is necessary to comprehend not only the relevance of law and judgment, but also the ultimate significance of evil. Evil requires judgment, for without some means of redressing the effects of evil, and of placing its occurrence within bounds, the human condition can possess no meaning at all. Life, whether in the primitive "state of nature" or the modern polity, is subject to various confusions and frustrations, but, unless there is some sense of an ultimate order by which evil is punished and good rewarded, there is nothing beyond the chaos of circumstances to lend it coherence. Hence, in *Leviathan*, Hobbes observes that the recognition of certain structural possibilities even within the "state of nature" can be exploited in order to effect escape to a better condition of life in which evil is checked and order imposed.[38] By giving these transitional postulates the character of "theorems" (that is, ratiocinations rather than externally imposed norms), Hobbes regards the worldly instantiation of peace and harmonious order as emphatically human achievements. Modern political thought has followed Hobbes in this, both secular philosophies *and* exponents of religiously inspired politics sharing the basic belief that justice in the world is realised through human action: either there is no God, and we stand alone as bringers of order to a chaotic world; or God exists and we are his instruments, effecting the suppression of evil in his name. In both cases too, justice (and by implication mercy) must be understood as *legal* virtues, for they seemingly represent a scale of values that cannot be understood apart from law.[39]

37 "Secular society", moreover, is not strictly speaking a *type* of society; it is rather an incomplete view of the whole of society. (It is in this sense on a par with "the tolerant society", "the wealthy society", etc. in which it is always possible to meet with intolerance, poverty and so on.)

38 See Hobbes, *Leviathan*, n. 15 above, ch. 14.

39 See e.g. Simmonds, "Judgment and mercy", n. 1 above, at p. 52: "mercy is not, as might first appear, a recognition of the extent to which non-juridical values such as that of love transcend the abstract and formal claims of law. Rather, mercy is itself inseparable from the framework of juridical thinking, exhibiting its distinctive and autonomous character only in the specific context of judgment."

Though I aim to dispute these connections, the position of "the moderns" in relation to justice unearths a valuable insight: that of the necessary relationship between justice and power. If justice is to exist as more than merely an abstract idea, but is actually to be done in the world, it must be exercised through law. Howsoever law might be said to place restraints and limitations upon power, its existence also depends upon and presupposes power. Justice in the world is thus not independent of power. Power in this sense is political power (in belonging to the realm of political concepts), and insofar as it requires enforcement it is also military or executive power. Power, therefore, though not an intrinsically evil idea, cannot be entirely divorced from evil ambitions and effects: for it is always and everywhere the projection of human ideals and interests which hold themselves out as sufficient or final centres of order in the world. Consequently, the imposition of justice by human agency does nothing to suppress or eliminate evil in the world (though it may effect the suppression of particular instances of evil), but actively perpetuates the struggle between opposing ideologies. This fact has received greater acknowledgment in the sphere of international politics than elsewhere (illustrated by contemporary concern over the Treaty of Versailles, for example), but its obvious implication has never successfully penetrated political consciousness: that the domain wherein human agency can achieve its goals is much more limited than has been supposed.

A proper understanding of justice and mercy thus requires a severing of the assumed connection between human agency and God's will (whereby agency is represented as an instrument for the ultimate triumph of good over evil). This is a bigger step than might be supposed in the context of "secular society", for it is easy to underestimate the extent to which the tradition of natural law thinking has shaped the modern juridical consciousness. Such thinking served to emphasise the Judeo-Christian religion as a *juridical* religion: the religious person lives his or her life under the guidance of moral *laws* finding an ultimate source in the divine will. Human evil is an affront to such laws, and thus to the authority of God; as such, all evil must ultimately be contained within a greater power which limits and judges it. This power, historically and theologically, is manifested not through direct intervention in the world, but is assumed to be effected by the earthly princes who serve as God's instruments. In this way, the foundation of political or prerogative power is presented as an extension of the divine authority. Such assumptions have not disappeared from the modern polity: the oath of allegiance and loyalty, at all levels of the political system, terminates in an act of recognition of monarchical authority deriving from the coronation. Nor is this a mere ceremonial relic, for the coronation is marked not by a political proclamation but by a religious rite that is central to its constitutional meaning. The detailed business of the day-to-day administration of the law does not, of course, pay attention to these ideas. But insofar as government and officers of the law continue to believe their exercise of power to be other than groundless and arbitrary, and so long as the organisation of their efforts is informed by considered values rather than random impulses, such underpinning assumptions have not receded utterly into the background but have simply undergone a transition. Religious impulse has not disappeared; it has merely elevated the "secular state" to the level of an object of faith and worship.

Suppose, in contrast, that we regard human societies not as the final, but only as premature and inadequate centres of order in the world. Here, we perceive the world in terms of a set of transcendent values that are not merely the possible projections of our present values. This is, inevitably, an eschatological standpoint: one that is not ultimately structured by social values at all, but instead exhibits belief that a scale of values exists that does not fully accord with the possibilities of human judgment. Leaving behind the immediate social meaning of justice and mercy (or more precisely their possible

meanings as social constructs), then, how might we think of those ideas within the realm of absolute meaning?

One possibility is presented by Christian theology. From this perspective, the world (in the form of nature, history etc) has no immediate overall moral significance, whilst yet possessing such meaning absolutely. The natural world and the world of human society are of course interpretable; but in occupying a space *within* history, we lack the perspective necessary to judge its final or overall significance. We are therefore in the position of knowing that the world has meaning, but not being able to comprehend fully what that meaning is.[40] Acceptance of this proposition is, in essence, the core of religious faith: to believe in the reality of a moral order in nature (i.e. the moral significance of natural and historical processes), and to reject the nihilistic possibility that life has no meaning at all beyond the fact of its subsistence. Faith of this kind requires the abandonment of belief in any straightforward, intelligible correlation of morality with natural and historical processes, for it entails the essential impartiality of divine justice in the following way. God's creation of the world, as something apart from himself, involves the realisation of freedom within it. But the creation of such freedom necessitates also the creation of ultimate limits in relation to the defiance this freedom implies. These limits must operate within the world (if the world is not to be dismissed as utterly spoilt and irredeemable), and not simply become present as judgment in the afterlife. However, in the absence of direct intervention, the justice by which evil is checked must fall indiscriminately as a judgment upon all: the good as much as the sinners, upon whom it rains and shines in equal measure. If any natural process is to be interpreted as belonging to this moral order (the death of an enemy from disease, the poor harvest, the shipwreck of a missionary voyage etc.) then it must be regarded as possessing no immediate or discoverable meaning, but rather an ultimate and incomprehensible meaning.

Yet the full meaning of the moral order is not exhausted by these ideas, for the very impersonality of justice (the wrath of God) seems incompatible with the idea of God as present within the Judeo-Christian tradition. The full meaning of the moral order is thus completed by God's *mercy*, manifested in the image of the crucified Christ. Mercy can therefore be explained in the following terms. Divine justice (the manifestation of God's power in the world) is an inescapable consequence of human freedom; but the nature of such justice is to be impersonal, so that the sun shines on both good and evil, and the rain falls on the good person and the bad alike. The justification for God's judgment is characterised by the fallen state of humankind: "the good man" is never absolutely good, the "worst of men" not irredeemably bad, and therefore (as with all justice) its imposition is deserved. As beings (according to the Christian story) we are imperfect, forever giving in to sin. Thus, if we are to be saved it cannot be justice which achieves this salvation, but rather mercy. And it is mercy that is manifested in the crucifixion of Christ: God the Father judges the world, but gives the world his only Son, and in submitting to rather than refusing agony, it is God the Son who "takes away the sins of the world".[41] The crucifixion then represents God's mercy (specifically that of Christ) in remitting the full consequences of justice by taking the final such consequences to himself. Where otherwise there would be inescapable damnation, there is the possibility of redemption.

I have set out these views about the nature of mercy because they seem to me to represent the only finally satisfactory understanding of the meaning of mercy. The paradoxes concerning the character of mercy are dissolved, because the framework in

40 See e.g. Neibuhr, "Christian witness", n. 11 above, at p. 14.

41 The words of the *Agnus Dei*, taken from John 1:29: "Agnus Dei, qui tollis peccata mundi, miserere nobis."

which mercy is finally comprehensible is not that of the attempted balancing of competing interests in society. Within that narrower framework, mercy seemed paradoxical for it did not make sense in terms of the conceptions of rationality that structure the framework.[42] Rather than accepting that framework as the ground for dismissing mercy as a coherent idea, I suggest that we instead retain mercy and dismiss the framework. Within the broader framework I have outlined, the tension between justice and mercy becomes finally explicable: mercy tempers justice, in mitigating its punitive consequences, but it does so by simultaneously standing as the culmination and fulfilment of justice.[43] I do not claim that in order to comprehend the value of mercy one must *accept* the Christian story (for of course many do not); I simply claim that an *understanding* of the story is a prerequisite for grasping the true meaning of mercy; for it is only in terms of this framework that (I believe) the idea of mercy is ultimately comprehensible. In the following section I shall sketch out some of the implications of this view of mercy for a jurisprudential understanding of law and society.

4 The role of mercy in jurisprudential understanding

I began this essay by mentioning the general absence of discussions of mercy in the arguments of modern jurisprudence. The reason for such lack of discussion can be put down to the general acceptance amongst jurisprudential scholars of a conceptual framework in which mercy has no obvious place. Modern jurisprudential arguments contain many fiercely competing understandings of the implications of this framework, but they do not often exhibit a willingness to make the framework itself an object of criticism. I have attempted in the foregoing discussion to bring into focus some of the main features of this framework: the focus on personal interests, the mechanism of justice, belief in rational solutions, the perfectability of society and so on. These ideas I have brought loosely together under the term "the idealisation of politics". Because the idealisation of politics involves the belief that the meaning of society itself is ultimately comprehended by a theory of justice, modern jurisprudential scholarship can be described without too much exaggeration as recommending the pursuit of justice "without mercy". I believe this to be an unfortunate and damaging direction of thought, and that its central claims, as well as its questions and focal concerns, ought to be given up.

