MANCHESTER 1824

The University of Manchester

# Student Law Review

VOLUME II | December 2013

# University of Manchester Student Law Review

The University of Manchester's
School of Law Student Journal

VOL. II

The *University of Manchester Student Law Review is* a student led peer-reviewed journal founded at the University of Manchester, School of Law. The journal exhibits the best academic work in law, criminology and ethics, on both the undergraduate and postgraduate levels, publishing it annually.

VOLUME II DECEMBER 2013

University of Manchester Student Law Review

© *University of Manchester Student Law Review*, 2013
All rights reserved

The *University of Manchester Student Law Review* and its editors are not responsible for the opinions expressed in this journal.

No part of this journal may be reproduced in any form without the expressed consent of The *University of Manchester Student Law Review* or the University of Manchester, School of Law.

All enquires should be addressed to:

*University of Manchester Student Law Review*
Williamson Building, School of Law
University of Manchester
Manchester
M13 9PL

review.manchester@gmail.com

Typeset in Baskerville Old Face
Printed in United Kingdom

# University of Manchester Student Law Review

**Editor-in-Chief**
Ben Adamson

**Associate Editors**
Christopher Markou
Craig Prescott
Jorge Nunez
Melissa Bone
Michal Kniec
Ugochukwu Orazulike

**Assistant Editors**
Ammar Lulla
Ashley Porter
Étienne Farmer Lacombe
Iman Nauman
Sarah Burke
Sonam Cheema

**Layout and Design**
Emile Abdul-Wahab

**Finance Officer**
Joseph Tomlinson

**Faculty Advisors**
Prof Geraint Howells
Prof Rodney Brazier
Prof Margot Brazier
Prof Toby Seddon
Dr Jackson Maogoto
Dinah Crystal OBE

This year's edition was generously funded by the
University of Manchester, School of Law.

The entire editorial team are sincerely grateful for the
School of Law's continuing support.

# Preface from the Head of the School of Law

On becoming Head of School I was happy to give the go ahead for this Review to be launched. It was an easy decision because of the enthusiasm of the students. It has also been one of the best decisions given the quality of what has been produced. It reflects the academic strength of the Law School and demonstrates that it spans the undergraduate and postgraduate communities and the disciplines of law, criminology and ethics.

Particular thanks this year go to Ben Adamson, the editor. It was a daunting task to follow our energetic founding editor, Michal Kniec, but I have had chance to see advanced copy and he has pulled off a tremendous success. Once again the Review has managed to showcase the quality of scholarly work in the School in a format which is appealing and accessible. There is a whole team that has worked with Ben from authors and editors to those involved in the management and production of the Review.

The Law School is proud to sponsor the Review and proud of our student community for producing work of such high quality.

Geraint Howells
Head of Law School

# Preface from the Editor-in-Chief

Familiar as I am with the prestige associated with student-led law reviews in the United States and their growing presence in the landscape of English legal education, I was very excited when I joined the University of Manchester Student Law Review in 2011 as an associate editor after its founding by Michal Kniec.

Upon assuming the mantle of Editor-in-Chief for the Review's second volume in 2012, I was eager to extend its already ambitious remit of publishing the finest examples of work from the School of Law's undergraduate and taught postgraduate students to also include work from postgraduate research students - reflecting the University of Manchester's membership of the Russell Group, and giving students a much sought after opportunity to publish original work stemming from their doctoral research.

In this, our second year, the response rate from students to the call for papers was astonishing and we received more high quality papers than we could possibly publish in two or even three volumes, let alone one. This presented the Editorial Board with a tremendously difficult task, narrowing down the papers to the selection presented in this volume. More importantly, the response represents a source of great pride in the School of Law and its students and one that should be shared by all - that so many students have produced work that is not only of high academic quality but also of high interest value and a genuine pleasure to read.

There are many people who deserve praise and thanks for their participation and support in the Review and it is regrettable that to list every one of them by name would be too time consuming an exercise. First and foremost I would like to thank Michal Kniec for all his hard work in founding the Review and for his continuing support throughout the second volume, allowing me to benefit from his considerable experience in student-led journals and from his passion and commitment to this Review. I would also like to thank Prof. Geraint Howells for his ongoing interest,

support and advice, and Dinah Crystal for her hard work in seeking and arranging funding. Without her priceless contribution, this book would not be in your hands today.

Of course without the editors, we would have no articles ready for publication and mere words cannot express my gratitude for the work of the Editorial Board – all of whom were hand picked from a high number of excellent applications. Offering up constructive critique on work that is already of a very high standard is no easy task and I am confident that every one of the editors this year has truly done justice to their respective authors and I hope that everyone involved is as proud of this Review as I am.

Volume two of the University of Manchester Student Law Review represents the culmination of a long and, at times, arduous journey for everyone involved from advisors to authors to editors and not least for myself but it also represents what has been a tremendously rewarding and enjoyable experience. The work published in this book truly sets the gold standard to which all University of Manchester law students should aspire and I hope that you derive as much inspiration and as much pleasure from reading it as we did from putting it together.