Mercy, on the view I have been suggesting, is irreducibly a religious idea. Its operations are therefore not historical (though they are manifest in history), but cosmic. The appearance of paradox within the character of mercy is the result of a failure to grasp this fact. Mercy seems paradoxical because it is thought to concern the adjustment of relationships amongst interests that have already been determined by the value of justice (and thus to be in conflict with justice). However, when properly understood, mercy does not concern the further refinement of the balance between personal interests, for it does not address such interests at all. Its concern is rather with the possibility of redemption. The rationality of mercy therefore transcends the rationality of interests, in terms of which its relationship to justice remains incomprehensible. Mercy (in the broader terms I am suggesting) does not *annul* justice, for justice remains historically present as a necessary absolute limit to evil in the world; yet it completes the eschatological meaning of that justice in offering salvation in place of unavoidable damnation. Insofar as justice and mercy are

42 Thus Murphy, in "Mercy and legal justice", n. 3 above, at p. 174, dismisses mercy as a juridical virtue in categorical terms, stating that there "is simply no room for mercy as an autonomous virtue with which [a judge's] justice should be tempered. Let them keep their sentimentality to themselves, for use in their private lives with their family and pets."

43 I borrow the expression from Neibuhr, "Christian witness", n. 11 above, at p. 30.

present as moral or political ideas, their operations must be analogues of these cosmic movements. But for that reason, any such concepts will always be *imperfect* analogues: for to promote ideas of justice and mercy (or indeed any moral concepts) to the status of absolute principles, "correct" understandings, or as offering definitive guidance to human endeavour is to elevate a form of human association to the level of an eschatological end-point or ultimate meaning. Modern philosophy might then be characterised by its failure to appreciate that human society can be comprehended only by reference to a deeper set of values, and that the equation of these values with "the ideal society" serves not to illuminate, but to prevent the full emergence of this meaning.

Contemplation of the value of mercy is essential to an understanding of the nature of law and society. This is so, not because an awareness of mercy suggests any particular set of social arrangements as necessary or desirable, but because it promotes a greater sensitivity to the mutable and imperfect nature of all "progress", whether theoretical or practical. From this awareness comes a different conception of the role of law in society: for having given up the belief in the idea of "the ideal society", law is no longer associated (whether directly or instrumentally) with the production of that happy but far-off condition. Instead, law might be seen simply as a means whereby conditions are created or preserved in which human beings have space to "flourish".[44] The notion of human flourishing is itself a philosophical problem, but in general terms it might be said to involve the act of living a social life, and of exploring the meaning of one's existence within the various commonalities that make this existence possible. It should be apparent from the foregoing reflections that the character of this "flourishing" is not determined by the extent to which a specific set of external social conditions has been established; thus, the nature and substance of the commonalities embodied within the law are never fixed or static, but subject to continual change and restatement. Nor should they be presumed to point in any specific direction, or to represent a cumulative advance in the *same* direction at different times.[45]

The purpose of jurisprudence, it seems to me, is not therefore to ascribe a particular character to the law, but instead to explore the meaning and to clarify the implications of the commonalities to be found within the heart of law. In doing so, the jurisprudential scholar might hope suggestively to relate the substance of legal understandings to a deeper set of values that are not finally social but rather transcendent and eschatological. Perhaps the most important insight that could serve such an endeavour is the constant awareness that in seeking to relate the fluid and transitive to what is absolute and unmoving, the utmost care must be taken to avoid the presentation of the fluidity of real events and arrangements as something itself fixed in their truth or direction. Something of this concern possibly lies at the back of the following words of Gadamer's:

> Is not conscious distortion, camouflage, and concealment of the proper meaning in fact the rare extreme case of a frequent, even normal situation? – just as persecution (whether by civil authority or the church, the inquisition or any other agency) is only an extreme case compared to the intentional or unintentional pressure that society and public opinion exert on human thought.[46]

44 This term has evolved a series of quite specific meanings (see e.g. J Finnis, *Natural Law and Natural Rights* (Oxford: Clarendon Press 1980), ch. 4); but I use it more loosely here.

45 For an exploration of some of these themes in the context of adjudicative practices, see A W B Simpson, "The common law and legal theory", in *Oxford Essays in Jurisprudence* 2nd series (Oxford: Clarendon Press 1973).

46 H-G Gadamer, "Hermeneutics and historicism", reprinted as an appendix in *Truth and Method* 2nd edn (London: Continuum Books 1989), pp. 507–45, at p. 535.

The moral to be drawn from this is the need to avoid the tendency to make absolute that which is in reality mutable and transitive. This applies as much to the moral ideas that we take as fundamental as to the pliable legal rules through which they receive varying expression. Morality is best understood as an active sensibility which addresses a continually disordered array of values and circumstances that are permanently in motion. The processes of detachment and abstraction that inevitably inform moral decision are naturally inclined to suggest a picture of morality as a juridical structure of rules, rights and principles. A representation of morality along these lines vastly impoverishes our ethical understanding;[47] but its most corrosive effects lie in the elimination of mercy from the evaluative judgments concerning human relationships. If we do pursue such an understanding, we may come to find that a certain destructive mercilessness also characterises our social institutions through which such values are projected, defying all attempts to perceive within them a full and satisfactory expression of even the most basic moral concerns.

47 See my discussion in *From Positivism to Idealism*, n. 9 above, ch. 2.

NILQ 60(3): 361–80

A pandisability analysis? The possibilities and pitfalls of indirect disability discrimination

OLIVIA SMITH

School of Law and Governance, Dublin City University

Introduction

The prohibition of indirect race and sex discrimination has received considerable comment ever since the birth of the doctrine of disparate impact[1] under Title VII of the United States' Civil Rights Act 1964.[2] Following the doctrine's adoption on this side of the Atlantic, the focus centred mainly on the interpretation and application of indirect sex discrimination at European and member state level.[3] With the advent of the European Union's Framework Employment Equality Directive (hereafter the Equality Directive)[4] and the Race Directive,[5] the European anti-discrimination project has been extended to cover the grounds of race, religion, age, sexual orientation and disability, and there has been considerable discussion of the ramifications of this broad extension.[6] However, discussion of indirect disability discrimination has been less developed,[7] which reflects the situation with the pioneering Americans with Disabilities Act 1990 (hereafter, the ADA).

1 The term disparate impact is utilised in US employment discrimination law and the term indirect discrimination is generally agreed to capture the concept in European jurisprudence. Although some differences between the two doctrines remain, the terms are used interchangeably in this article.

2 *Griggs* v *Duke Power* 401 US 424 (1971). For a useful account of the uneasy place occupied by the disparate impact doctrine in US civil rights jurisprudence, see G Rutherglen, "Disparate impact, discrimination, and the essentially contested concept of equality" (2006) 74 *Fordham Law Review* 2313.

3 Some EU member states had introduced race discrimination legislation long in advance of EU initiatives.

4 Council Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation: OJ L303/16.

5 See Council Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin: OJ L180/22.

6 See E Barry and C Costello (eds), *Equality in Diversity: The new equality directives* (Dublin: ICEL 2003) and H Meenan, *Equality Law in an Enlarged European Union: Understanding the Article 13 Directives* (Cambridge: Cambridge UP 2007).

7 There is some discussion in R Whittle, "The Framework Directive for Equal Treatment in Employment and Occupation: an analysis from a disability rights perspective" (2002) 27 *European Law Review* 303 and L Waddington, *Implementing and Interpreting the Reasonable Accommodation Provision of the Framework Employment Directive* (Brussels: EU Network of Experts on Disability Discrimination 2004).

What accounts for this inattention to indirect disability discrimination in the literature and the case law when it generates such attention in the gender and race contexts?[8] It is suggested here that it has to do with the different matrix of disability discrimination laws, within which the much-discussed duty of reasonable accommodation predominates. Despite differences in approach, the reasonable accommodation duty, which first emerged in the ADA, is common to nearly all disability discrimination statutes. Described in general terms here, it requires reasonable adjustments to employment policies, practices and structures so as to take into account the particular needs of an individual with a disability in order to accord the individual equal employment opportunities. It is viewed as core to disability discrimination law and, given its absence from other discrimination law frameworks, its presence is determinative of the "canonical"[9] differences said to exist between disability discrimination legislation and race and sex discrimination legislation.[10] While disparate impact or indirect discrimination – generally described as the unjustified differential and disadvantageous impact of "neutral" employer policies and practices on members of particular "outgroups" – has long dominated US civil rights discourse,[11] the ADA's reasonable accommodation duty now attracts a comparable level of judicial and academic scrutiny.[12] Leaving aside the thorny issue of disabled status,[13] the vast majority of other cases under the ADA, and under its equivalents in Ireland and the UK, concern the reasonable accommodation duty, not indirect discrimination. Courts[14] and commentators[15] have largely ignored the ADA's disparate impact doctrine and this position has so far been replicated in many jurisdictions across the European Union. Echoing the earlier debate on disparate impact, much of the US literature has focused on the legitimacy or otherwise of the reasonable accommodation duty within the existing orthodoxy of an equality of opportunity framework.[16]

8 But see C Tobler, *Limits and Potential of the Concept of Indirect Discrimination* (Brussels: European Commission 2008) which discusses aspects of the relationship between indirect discrimination and reasonable accommodation.

9 C Jolls, "Antidiscrimination and accommodation" (2001) 115 *Harvard Law Review* 642.

10 Reasonable accommodation first appeared in the US Equal Employment Opportunity Commission's regulations published under Title VII in the context of religious discrimination.

11 Disparate impact law has been attacked in the US because, inter alia, of its lack of foundations in the Civil Rights Act. See Rutherglen, "Disparate impact legislation", n. 2 above.

12 Much of the debate in the US literature concerns the normative basis of anti-discrimination and accommodation principles. Various authors have argued that traditional aspects of the anti-discrimination structure are in essence accommodation mandates. See Jolls, "Anti-discrimination", n. 9 above. Cf. S Schwab and S Wilborn, "Reasonable accommodation of workplace disabilities (2003) 44 *William and Mary Law Review* 1197. Others argue that the duty is an example of a "legally mandated form of positive action" in favour of disabled people. See P Karlan and G Rutherglen, "Disabilities, discrimination and reasonable accommodation" (1996) 46 *Duke Law Journal* 1, at p. 9. Cf. M Crossley, "Reasonable accommodation as part and parcel of the antidiscrimination project" (2004) 35 *Rutgers Law Journal* 861.