Ben Adamson
Editor-in-Chief

# Contents

# Style over Substance? A Comparative Analysis of the English and French Approaches to Fault in Establishing Tortious Liability

Danny Watson

## Abstract

*The English law and French law methods for establishing tortious liability are stylistically divergent. Casuistry prevails in English tort law, whereas the French law of delict (equivalent to English tort law) proceeds rigidly from a general principle of liability in three spartan articles of the Code Civil. The purpose of this article is to examine, with regard to the role of fault in tortious liability for harm caused by things under one's control, whether these different methods produce substantively different results. The English and French approaches to fault-based liability have evolved near-concurrently. Through examining these policy-based evolutions, it will be shown that both systems have sought to attenuate fault-based liability for harm caused by things under one's control, in favour of a stricter liability regime. This implies that the inflexible theoretical basis of liability in French law has not prevented the French system from incorporating policy just as fluidly as the English system. An exposition of the diminution of fault-based liability thus provides a salient example of how two legal systems with differing methodologies may coalesce in attaining practical solutions.*

## I. Introduction

The French law of liability for harm caused by things under one's control is founded on a single article[1] of the *Code Civil*. English law relies on the tort of negligence espoused in *Donoghue v Stevenson*[2], as well as specific nominate torts, to establish liability for things under one's control, demonstrating a very different approach. As will be seen, the broad provision of the French system has been

1 Article 1384.
2 [1932] UKHL 100.

narrowed to establish liability where the courts see fit, just as the casuistic approach of the English system *creates* torts where policy demands it. Despite their methodological differences, both systems produce remarkably similar outcomes in the majority of cases. Subjective (i.e. moral) fault traditionally played an imperative part in founding liability in both systems; however, its role diminished significantly throughout the Industrial Revolution, on the justification that certain activities should engender responsibility for harm irrespective of moral culpability. The similarities and differences between the modern English and French methodologies, and their practical results, will now be examined.

## II. The origin and development of strict(er) liability

The policy of *risque-profit* dictates that liability without fault should be imposed on someone who profits from a thing, because he who takes its benefit should bear its potential burden. Similarly, *risque-créé* advances the idea that someone who creates a risk should bear its potentially harmful consequences. Both of these arguments, as well as a general humanitarian concern[3], underlie the modern English and French approaches to liability for things under one's control.

Unlike French law, there is no general principle of strict liability in English law, the courts relying instead on the principles of negligence [4] to establish liability for unintentional harm, with occasional statutory[5] and casuistic intervention[6]. Unlike English law, French law does not refer to the concept of a duty in order to establish liability. However, a breach of a statutory duty for things under one's

3 Paula Giliker, 'Codifying tort law: lessons from the proposals for reform of the French Civil Code' (2008) ICLQ 582.

4 Duty, breach and causation.

5 Consumer Protection Act 1987.

6 *Rylands v Fletcher* (1868) LR 3 HL 330.

control will engender liability. In English law[7], breach of statutory duty was previously seen as a branch of negligence, but now engenders liability in its own right, distancing breach of statutory duty from fault-based liability.

Liability in English law for things under one's control generally depends on the existence of a duty of care, breach of this duty, and causation: the existence of a duty being based on foreseeable damage, relational proximity[8] and it being 'fair just and reasonable' to impose a duty to act as reasonable man[9] (similar to the *homme avisé* in French law) in a given situation. The level of care required will vary according to the probability of harm occurring in a given situation[10]. The seriousness of the potential harm[11] and the utility of a risky activity[12] will also affect the standard of care. Thus, policy factors weigh heavily on the establishment of liability in negligence, meaning that the English courts may impose strict liability on prevailing policy and/or social justice grounds as they see fit.

To move towards stricter liability the French *Cour de Cassation* established that art.1384.1 transferred the burden of proof for damage caused by things to the defendant[13]; though the 'presumption of liability' did mean that *lack* of fault obviated liability (a similar rebuttable presumption exists in English law under s.2 (1) Misrepresentation Act 1967, and under the doctrine *res ipsa loquitur*[14]). Then in 1914, it was established that the custodian of a thing could escape liability only by proving that the damage was due either to *force majeure*, contributory negligence or the act of a third party[15].

---

7 *Anns v Merton LBC* [1978] AC 728.
8 Between claimant and defendant.
9 *Blyth v Birmingham Waterworks* (1856) 11 Ex 781.
10 *Bolton v Stone* [1941] AC 850.
11 *Paris v Stepney BC* [1951] AC 367.
12 *Watt v Hertfordshire CC* (1954) 1 WLR 835.
13 Cp req 30 mars 1897.
14 *Ward v Tesco Stores* [1976] 1 WLR 810.
15 Req 19 Jan 1914 s 1914 I 128.

In this way, the courts 'have made extraordinary changes to fault-based and strict liability without any real modification of the wording of these articles [1382-4]'[16] in order to meet demands of social justice.