13 A considerable number of ADA cases are dismissed on summary judgment in favour of defendants on this issue. See C Feldblum, "The definition of disability under federal anti-discrimination law: What happened? Why? And what can we do about it?" (2000) 21 *Berkeley Journal of Employment and Labour Law* 91. This is not as problematic an issue under the Irish legislation given the width of its definition of disability. While the Equality Directive fails to define disability, note the disappointing ECJ decision in *Chacon Navas* v *Colectividades* [2007] ICR 1.

14 In an isolated reference, the US Supreme Court in *Raytheon* v *Hernandez* 540 US 44, 53 (2003) recognised that "disparate impact claims are cognizable under the ADA". This comment was, however, *obiter*.

15 A recent exception is the discussion by M Stein and M Waterstone, "Disability, disparate impact, and class actions" (2006) 56 *Duke Law Journal* 861.

16 See Crossley, "Reasonable accommodation", n. 12 above, and Jolls, "Antidiscrimination", n. 9 above.

A similar situation persists in the UK, though for somewhat different reasons: under the UK's Disability Discrimination Acts 1995–2005 (hereafter the DDA), traditional indirect discrimination claims are not justiciable due to the absence of a separate prohibition on indirect discrimination.[17] Despite a number of amendments to the DDA, many driven by EU law, the UK legislature still opted against a "pure" implementation of the indirect discrimination provision. Instead, it took advantage of the provision in the Equality Directive,[18] discussed below, which allows member states to choose to implement the indirect discrimination provision by means of the duty to make reasonable accommodation (adjustments).[19] It has been argued that the DDA's reasonable adjustment duty embraces the indirect discrimination provision.[20] This position will be interrogated below as part of the discussion of what role, if any, a specific prohibition on indirect disability discrimination has in disability discrimination protection.

A further contrast is the position under Irish law, where, despite the express inclusion of a statutory definition of indirect disability discrimination alongside the reasonable accommodation duty in the Employment Equality Acts 1998–2004 (hereafter, the EEA), the former provision, like its US counterpart, has remained under-utilised.[21]

There is room, therefore, for further discussion on the role and reach of the indirect disability discrimination provision and also its relationship to the reasonable accommodation duty. A recent discussion by Stein and Waterstone in the ADA context raises interesting issues for the European experience: these commentators reconsider the ADA's disparate impact doctrine in light of the original purpose of the statute.[22] They highlight the doctrine's capacity to offset the individualised approach of the accommodation duty through its focus on workplace practices and norms, which disadvantage groups of disabled people. While the central premise they forward – that a disparate impact analysis usefully captures the inequalities experienced by a considerable number of disabled people – is attractive, there are a number of practical problems with its application under the current matrix of European and Irish anti-discrimination law. As will become clear below, Ireland's legislative framework provides a useful site for the discussion because it includes *both* indirect discrimination and reasonable accommodation in the disability context.

This article proceeds as follows. It outlines the premise and scope of the reasonable accommodation duty. While the duty is undeniably important in the scheme of disability rights, it is not without limitations. Indeed, the recognition of these limitations has

17 The DDA originally defined unlawful disability discrimination in two ways: first, less favourable treatment for a reason related to disability which was capable of justification; and second, discrimination as a failure to comply with the duty to make reasonable adjustments. The transposition of the Equality Directive ensured the inclusion of a "pure" direct discrimination provision, which is incapable of being justified.

18 Article 2(2)(b)(ii).

19 During council negotiations on the directive it was felt that the "reasonable accommodation" duty was a sufficient means of addressing indirect discrimination "since many if not all of the obstacles that arise through indirect discrimination can be removed by invoking such an obligation": *Disability Discrimination Law in the EU Member States*, Baseline Study (Brussels: EU Network of Independent Experts on Disability Discrimination 2004), p. 14.

20 "The functional equivalent of indirect discrimination is the section 6 duty to make reasonable adjustments.": M Connolly, *Townshend-Smith on Discrimination Law: Text, cases and materials* 2nd edn (London: Cavendish 2004), p. 492.

21 A review of the EEA's disability cases reveals only two (unsuccessful) cases litigated under the indirect disability discrimination provision. See O Smith, "Side-stepping equality: disability discrimination and generally accepted qualifications" (2008) 30 *Dublin University Law Journal* 279.

22 The preamble to the ADA implores employers to remove all "artificial, arbitrary, and unnecessary barriers" and to eliminate the "built-in headwinds" of the conventional work environment: 42 USC §12101.

prompted commentators to consider more closely the role of indirect discrimination in the disability context. The heuristic device of "pandisability", discussed below, has been suggested as capable of injecting renewed purpose into the ADA's disparate impact provisions. The basis of this device is outlined and then considered against the backdrop provided by the Equality Directive and Ireland's employment equality legislation.

Disability discrimination and the predominance of reasonable accommodation

Traditional employment discrimination law scholarship draws a bright line between race and sex discrimination on the one hand and disability discrimination on the other.[23] The view that disability discrimination is inherently different is so prevalent that one scholar has described it as "canonical".[24] The difference between the two categories of statutory schemes (race/sex and disability) can be traced to discussions around the "relevance" of such characteristics to the issue of equal employment opportunity. Basically put, the common belief now prevailing with respect to race and sex discrimination is that this phenomenon is unjust because these characteristics generally bear no relation to issues such as job capability, competence, achievement of qualifications and suitability.[25] The liberal principles of individualism and meritocracy decree that persons competing for market positions should be considered objectively on the basis of their individual merits, and not on the basis of irrelevant characteristics such as race or sex.[26] The process, in short, should be colour blind and gender neutral. In employment discrimination law this idea is formulated in the prohibition against direct discrimination[27] across an array of grounds and with regard to various stages and aspects of the employment relationship. The basis of this provision is to prohibit "less favourable treatment" on the grounds of race, sex etc. of individuals whose circumstances are not materially different.[28]

On the other hand, disability and impairment retain relevance to the enquiry because, it is argued, disability is a biological reality that equates with lower productivity.[29] This point presupposes that the equal treatment principle is inapplicable to disability because disabled individuals, due to their impairments, are generally not similarly situated with their non-disabled counterparts. This account can be challenged as being over-inclusive and need not sidetrack the discussion at this point.[30] However, circumstances remain where the strict comparability approach alone would be insufficient to capture the nature of disability-based discrimination occurring in, for example, the interaction between an impairment and a feature of the working environment, which makes disability "relevant" to the enquiry. In such cases, the response needs to move beyond the narrowness of formal equality and

23 Stein and Waterstone, "Disability", n. 15 above, at p. 895.

24 Jolls, "Antidiscrimination", n. 9 above, at pp. 643–44.

25 One major exception is the bona fide occupational requirement where the sex of an individual may be a requirement of the job: section 25 of the EEA.

26 For a critique of the "neutrality" of these merit standards, see I M Young, *Justice and the Politics of Difference* (Princeton: Princeton UP 1990), ch. 7.

27 The term disparate treatment is used in the United States and generally equates to the concept of direct discrimination in European law, although differences remain in application.

28 See s. 6(1) of the EEA which sets up this approach.

29 See Schwab and Willborn, "Reasonable accommodation", n. 12 above. There are problems with this type of generalised assertion with respect to causality and, consequently, in terms of responses to disability inequality. See Crossley, "Reasonable accommodation", n. 12 above.

30 The given nature of the previous statement can be attacked. It may remain the case that a disabled person and a non-disabled person are similarly situated in the governing context, i.e. similar in terms of qualifications, experience, capability, notwithstanding the presence of an impairment, which may not require an accommodation. The problem here tends to be in respect of assumptions, fuelled by stereotyping or prejudice, that this is not the case.

recognise that disability *can be* relevant to employment opportunity and, where this is the case, it ought to be taken into account accordingly.[31]

However, because of dominant perceptions which equate it with incapacity, the relevance of disability is most often viewed in negative and exclusionary terms, so that the presence of an impairment resulting in different levels of functioning often forecloses employment opportunity altogether.[32] For a long time disability and employment had been – and, in many cases, continue to be – viewed as mutually exclusive.[33] This entrenched construction of disability in terms of personal incapacity, misfortune and economic dependence has been widely refuted by social and minority model theorists within disability discourse.[34] This analysis, according to disability theorists, locates the problem of disability within the person and the contributory force of external factors, which operate to disable and exclude individuals with impairments are ignored. In short, the biological account of disability is incomplete and, therefore, challengeable. There are obvious parallels between this analysis and formerly held dominant social conventions that viewed women as less capable than men, and race as a biological absolute.[35] While such (widespread, at least) views of sex and race have, in the main, been left behind, they have been much more difficult to dislodge in the disability context. Courts and many commentators retain considerable difficulty in seeing disability as anything other than an identifiable and verifiable medical problem of the individual.[36]

What is clear from the above account is that any disability equality initiative predicated solely on the equal treatment principle would be an impoverished one. Developments at the level of equality theory and within disability discourse have combined to deconstruct both the "equality as sameness" paradigm and the "neutral" forces that view disability as an individual problem, which explain and legitimate non-participation in the labour market and other aspects of social life. The particularities of the disability issue have helped contribute to the development of more substantive equality norms. In this context the focus has rested almost exclusively on the individual reasonable accommodation duty.

Article 5 of the Equality Directive outlines the rationale of this development: it equates the equal treatment of persons with disabilities with the provision of "reasonable accommodation" and

> [t]his means that employers shall take appropriate measures, where needed in a particular case, to enable a person with a disability to have access to, participate in, or advance in employment or to undergo training, unless such measures would impose a disproportionate burden on the employer.

31 G Quinn, "Rethinking the place in difference in civil society – the role of anti-discrimination law in the next century" in R Byrne and W Duncan (eds), *Developments in Discrimination Law in Ireland and Europe* (Dublin: ICEL 1997), p. 64, at p. 77.

32 This statement is not intended to exclude those with impairments with little functional impact (e.g. physical disfigurements) but who, nevertheless, endure exclusion due to stigma.

33 See A Silvers, "Formal justice" in A Silvers, D Wasserman and M B Mahowald, *Disability, Differences, Discrimination: Perspectives on justice in bioethics and public policy* (Lanham: Rowman and Littlefield Publishers 1998), p. 13.

34 See M Oliver, *The Politics of Disablement* (London: Macmillan 1990). For discussion of disability as a minority group issue, see H Hahn, "Anti-discrimination laws and social research on disability: the minority group perspective" (1996) 14 *Behavioral Sciences and Law* 41.

35 Stein and Waterstone, "Disability", n. 15 above, at pp. 895–6 for discussion.

36 See comments of the Irish Supreme Court in *Re Article 26 and the Employment Equality Bill 1996* [1997] 3 IR 321, at p. 367.

The idea behind the duty is to tackle established workplace and social norms, which erect barriers to the equal employment opportunities of disabled people.[37] The response, therefore, requires alteration of the work environment and/or practices, where reasonable, in order to accommodate individuals across a wider range of functioning levels.[38]

Notwithstanding the real significance of reasonable accommodation as a disability equality initiative, it is constrained by factors internal to its statutory design (such as the limits inherent in the qualifier "disproportionate burden") and by factors pertinent to discrimination law enforcement more generally. Despite the assumed group objectives attributed to legal rights at the policy level, at the operational level, the duty is highly individualistic in its assessment of the kind of accommodation necessary to achieve the participatory goals of a particular disabled person in the particular workplace at issue. In this sense, reasonable accommodation, where required, "always involves an individual assessment and a tailored individual solution".[39] The group dimension appears to be ad hoc at best: in the strongest, almost altruistic sense, the duty may be described as anticipatory only if it inspires employers to think about accommodation requirements in advance and the alteration of practices and policies so as to be as inclusive as possible to all. In this sense, there is some equivalence between the accommodation duty and the principles of Universal Design.[40] In practical terms, the duty does not operate in this manner in many statutes; for example, under the UK's DDA, the duty only becomes operable upon an employee or applicant specifically making an accommodation request.[41] Thereafter, "an interactive dialogue" between the parties should, ideally, take place. The more realistic perspective is that employers overlook the underlying broader objectives attributed to the duty and react only when faced with an accommodation request from a protected individual. On this view, the basis of the duty, therefore, is to accommodate individuals into existing structures without systematically challenging the structures, institutions and norms on which they depend. Further, the enforcement mechanisms institutionalised within discrimination law support this analysis. If a complainant successfully pursues a failure to accommodate claim against his or her employer, where the suggested accommodation imposes no disproportionate burden, then the specific workplace policy is, in theory at least, altered for that individual. There is little to no discussion whether other individuals with disabilities beyond the complainant may benefit by the accommodation duty.[42] This atomistic, individualistic conception of the accommodation duty, combined with the longstanding problematic enforcement strategies of discrimination law, limits its transformative effect.[43] However, as is discussed below, the assumed inevitability of individual relief under the accommodation duty may be overstated as "[m]any requests for accommodation dealing

37 For a robust critique of the duty, see S Day and G Brodsky, "The duty to accommodate: who will benefit?" (1996) 75 *Canadian Bar Review* 433.

38 It was expected that the duty would help tackle the consistent and endemically low employment rate of disabled people. In Ireland in 2004, 37% of people of working age with a disability/longstanding health problem were in work, compared to 67% of other working age adults: National Disability Authority, *A Strategy of Engagement – Towards a comprehensive employment strategy for people with disabilities* (Dublin: NDA 2007), at p. 3.

39 Waddington, *Implementing and Interpreting*, n. 7 above, at p. 8.

40 See text to n. 49 below.

41 This presupposes that employees/job applicants are aware of the duty and is problematic because of reported low levels of awareness of equality rights: ESRI/Equality Authority, *The Experience of Discrimination in Ireland: Analysis of the QNHS Equality Module* (Dublin: ESRI 2008).

42 Stein and Waterstone, "Disability", n. 15 above, at p. 899.

43 See generally, Brodsky and Day, "The duty to accommodate", n. 37 above.

with physical accessibility and environmental design, for instance, can be commonly remedied by use of universal design principles".[44]

Research in the US demonstrates that the reasonable accommodation duty has not dramatically transformed the employment situation of disabled people.[45] Significant reductions in the unemployment rates of disabled people in Ireland do not appear to be emerging despite the existence of disability equality provisions for near on a decade.[46] How useful the duty is to job applicants with disabilities seeking access to the labour market remains unclear. A clear pattern appears to be emerging from the body of caselaw generated under the Irish legislation: most complainants raising the reasonable accommodation duty are job incumbents, or dismissed former employees, and not job applicants. It seems far easier for a job incumbent to rely on the reasonable accommodation duty than it is for a job applicant.[47] The system relies on an individual complainant perceiving his or her situation in litigation terms, which, if successful, may result (more often than not) in an award of compensation for discriminatory dismissal or discriminatory treatment consequent upon a failure on the part of respondents to consider their obligations under the reasonable accommodation duty. There are only a few examples in the caselaw of courses of action ordered under the accommodation duty which allow an individual to continue in employment.[48] For the most part, litigation reveals that the duty often boils down to a consideration of the legitimacy of an employer's actions: either to accommodate a particular individual employee or not, and whether the decision taken was legitimate given the parameters of the duty as set out in the particular statute. Where an accommodation is requested by a disabled individual and is refused, the status quo remains.[49]

44 Stein and Waterstone, "Disability", n. 15 above, at p. 903. Universal Design is an architectural concept and its goal is to design products that are usable by all people, to the greatest extent possible, without the need for adaptation or special design. Cited in text to n. 143 at p. 893. See Center for Universal Design, Principles of Universal Design at www.design.ncsu.edu/cud/about_ud/udprinciples.htm (last accessed 6 October 2009).

45 D Stapelton and R Burkhauser, *The Decline in Employment of People with Disabilities: A policy puzzle* (Lanham: Rowman & Littlefield Publishers 1998).

46 In the 30-month period between the two Central Statistics Office surveys on disability in 2002 and 2005, the employment rate for people with a disability fell from 40.1% to 37.1% despite overall employment growth of 5.6% over the period: NDA, *A Strategy*, n. 38 above, "Executive summary", p. 3. It is simplistic to assert that the introduction of disability equality law is a sufficient means of effecting the integration of disabled people into employment. Other issues remain pertinent and include: access to education; the issue of disability benefits, in particular the "benefit trap"; working time issues; health issues; not forgetting the prevailing economic climate. See ibid. and M Weber, "Beyond the Americans with Disabilities Act: a national employment policy for people with disabilities" (1998) 46 *Buffalo Law Review* 123.

47 See the narratives discussed in M Russell, *Beyond Ramps: Disability at the end of the social contract* (Monroe: Common Courage Press 1998), p. 118.

48 In *Feore* v *Alzheimer Society of Ireland* DEC-E2006-101, the remedy ordered by the Equality Tribunal included the offer of the position the complainant would have held had she not been absent from work on sick leave. Most failure to provide reasonable accommodation claims have arisen as a consequence of discriminatory dismissals and the remedy awarded is generally compensation, although reinstatement and re-engagement are options available to the tribunal under the Acts.

49 In *A Computer Component Company* v *A Worker* ED/OO/8, the complainant was sacked due to the apparent incompatibility between her impairment and a rather infrequent duty of her position – to operate a particular machine. For breach of the duty, which would have excused the complainant from using this machine, she was awarded compensation and not the job. This occurrence is prevalent across the caselaw.

Due to the predominance of these individualistic issues, a group dimension, the often-attributed benefit of indirect discrimination[50] seems less obvious in respect of the accommodation duty. As Stein and Waterstone point out

> Much heavy weather is made of the heterogeneity of disability with the result that, rather than being viewed as systematically excluded by the environment, disability is held to be the by-product of individual workers not fitting into particular workplace circumstances. Consequently, assertions of disability discrimination have been closeted into a narrow category that examines the reasonableness of a particular accommodation to a single individual rather than questioning the larger issue of whether a hostile workplace environment was constructed that excluded employees with disabilities.[51]

These commentators reference a combination of factors which explain the reluctance to formulise failure to accommodate claims under a theory of disparate impact. One prominent reason is that the failure to accommodate claim is viewed as a stand-alone substitute for disparate impact litigation. A second reason for the US federal courts' reluctance to engage with disparate impact under the ADA is the belief that Title I's statutory basis of disparate impact is less clear than that which exists under Title VII of the Civil Rights Act. As shall be discussed below, both of these explanations hold little weight in the Irish context given the EEA's express inclusion of a stand alone indirect disability discrimination provision, alongside the reasonable accommodation duty. It is submitted that factors relating to the "difference" of disability discrimination, difficulties in surmounting the individual, as opposed to group, effects of the experience of disability exclusion, alongside specifics internal to the statutory design of the indirect discrimination, account for the under-use of the indirect disability discrimination provision.

The nature of the reasonable accommodation duty

Some discussion has taken place on the nature of the duty as a positive or negative right: in other words, whether it is a positive duty placed on employers, which implies a right to an accommodation, or a negative provision where a refusal of an accommodation amounts to discrimination. This issue has been discussed in detail by Waddington.[52] The angle of interest for this discussion is the consequence of a failure to accommodate and whether this amounts to actionable discrimination of a particular kind, whether direct, indirect or a so-called "third way".[53] This reveals the fuzzy relationship between the orthodox constructions of discrimination law (namely, direct and indirect) and the reasonable accommodation duty. In EC law, Article 5 of the Equality Directive expressly provides a relationship between the principle of equal treatment and the duty to provide reasonable accommodation. However, the directive does not explicitly call the denial of reasonable

50 The group benefits of indirect discrimination have been questioned: Schiek argues that "group disadvantage has always been the starting point for indirect discrimination, but it has never established group rights": D Schiek, "Indirect discrimination" in D Schiek, L Waddington and M Bell (eds), *Cases, Materials and Text on National, Supranational and International Non-Discrimination Law* (Oxford: Hart Publishing 2007), p. 323, at p. 330. On this point, there is a real distinction between the doctrinal path suggested by Stein and Waterhouse, discussed below, and its applicability to the European context, which is the absence of a tradition of class action litigation in European discrimination law.

51 Stein and Waterstone, "Disability", n. 15 above, at p. 897.

52 See Waddington, *Implementing and Interpreting*, n. 7 above, at pp. 41–4.

53 See L Waddington and A Hendricks, "The expanding concept of employment discrimination in Europe: from direct and indirect discrimination to 'reasonable accommodation' discrimination" (2002) 18 *International Journal of Comparative Labour Law and Industrial Relations* 402, who argue that a failure to provide a "reasonable accommodation" is a form of discrimination *sui generis*.

accommodation a form of discrimination.[54] Tobler argues that the directive treats the duty as "a specific obligation of the employer, to which corresponds a specific right on the side of the employee with a disability" and she deems it unnecessary and unadvisable to label the refusal to provide reasonable accommodation "indirect discrimination".[55] Yet, as is discussed below, there is a link in EC law between indirect discrimination and reasonable accommodation under Article 2(2)(b)(ii).

In Irish domestic law the status of the EEA's reasonable accommodation was not originally clarified. Not surprisingly, it has yet to be considered whether a failure to accommodate claim can be analysed in disparate impact terms. In its early stages, despite no explicit connection between the statute's definitions of discrimination (both direct or indirect), the interpreting tribunals had mainly viewed failure to accommodate cases as an independent limb of actionable discrimination, without characterising it as either direct or indirect. A purposive interpretation of the statute clearly demanded this approach because the statutory provisions dealing with remedies are dependent on a finding of "discrimination".[56] However, in *A Worker (Mr O)* v *An Employer (No 1)*,[57] the Labour Court expressly rejected the proposition that the EEA's original accommodation duty amounted to a freestanding right: that is, that an individual has a right to a reasonable accommodation where the burden on the employer did not give rise to any extra-nominal costs.[58]

In section 6(1), discrimination is defined in the traditional manner of less favourable treatment on the basis of any of the protected grounds, including disability. Indirect discrimination is provided for later in the statute.[59] However, section 16(1) provides a defence to employers by providing that the Act does not require employers to recruit, promote, retain or train an individual for a position if that person is not fully competent and fully capable of undertaking the duties of a position. However, the section 16(1) defence is itself qualified by section 16(3), which set out, in a rather convoluted manner, the original reasonable accommodation provision: in essence it provides that a person with a disability shall be considered fully competent for a position if with special treatment or assistance they would be fully capable for the position and an employer is required to do all that is reasonable to accommodate the person by providing such special treatment or facilities.

In *A Worker (Mr O)* v *An Employer (No 1)*, the Labour Court interpreted the scope of the reasonable accommodation duty under the 1998 Act strictly within the confines of section 16. It noted that notwithstanding section 6(1), section 16(1) allows an employer to treat a disabled person less favourably than others. This defence is available where a disabled person is not fully capable of carrying out "all the duties attached to the job for which they applied". However, the employer defence is, in turn, qualified by subsection 3, outlined above, which imposes a duty on employers, where reasonable, to provide such special treatment/facilities so as to render the disabled person competent and capable of the job required of them. This link between section 16(3) and section 16(1), according to the Labour Court, made it clear that the reasonable accommodation duty did not found "an independent cause of action for an employer's failure to provide special treatment or

54 The definition of discrimination under Art. 2 of the UN Convention on the Rights of Persons with Disabilities (CRPD) expressly includes the denial of reasonable accommodation.

55 Tobler, *Limits and Potential*, n. 8 above, at Part IV, para. 5.

56 S. 77 of the Act as amended specifically grants a person who claims "to have been discriminated against by another in contravention of this Act" the right to seek redress in various fora: s. 77(1)(a).

57 [2005] ELR 132.

58 This was the original standard set out in the 1998 Act before the Equality Act 2004 altered it to the European standard of "disproportionate burden".

59 See ss 22 and 31 of the EEA 1998–2004.

facilities". Further support for this position, according to the Labour Court, is available because no part of the Act defined discrimination as including a failure to fulfil the reasonable accommodation duty. Thus, prior to the amendments introduced by the Equality Act 2004, the only way the duty to accommodate became operative was as an antidote to an employer's reliance on the defence in section 16(1) allowing it to treat a disabled person less favourably. Consequently, where the employer did not seek to rely on section 16(1), the duty had no application. The amendments to the 1998 Act, brought in by the Equality Act 2004, do appear to allow a freestanding cause of action. This is because the language in the amending clause places an independent duty (unlinked to section 16(1)) on employers to provide reasonable accommodation to disabled workers and applicants.[60]

Despite this positive development, the amendments do not expressly link the failure to accommodate with the possibility of an indirect discrimination claim – but neither do they expressly forbid this approach. This ambiguity is not unique to the Irish statutory framework and it permeates others, such as the equivalent provision under the ADA. One reaction to the perceived limitations of the duty has been an increasing inquisition into its purpose and effectiveness. As referred to already above, a number of studies in the United States[61] demonstrate that the ADA has failed to challenge the in-built headwinds facing disabled people accessing employment. There are, of course, many contributory factors, not all of which can be attributed to the "failure" of discrimination law protections. However, increasingly, the structural foundations of the ADA's reasonable accommodation duty are under scrutiny from which a "backlash" against the federally driven disability provisions has emerged.[62] Simultaneously, however, this can create room for discussion of a more constructive nature: in particular, it leaves space to consider the role of disparate impact in the disability context and its interaction with or influence on the reasonable accommodation duty.

The next section goes on to outline the genesis of the indirect discrimination doctrine and, at a general level, considers the possibilities for the provision in the disability context. Given that one of the key ideas behind indirect discrimination has been to expose the adverse effect of putative neutral practices on groups, the discussion considers whether the doctrine can offset the individualised focus of the accommodation duty.

The development of indirect discrimination

Within equality theory the traditional construct of the equal treatment principle has long invited a range of criticisms.[63] For one, because the principle of equal treatment concentrates on excising differential treatment stemming from racial prejudice and gender stereotyping, it is predicated on a version of equality which generally accepts the given backdrop of existing labour frameworks and other institutional environments and does not demand that they are radically restructured. Thus, even if unequal treatment was successfully tackled through the prohibition on direct discrimination, powerful and established institutional and societal arrangements maintain existing distributive patterns and hierarchies, which benefit certain groups at the expense of others.[64] As Collins continues, these institutional arrangements are both formal and informal; for example,

60 S. 16(3)(b) of the EEA 1998–2004.

61 See R Colker, " The Americans with Disabilities Act: a windfall for defendants" (1999) 34 *Harvard Civil Rights–Civil Liberties Review* 99.

62 See L Hamilton Krieger, "Backlash against the ADA: interdisciplinary perspectives and implications for social justice strategies" (2001) 21 *Berkeley Journal of Employment and Labor Law* 1.

63 See N Lacey, "From individual to group? A feminist analysis of the limits of anti-discrimination legislation" in *Unspeakable Subjects: Feminist essays in legal and social theory* (Oxford: Hart Publishing 1998), at p. 19.

64 H Collins, "Discrimination, equality and social inclusion" (2003) 66 *Modern Law Review* 16, at p. 30.

formal rules requiring full-time work for entry to pension schemes impact adversely on part-time workers who are disproportionately female because of more informal social norms that see women taking primary responsibility for caring obligations.[65] Likewise, full-time work norms can disproportionately impact on those disabled people unable to commit to full-time work due to medical reasons. Minimum occupational requirements such as formal education qualifications – for example, state examinations – can impact adversely on certain disabled people denied mainstream education because of accepted social norms that deem many disabled people as uneducable.

Indirect discrimination, it is argued, purports to capture the systemic differential impact which longstanding, assumed ordinary and "neutral" practices can have on different groups. Recognised by the US Supreme Court in *Griggs* v *Duke Power Company*,[66] it was held that Title VII of the Civil Rights Act 1964

> . . . proscribes not only overt discrimination but also practices that are fair in form, but discriminatory in operation. The touchstone is business necessity. If an employment practice . . . cannot be shown to be related to job performance, the practice is prohibited.[67]

The first recognition of the *Griggs* principle on this side of the Atlantic was in the UK's Sex Discrimination Act 1975. At the European level, indirect sex discrimination was eventually recognised in community law through judicial interpretation of Article 141 of the EC Treaty.[68] The European Court of Justice jurisprudence was codified in 1997 in the Burden of Proof Directive in cases of sex discrimination.[69] This definition has recently been reformulated in the Amended Equal Treatment Directive.[70] A further definition of indirect discrimination applicable to the newer grounds is set out in the Framework Employment Equality Directive.

The major strength of indirect discrimination is that it conceptualises discrimination as involving more than intentional, episodic or individualised wrongdoing.[71] It attempts to expose the discriminatory by-product of so-called neutral practices and it requires employers either to justify the continued operation of practices which deny individuals from particular groups equal employment opportunities or, where justification cannot be made out, to desist from their use.[72] The doctrine's approach would, prima facie, suggest it is underpinned by a more substantive conception of equality as compared with the individualistic modification of practices for a single employee under the accommodation analysis. However, the translation of these ideas into law has been "extraordinarily problematic".[73] These difficulties are multifaceted and include issues which have long plagued the gender equality agenda and which are not developed here, such as: the differing

65 Collins, "Discrimination", n. 64 above. This policy was originally challenged as indirect sex discrimination prior to the introduction of Council Directive 97/81/EC Framework Agreement on Part Time Work OJ L 14, 20.1.1998.

66 *Griggs* v *Duke Power* 401 US 424 (1971).

67 Ibid., at p. 431.

68 *Bilka Kaufhaus* v *Weber von Hartz* [1986] ECR 1607.

69 Art. 2(2) Directive 97/80/EC on the burden of proof in cases of discrimination based on sex [1998] OJ L14/6.

70 Directive 2002/73, OJ [2002] L269/15: Art. 2(2).

71 C Gooding, *Disabling Laws, Enabling Acts: Disability rights in Britain and America* (London: Pluto Press 1994) pp. 38–9.

72 "The standard judicial remedy in a Title VII disparate impact case requires the employer to change the policy or standard for everyone, not just the protected group.": Schwab and Willborn, "Reasonable accommodation", n. 12 above, at p. 1238.

73 A McColgan, *Discrimination Law: Text, cases and materials* 2nd edn (Oxford: Hart Publishing 2005), p. 76.

definitions of indirect discrimination;[74] the constant amendment to the definitions; proof of the constituent elements of disparate impact, including judicial wrangling over the tests utilised to prove acceptable degrees of disparate impact;[75] the use and (non)-availability of statistical evidence;[76] and the breadth of the justification defence. Further, the group effects of indirect discrimination have also been queried, in particular, the extent to which the principle has inclusionary aspects which oblige employers to take into account and accommodate the needs of individuals with certain characteristics. Consequently, there has been a measure of disappointment with the indirect discrimination doctrine, which Schiek attributes to

> overburdening the prohibition of indirect discrimination with the expectations of achieving all the substantive aims of discrimination law and policy at large, although indirect discrimination law is only a small part of equality law and policy.[77]

Background to indirect disability discrimination

According to one US federal judge, the disparate impact doctrine emerged in response to ". . . situations where, through inertia or insensitivity, companies were following policies that gratuitously . . . excluded black or female workers from equal employment opportunities".[78] In the US neither judges nor scholars have felt comfortable replacing "black or female workers" with "disabled workers".[79] Yet the ADA includes statutory provisions consistent with the theory of disparate impact as recognised under Title VII.[80] Further, both European and member state legislatures expressly included a definition specific to the disability ground, although the EC adopts an alternate optional approach in Article 2(2)(b)(ii). Presumably, those legislative bodies assumed that the doctrine had some role to play in addressing disability discrimination so as to merit its inclusion in the statutory framework. Put another way, the express inclusion of the indirect discrimination doctrine gives rise to a presumption that the inequality endured by disabled people can manifest itself in a manner conducive to a disparate impact analysis. In fact, the traditional understanding of the way ordinary and commonplace practices, structures and institutional arrangements indirectly and adversely impact on members of particular groups seems a very cogent way of explaining the type of disadvantageous exclusion endured by disabled people. The well-documented experiences that disabled people have of discrimination and inequality are less often direct. Disability inequality is often the product of practices and procedures that have been designed with the needs and requirements of the average, able-bodied individual in mind.[81] For example,

74 Ireland's EEA originally contained four definitions of indirect discrimination, delineated by pay, other aspects of the employment relationship, gender and the other eight protected grounds.

75 On the issue of adverse impact, see R v *Secretary of State for Employment, ex parte Seymour Smith and Perez* [1999] ECR 623 (ECJ) and [2000] 1 All ER 857 (HL).

76 Ibid.

77 See Schiek, "Indirect discrimination", n. 50 above, at p. 332.

78 *Finnegan* v *Transworld Airways Inc.* 967 F2d 1161, 1164 (7th Cir 1992).

79 Stein and Waterstone, "Disability", n. 15 above, at p. 886.

80 No single provision under the ADA deals with disparate impact. Rather, there are a number of provisions consistent with disparate impact theory developed under Title VII. For example, the ADA prohibits employers from "utilising standards, criteria or methods of administration . . . that have the effect of discriminating on the basis of disability". It also prohibits employers from using qualification standards, tests or other selection criteria that screen out or tend to screen out disabled people unless the standard, test or other selection criteria is shown to be "job-related . . . and . . . consistent with business necessity".

neutral rules which permit access to benefits based on continuity of service can exclude individuals who take medical leave; rules which exclude part-time work opportunities can exclude disabled people who traditionally have greater health needs; rules which set break times at pre-designed slots can affect individuals with medical requirements; rules which require communications to be effected in a particular manner can exclude individuals with cognitive or sensory impairments. Whether these are issues of accommodation or disparate impact and the effect of this characterisation is explored below.

Given the prima facie utility of indirect discrimination principles to the disability ground, it is necessary to consider what drives this reluctance to conceptualise disability-based discrimination in disparate impact terms. Perhaps the most commonly expressed reason for this limited consideration of the disparate impact doctrine is the lack of homogeneity among disabled people as a class. The heterogeneous nature of disability is thought to render a disparate impact analysis less potent and the individualisation of disability as a personal issue, particularly in employment litigation, is critical to this argument. For example, Rutherglen attributes the rarity of ADA class actions to "the predominance of individual issues, such as the nature and extent of a person's disability and the cost of accommodations".[82] Unlike sex discrimination cases, where assumptions are generally made (often leading to charges of essentialism) about the impact of particular policies or practices on women as a class, the disability category, it is argued, is simply too diffuse to ground a disparate impact analysis.

It has been suggested, however, that this barrier to disability indirect discrimination can be surmounted. Borrowing from the term "panethnicity" utilised in US class action suits, Stein and Waterstone develop a similar concept of "pandisability" in response to the heterogeneous issue. The notion of panethnicity was key to early Title VII disparate impact and class action litigation in the United States; it describes "the heuristic processes through which ethnic minority groups that might internally consider themselves heterogeneous are externally perceived by the non-group majority as homogeneous".[83] Panethnicity involves the grouping together of different nationalities/ethnicities under the banners of African-Americans or Asian-Americans in order to allow diffuse groups to pursue race-based disparate impact litigation.[84] Instead of viewing the perceived heterogeneity of the disability class as a barrier to disparate impact analysis, the pandisability approach sees disability as a unifier: it is a minority group issue which allows individuals with diverse impairments to be treated alike for group identification purposes. In this way, wheelchair users, individuals with learning disabilities, individuals with cognitive impairments, despite differences at the bio-medical level, can be categorised as forming a class on a more fundamental level because of their shared experiences of exclusion, animus and prejudice. There are clear links between the pandisability concept and the social model of disability

81 Brodsky and Day have queried this account of the neutrality of ordinary standards and common practices as unintentional. The whole scale exclusion of disabled people from mainstream social, economic and cultural practices is a systematic phenomenon that can hardly be described as unintentional or as a simple by-product of development: "Duty to accommodate", n. 37 above, at p. 458.

82 Rutherglen, "Disparate impact", n. 2 above, at p. 2319, at n. 43: "the few disparate impact claims under the Americans with Disabilities Act of 1990 that are likely to be successful concern denial of access to government services or public accommodations under Titles II and III". See *Sunrise Dev Inc* v *Town of Huntingdon* 62 F Supp 2d 762, 766–7 (EDNY 1999).

83 Stein and Waterstone, "Disability", n. 15 above, at p. 870, relying on Y Le Espiritu, *Asian American Panethnicity: Bridging institutions and identities* (Philadelphia: Temple UP 1992).

84 Courts in the United States have allowed groups such as Korean, Japanese and Chinese Americans to be represented under the banner of Asian-Americans in class action litigation. This approach has also been utilised by Latinos/Hispanics.

within disability discourse: antecedents of the pandisability concept can be traced to basic social model theorising where disability is viewed as a form of social oppression imposed on individuals with impairments. It is the social consequences of disability, as expressed in the commonality of prejudicial experience, which in turn reinforces disparate individuals' membership in the disability classification.[85] Viewed in this way differences in impairments and degrees of functional impact can be put to one side. This development of pandisability is key to the analysis of the failure to accommodate limb of disability discrimination law in disparate impact terms and is considered below.

A disparate impact analysis of failure to accommodate claims

The issue for discussion here is whether a failure to accommodate an individual with a disability can ground a cause of action of indirect disability discrimination. In other words, whether a neutral provision or practice which places persons with a disability at a particular disadvantage and which an employer has not offset by way of reasonable accommodation can be analysed in indirect discrimination terms. This question is interesting because of the responses generated following each cause of action. Under the traditional reasonable accommodation duty (where successful), the response requires the adjustment of a policy/practice in favour of a single individual. Otherwise, the policy or practice concerned remains untouched. However, the response in the case of a policy found to have a disparate impact would be striking down or adjusting the policy for all members of the group: otherwise there is a risk of further litigation from members of the same class.[86]

There is some discussion in the literature on the similarities between the two types of statutory protection. One strand of the discussion of the accommodation duty in the US has arisen from the need to stabilise it within the boundaries of equality of opportunity theory in order to insulate it from judicial attack and public and academic backlash.[87] To this end, Christine Jolls' seminal work demonstrates the similarities – in terms of doctrinal reach and practical response – between the ADA's accommodation mandate and the disparate impact doctrine under Title VII.[88] Jolls' analysis of several cases under orthodox anti-discrimination law shows that a successful disparate impact suit involves the alteration of pre-existing employment practices in favour of the protected group: she argues that "important aspects of disparate impact liability are in fact accommodation requirements".[89] In other words, an accommodation type of remedy was provided in many Title VII disparate impact cases. The main success of disparate impact in the US has been to tackle the use of job specifications that are unrelated to job performance but which hinder the uptake of employment opportunities and benefits on the part of women and ethnic minorities. For example, Jolls has highlighted caselaw which has successfully attacked the disparate impact of "neutral" grooming rules – no-beard rules on black men, job selection criteria, including standardised tests, on African-Americans generally, rules excluding the

85 Stein and Waterstone, "Disability", n. 15 above, at p. 900.

86 One objection to the disparate impact analysis is the need to alter the practice in different ways due to the differing impact of the practice on persons with different types of disabilities. However, this approach raises the issue of remedy, as opposed to the actual cause of action. In the history of class action litigation in the US, where disparate classes united to challenge the effects of particular employment practices, cases were not hindered by the fact of diverse remedies being provided, where successful.

87 In *Board of Trustees of the University of Alabama* v *Garrett* 531 US 356 (2001), the Supreme Court held that the states are not subject to private suits for monetary damages under Title I of the ADA. See also, Hamilton Krieger, "Backlash against the ADA", n. 62 above.

88 See generally, Jolls, "Anti-discrimination", n. 9 above.

89 Ibid., at p. 651.

use of leave and their effect on pregnant women, and English-only rules for their effect on ethnic minorities.[90]

What Jolls omits to consider is whether the ADA's failure to accommodate cause of action can itself be subject to a similar disparate impact analysis. Stein and Waterstone go on to present the possibilities opened up by a disparate impact analysis under Title I of the ADA. This analysis, they argue, has the capacity to offset the hitherto individualised focus of the reasonable accommodation duty by pointing to employer practices and procedures that disproportionately affect groups of disabled people. In this sense, litigation moves beyond a consideration as to whether a particular reasonable accommodation should have been provided and considers whether the workplace norms adopted by the employer adversely impact other disabled individuals beyond the claimant. These commentators then adapt Joll's Title VII disparate impact scenarios with accommodation remedies to ADA disparate impact examples, such as:

(1) rules on the use of physically inaccessible venues when alternative accessible venues are available (such as placing a workstation up a flight of stairs where it is dangerous for someone with a balance disorder);

(2) job selection criteria that tend to exclude people with disabilities (e.g. quadriplegic bank teller to lift boxes, or standardised testing or written tests that severely dyslexic persons may not be able to perceive);

(3) rules that require the use of specific systems that adversely effect individuals using alternative formats (e.g. Braille, large print etc.);

(4) refusals of leave time/alternative work venues.[91]

They argue that applying basic disparate impact theory to the accommodation duty would enable people with disabilities to challenge the type of "neutral" workplace policies and practices commonly challenged under Title VII. It may be that the consequence of a successful disparate impact analysis will be the alteration of the practice, criterion or provision that gives rise to the disparate impact, where it remains unjustifiable; this has the advantage of providing a remedy which is inclusive and redistributive as opposed to compensatory and exclusive. The discussion below considers whether these arguments translate to the European and Irish practice of indirect discrimination and reasonable accommodation.

From theory to practice: the statutory design of indirect disability discrimination

The EEA and indirect discrimination

Section 31 of the EEA, as amended by the Equality Act 2004, sets out the definition of indirect discrimination applicable to the non-gender grounds. This is a technical amending provision, the net effect of which is to apply the definition of indirect discrimination that operates in respect of the gender ground to the other eight grounds. The definition of indirect discrimination applicable to disability reads:

> Indirect discrimination occurs where an apparently neutral provision puts persons of of a particular disability at a particular disadvantage in respect

90 Jolls, "Anti-discrimination", n. 9 above, at pp. 652–66.

91 Stein and Waterstone, "Disability", n. 15 above, at pp. 915–16.

> of any matter other than remuneration compared with other employees of their employer.[92]

Where this paragraph is satisfied, the employer is deemed to be discriminating against the claimants unless the provision is objectively justified by a legitimate aim and the means of "achieving that aim are appropriate and necessary".[93]

On its face, the definition may attempt to guard against some of the obvious disadvantages faced by many disabled people within a society predicated on ostensibly neutral yet exclusionary norms. For example, minimum educational requirements are demanded on the assumption that educational opportunity is generally available to all – unnecessary educational qualifications have long been recognised as capable of giving rise to disparate impact;[94] full-time work norms are demanded because historically this has been a cornerstone of capitalist production and it places "equal" demands on all constituents of society. However, many neutral policies, procedures and practices of the working environment have differential effects on groups of individuals who depart from the full-time, face-time norm of the paradigm worker.[95] Much use has been made of the indirect discrimination principle in the gender context where putative neutral policies such as height and weight requirements,[96] unnecessary length of service/continuous service requirements[97] and full-time work norms[98] have been found to bear more heavily on women, particularly older women and women with caring responsibilities. However, there have been numerous gaps in the coverage of indirect sex discrimination which has lead to a reassessment of the principle's potential to transform established workplace and social norms.[99] Many of these issues appear to repeat themselves in the disability context. Further, there are aspects to the definition and further provisions within the Act which appear to threaten the operability of the provision in the expansive manner contemplated above.[100]

It is not entirely clear whether the current definition can facilitate a pandisability analysis of disparate impact claims as developed by Stein and Waterstone in the US context. The view expressed here is merely tentative and is based on a combination of factors that influence the construction of disability in assessing equality claims.[101] Many commentators have taken the view that "reasonable accommodation" discrimination is different because the

92 S. 31(1)(a) of the EEA as amended. The definition of indirect discrimination in the Equality Directive includes contingent harm, i.e. it covers the possibility of adverse impact as opposed to the actual occurrence of adverse impact. The EEA definition does not reflect this aspect of the directive. See E Ellis, *EU Anti-Discrimination Law* (Oxford: OUP 2005), pp. 91–4.

93 S. 31(1)(b) of the EEA 1998–2004. A similar provision extends the indirect discrimination protection to pay practices, but this is subject to s. 35(1).

94 See *Griggs* v *Duke Power* 401 US 424 (1971).

95 See M Travis, "Recapturing the transformative potential of employment discrimination law" (2005) 62 *Washington and Lee Law Review* 3.

96 *Allcock* v *Chief Constable Hampshire Constabulary* ET case no 3101524/97.

97 *Falkirk Council* v *Whyte* [1997] IRLR 560.

98 *Inoue* v *NBK Designs Ltd* [2003] ELR 98.

99 See C Barnard and B Hepple, "Substantive equality" (2000) 59 *Cambridge Law Journal* 562 and Tobler, *Limits and Potential*, n. 8 above, at V.5.

100 Specifically, s. 36 of the EEA 1998–2004. For discussion, see Smith, "Side-stepping equality", n. 21 above.

101 For a discussion of the effect of textualism as the dominant mode of interpretation and its influence on disability rights, see W Parmet, "Plain meaning and mitigating measures: judicial interpretations of the meaning of disability" (2000) 21 *Berkeley Journal of Employment and Labor Law* 53.

> disadvantage is not necessarily experienced by all or most members of a particular group, but is . . . experienced on the individual level, depending on both individual and environmental factors.[102]

In response the pandisability approach borrows from the heuristic device of panethnicity; it allows otherwise heterogeneous groups to form a homogeneous class on the basis of a shared experience of prejudice, stereotyping and exclusion for the purposes of disparate impact litigation. However, it may be the case that the language of section 31 defies this type of argument: there must be a neutral provision which "puts persons of a particular disability" at a particular disadvantage compared with other employees of the employer.[103] The critical issue will turn on how "persons of a particular disability" is interpreted. If interpreted strictly and literally, it may limit indirect discrimination claims to instances of "particular disadvantage" demonstrated only by individuals with similar impairments.[104] It is interesting to consider whether the approach would be different if the statutory provision provided that the neutral provision would put "disabled people" at a particular disadvantage. Uncertainty remains as to how the term "persons of a particular disability" will be interpreted. It would be unsurprising if this aspect of the statutory provision were to be interpreted from a bio-medical perspective on disability which would demand that each member of the class of "individuals with disabilities" deviate from the perceived bodily norm in the same way. A broader, yet still medically inspired classification could be utilised, such as the approach taken in the statutory definition of disability.[105] A more generalised categorisation system would support a social model perspective on disability and move away from the minutiae of individual impairments and medical labelling. This approach would embrace the pandisability analysis, as suggested by Stein and Waterstone above. It has the added advantage of focusing on the impact and effects of the practice, which denies groups of individuals employment opportunity, as opposed to the characteristics of the complainant, a focus that is unfortunately inverted in many jurisdictions with respect to the definition of disability.[106] However, the likelihood of this broader interpretation must be considered against the backdrop of a consistent judicial deference to medical conceptions of disability.[107]

Aside from this issue, other barriers remain to a pandisability analysis of disparate impact claims under Irish law. The need to isolate a particular practice giving rise to the adverse impact is another hurdle which must be surmounted. In other words, it is simply insufficient to survey the demographic make-up of a particular place of employment pointing to a statistical absence of disabled people in relation to their presence in the available labour market: as Stein and Waterstone put it, there is little scope to argue that the environment generally is hostile to the presence of disabled people.[108] Applicants must demonstrate the existence of a particular employment practice which causes the disparate

102 Waddington, *Implementing and Interpreting*, n. 7 above, at p. 82.

103 Given the range of impairments and their varying functional effects, some individual claimants may have "great difficulty in identifying a particular group that is disadvantaged by the relevant provision, criterion or practice in the same manner as they are": Whittle, "The Framework Directive", n. 7 above, at p. 308.

104 An ADA plaintiff may prove a disparate impact by demonstrating that an employer's policy "screens out or tends to screen out an individual with a disability or a class of individuals with disabilities". The plaintiff need not demonstrate an adverse effect on a class of persons with disabilities. See 42 USC s. 12112(b)(6).

105 See the definition of disability in s. 2(1) of the EEA 1998–2004.

106 Most notably the United States and the United Kingdom.

107 See generally, Parmet, "Plain meaning", n. 101 above. Note also the views of the Irish Supreme Court in *Re Article 26 and the Employment Equality Bill 1996* [1997] 3 IR 321 and the ECJ in *Chacon Navas* v *Colectividades* [2007] ICR 1.

108 Stein and Waterstone, "Disability", n. 15 above, at p. 897.

impact for disabled people. Where explicit formal practices are in operation, this task is less formidable – formal entry qualifications, promotion procedures etc. However, where a combination of informal norms and subjective practices collide, it becomes less easy to extract a specific practice or criterion as giving rise to the disparate impact. It may simply be a repetition of the inaccessibility of indirect discrimination norms to informal criteria that are often determinative of access to benefits and promotions in senior and professional positions on other grounds. This is where a disparate impact analysis of failure to accommodate can operate to tackle workplace culture and more hidden, subjective norms that combine to exclude disabled people from employment opportunities.[109]

The relationship between reasonable accommodation and indirect discrimination at European level

This part moves on to question the applicability of the pandisability argument to failure to accommodate and indirect discrimination at the level of the European Union. The key provision for consideration is Article 2(2)(b)(ii), which provides that a provision, criterion or practice will amount to indirect discrimination where

> it would put persons having a particular disability . . . at a particular disadvantage compared with other persons unless i) [it is] objectively justified by a legitimate aim, or (ii) as regards persons with a particular disability . . . the employer . . . is obliged to [make a "reasonable accommodation"] in order to eliminate the disadvantages entailed by such provision, criterion or practice.[110]

This provision allows member states to choose to use the duty to make reasonable accommodation to remove the barriers and disadvantage caused by the indirectly discriminatory practice. This reading of Article 2(2)(b) suggests that, as long as the employer provides a reasonable accommodation to a disabled employee/applicant, the employer will be entitled to maintain the provision that puts disabled people more generally at a particular disadvantage. Thus, an employer can continue to apply the indirectly discriminatory practice against the group, as long as the employer offsets the disadvantage accruing to the specific individual with a disability by way of a reasonable accommodation.[111] This aspect of the directive is problematic and is an attack on the scope and potential of the indirect discrimination principle in the disability context. As Whittle comments, its effect "is to remove any group benefits that may have otherwise accrued from a successful action in this regard".[112] This approach upsets the orthodox practice of requiring the policy to be changed for the benefit of all, though in the disability context the possibility of a multitude of employment practices giving rise to disparate impact on the grounds of disability may be one of the reasons for the original introduction of the

109 Discussed further in Stein and Waterstone, "Disability", n. 15 above, at pp. 917–20.

110 Art. 2(2)(b)(ii) of the Equality Directive.

111 It is worth noting that the procedural operation of the accommodation duty in some jurisdictions also weakens the section's provision for indirect discrimination. Under the UK's DDA 1995–2005, the duty becomes operable following a request from a covered individual. As Bamforth et al. point out, the requirement to refrain from indirect discrimination could require action even in the absence of a specific accommodation request: N Bamforth, M Malik and C O'Cinneide, *Discrimination Law: Theory and context* (London: Sweet & Maxwell 2008), p. 1006.

112 Whittle, "The Framework Directive", n. 7 above, at p. 310.

provision.[113] More broadly, as De Schutter points out, these "[a]d-hoc, individualised compensation measures risk becoming substitutes for wider scale modifications especially in the built environment or the organisation of work".[114] Under the EC structure, the individualised reasonable accommodation provision could allow the continuation of instances of indirect discrimination against groups of disabled people more generally. Not all member states have opted to implement the accommodation duty in substitution for the indirect discrimination principle.

Notwithstanding this point, the elements hostile to a disparate impact analysis of failure to accommodate claims have already been highlighted using the example of Ireland's equality legislation, which includes a stand-alone indirect disability discrimination provision. Further, given the absence of the orthodox elements of indirect discrimination under the Framework Directive, the suggestion by Stein and Waterstone of a disparate impact analysis of failure to accommodate claims across a wider range of sub-groups of disabled people also appears foreclosed.

Conclusion

The essence of the pandisability argument is that failure to accommodate claims can be analysed in ways beyond the current atomistic, individualistic approach. A more potent approach recognises that the failure to accommodate claims may not only affect the individual requester, but potentially illustrates the existence of employment practices and procedures which disproportionately exclude groups of disabled people from employment practices. In other words, the indirect discrimination analysis appears applicable to the reasonable accommodation duty because the failure to accommodate cases may have an impact beyond the specific circumstances of the individual with a disability and raise questions as to what extent such practices adversely impact groups of disabled people.

While such an interpretation of the reasonable accommodation duty may be raised in principle, real practical barriers remain. Interpreting the reasonable accommodation duty from an indirect discrimination perspective seems to have been closed off in two distinct ways. First, there are real doubts about the ability of the indirect discrimination norm to reconstruct institutions and practices in as inclusive a manner as possible – in essence, the pandisability argument simply overloads an already burdened discrimination norm with the impossible job of transforming more widely indirectly discriminatory practices which impact on groups of disabled people. This argument is predicated on the ability of indirect discrimination norms to promote transformative, pre-emptive practices on the part of employers by means of encouraging them to assess existing employment practices for possible disparate impact. However, this analysis is simply overstated. Except in the aftermath of the Dred Scott era and the period immediately after the *Griggs* decision in the US, employers, particularly, private sector employers, have rarely been involved in wholescale policy and procedure review from a disparate impact perspective. It does not demand any radical restructuring of employers' practices which remain justifiable. Second, despite the advance inherent in the notion of pandisability, there are barriers to surmount in respect of the constituent elements of the statutory definition of indirect discrimination

113 Tobler notes that "Art.2(2)(b)(ii) is intended to reinforce the duty to provide reasonable accommodation, in that the alleged discriminator can rebut a presumption of indirect discrimination by pointing to reasonable accommodation . . . Seen in this way, the consequence . . . is to help the victims of the alleged indirect discrimination to obtain reasonable accommodation, whilst giving the employer a certain degree of flexibility.": *Limits and Potential*, n. 8 above, at V.5.

114 O De Schutter, "Reasonable accommodations and positive obligations" in C Gooding and A Lawson (eds), *Disability Equality in Europe from Theory to Practice* (Oxford: Hart Publishing 2005), p. 35, at p.63.

as discussed above. Further, even if it were possible to surmount the possible statutory barriers raised above, one issue which has so far been overlooked is that a prima facie case of indirect discrimination always remains justifiable.[115] One of the major downfalls of the indirect discrimination principle has been its poor relation to the employer's interests in continuing with the practice or policy giving rise to disparate impact: in other words, it may often cede to powerful and established institutional rules that reflect the way things have always been done.[116] In light of these difficulties, Ellis has described indirect discrimination as "essentially . . . non-dynamic and [non]-redistributive . . . because it does little to dismantle hidden obstacles facing protected groups" or to "change customarily stereotyped roles".[117] Thus, both indirect discrimination and reasonable accommodation have their own separate limitations and their own spheres of operation. As such, the attempt to infuse new life into these doctrines may be little more than an elaborate discussion that repeats and reasserts many of the pre-existing limitations that attach to their interpretation and application in the struggle for disability equality.

115 See *Bilka-Kaufhaus GmBh* v *Weber Von Hartz* Case 170/84 [1986] ECR 160.

116 Though note the more robust enquiry into the employer's justification demanded by the proportionality principle.

117 Ellis, *EU Anti-Discrimination Law*, n. 92 above, at p. 115.

NILQ 60(3): 381–2

Book review

KATE BLOMFIELD*

Queen's University Belfast

Consociational Theory: McGarry and O'Leary and the Northern Ireland conflict, Rupert Taylor (ed.) (London: Routledge 2009), 400pp., £80

The murders of two soldiers and a police officer in Northern Ireland this past March 2009 drew international attention. Fears flared that the region's decade-old peace agreement, and the power-sharing government that institutionalised it, might be under threat. Importantly, Robinson and McGuinness, the (fraternal) twinned First Minister and Deputy First Minister, spoke in unison, denouncing the attacks and the citizenry of Northern Ireland appeared united against the prospect of returning to "troubled" times. This flashpoint, however, gives cause to think about the past, present and future of Northern Ireland's conflict and governance – was the power-sharing government instrumental to the end (and has it ended?) of sectarian violence? Is a government that houses such different views of Northern Ireland's future sustainable and can it serve the public well? What does and should the future hold for Northern Ireland's government in order to preserve peace and promote prosperity?

These questions are addressed in the recently published *Consociational Theory: McGarry and O'Leary and the Northern Ireland conflict*, edited by Rupert Taylor. Taylor assembles a symposium-style discussion on consociationalism, an applied political theory that advocates a specific form of power-sharing that promises democratically to resolve entrenched ethnic division and political conflict. The book focuses specifically on Northern Ireland's government, born of the 1998 Belfast/Good Friday Agreement (the "Agreement"), and is framed around an argument by John McGarry and Brendan O'Leary, leading proponents of consociationalism. McGarry and O'Leary outline their defence of the consociational Agreement and are met with 16 responses, ranging from wholly supportive to extremely critical, by notable authors from fields including political science, law, international relations and sociology. McGarry and O'Leary then have the final word, responding to each of the commentaries in turn.

The format of the book works well. It is apparent that a few of the authors have been in long intractable "discussions" over the benefits and failings of consociationalism and that

* Kate Blomfield is a student at Queen's University Belfast, taking her LLM in Law and Governance. She expresses thanks to the Commonwealth Scholarship Commission and the Law Foundation of British Columbia for their support of her studies.

their positions are entrenched and the debate growing weary. Much of the writing, however, imparts a sense of genuine interest and fresh perspective which keeps the volume relevant and energetic. Though dense, the tone of the essays, and particularly the replies and sur-replies, is often lively and the multitude of voices compelling.

McGarry and O'Leary's argument first categorises and dismisses their integrationist critics. Integrationists prefer fluid identities and mixing of segregated groups, whereas consociationalists see identities as more rigid and thus requiring accommodation. McGarry and O'Leary continue by looking at consociational theory and the Agreement, noting in particular the ways in which the Agreement differs from classic/theoretical consociational design. Finally, the primary areas of critique of consociationalism in general and the Agreement in particular are discussed – relating to stability, fairness and democracy. McGarry and O'Leary conclude that consociationalism was necessary to bring Northern Ireland out of conflict, and should remain integral to its governance until it "decays organically".

The responses to McGarry and O'Leary are at their best when they focus on evaluating in concrete terms the current operation of the consociational government (for example, Farry, chapter 7) and on emphasising the need (and opportunity) to begin to move beyond conflict management to conflict transformation (see O'Flynn, chapter 14, and Taylor, chapter 17). The discussions bring to light the ways in which the Northern Ireland settlement makes important use of transnational arrangements and implements a unique quasi-constitutional arrangement.

With respect to the need for consociationalism and the benefits that it has brought to date, McGarry and O'Leary emerge from the "'friendly' and 'not so friendly' fire" largely unscathed. None of the contributors successfully challenge their assertion that consociationalism, in the form in which it was introduced in Northern Ireland, was the best way to move the region out of violence and sectarianism to a period of peace and communication across sectarian lines. Though some argue (see, for example, Wilson, chapter 11, and Taylor, chapter 17) that life has been worse for the general public since and as a result of the Agreement, they are not convincing. The essays as a whole make clear the significant strides that Northern Ireland has made over the past 10 years.

This progress was evident in March 2009. The leading political parties, despite their vast differences, are clearly committed to working within the system and promoting stability. Lasting peace is obviously of fundamental importance, but, for Northern Ireland to prosper and to move towards becoming a more dynamic and egalitarian society, it is insufficient. A number of the book's chapters demonstrate the key weakness of McGarry and O'Leary's argument – its vision for the future. Consociationalism may have been useful to bring Northern Ireland to where it is now, but where should it go from here? And how will it get there?

Steiner (chapter 9) and Morison (chapter 15) are among those who discuss the need for greater citizen involvement and deliberation in Northern Ireland's governance. McGarry and O'Leary dismiss this suggestion abruptly, but should pay closer attention. Consociationalism operates thanks to, but then remains overly focused on, the political elite. The result is a government that is stable, but prone to gridlock. Attention needs to turn to the public at large, who are learning to live together. The leadership must be in more active dialogue with its citizens in order to chart a common path for the future.

Taylor's *Consociational Theory* is an important contribution to the extensive literature on consociationalism, as well as on Northern Ireland's conflict and settlement – it is highly recommended.