Originally, restrictions were imposed on the situations in which one could be held liable under art.1384.1. Firstly, that the thing which gave rise to liability must have been defective[17]. This was reversed in *Gare de Bordeaux*[18]. Other restrictions, such as that the thing in question must not have been guided by human hand, and that the thing must be dangerous, were rejected in *Jand'heur*[19]. Thus the current law states that wherever something is under one's control, one is presumed liable for damage it causes (subject to defences discussed below). All that is required of the 'thing' is that it contributed to the harm[20]. 'Liability... is founded solely on the question of custody (*garde*[21]) of the object and therefore any attempt to distinguish on the basis of the origin of the damage is irrelevant.'[22] It follows that liability may arise for both inert and moving things under one's control.

## III. Liability for inert and moving things

In *Colmar*[23], a French case, a woman fainted and injured herself against a scalding pipe in a public bath. It was held that it was 'reasonable' that the pipe was positioned where it was, and there was therefore no liability. However, in *Pialet,* a boy injured himself in a café after having tripped

16 Giliker (n 3) 568.

17 *Teffaire*; Cp Req 30 Mars 1897.

18 Civ 7 novembre 1922.

19 Ch réun 13 février 1930.

20 Civ 9 June 1933 DH 1393, 449.

21 Defined as 'usage, direction and control': *Connot c Franck* Ch reun 2 Dec 1941, S 1941 I 217.

22 R Redmond-Cooper, 'No fault liability on the French roads' (1995) JPIL 293.

23 John Bell, Sophie Boyron, Simon Whittaker, *Principles of French Law* (2nd edn, OUP 2008) 266.

over a chair that was lying in his way[24]. The inertia of the chair was held to be irrelevant: a *gardien*[25] is subject to a presumption of responsibility for damage caused by inert things, which is only rebutted if they prove that what happened could neither have been foreseen nor avoided. It is sufficient that the thing causes damage which would not otherwise have been produced, provided that its positioning is in some way abnormal[26].

Two further cases[27] reinforce the necessity of the inert thing having an 'abnormal' facet for the *gardien* to be held liable. Therefore the apparently strict liability for things under one's control is heavily qualified in the context of inert things by the notion of abnormality, which is a notion 'tainted with morality'[28], and is thus a fault-based approach. Moving things are dealt with via strict liability approach: in a case where a car being driven in a normal way killed a pedestrian in an accident[29], the driver was held liable. Thus with moving things we see a move away from fault-based liability, towards an approach where all that need be demonstrated is that the thing of which the defendant was *gardien* was in motion and impacted on the person or property damaged. In English law, liability for inert and moving things relies simply on the principles of negligence (subject to the special liability regimes discussed below), with liability depending upon the variable factors[30] weighed by the courts in establishing the existence of duty, breach, and causation, which in the context of inert and moving things will bring about very similar results[31] to the French system.

---

24 ibid 267.

25 Someone with 'the use, direction and control' of a thing: *Connot c Franck* Ch reun 2 Dec 1941, S 1941 I 217.

26 Civ (2) 19 March 1980, JCP 1980 IV 216, D 1980 IR 414.

27 Civ (2) 24 Feb 2005.

28 Jean Carbonnier, *Droit Civil* Vol 4, *Les Obligations* (22nd edn, 2000) 2369.

29 Civ 29 May 1964.

30 Such as it being 'fair, just and reasonable' to impose a duty.

31 Giliker (n 3) 565.

## IV. Defences

There exist two general defences to damage caused by things under one's control in French law: *force majeure*[32], where an unforeseeable/unavoidable event, external to the thing[33], causes harm (including the act of a third party); and contributory negligence. *Force majeure* was dealt with in *Trichard v Piccino*[34], where a driver was involved in an accident while having an epileptic fit. The *Cour de Cassation*[35] held that a mental disorder cannot be 'external' to a *gardien*; therefore the epileptic fit did not exonerate *Trichard*. In contrast, English law reduces liability by the extent to which a mental defect makes, for example, a driver lose control of a vehicle[36]. As in the French system, an unforeseeable intervening act such as an act of nature may break the chain of causation, thus either reducing or nullifying the defendant's liability[37]. Furthermore, a claimant's voluntary assumption of risk precludes liability under English law[38].

The defence of contributory negligence does not impinge on the extent of liability where for example a negligent pedestrian is hit by a car in French law[39], unless the fault of the victim was both 'inexcusable' and the exclusive cause of injury; however, the defence is entirely excluded in the case of those over seventy or under sixteen[40], unless they voluntarily sought the injury[41]. Thus stricter liability is encouraged in such cases, on the rationale of *risque-créé*. In

32 Or cas fortuit.

33 Req 22 Jan 1945, S 1945 1 57.

34 Civ 18 December 1964, D 1965, 191.

35 ibid.

36 *Roberts v Ramsbottom* [1980] 1 WLR 823.

37 *Carslogie Steamship Co Ltd v Royal Norwegian Government* [1952] AC 252.

38 *Darby v National Trust* [2001] PIQR 372.

39 21 July 1982, D 1982, 449, 487.

40 Unless the victim voluntarily sought the injury: *Desmares, 21 juillet 1982, D 1982, 449 & 487*.

41 ibid.

English law, contributory negligence reduces damages to the extent to which the claimant was contributorily negligent[42], provided that injury was caused by the risk to which the claimant exposed himself through his negligence[43]. The same degree of moral blameworthiness/negligence will engender different results according to the causative potency of the negligence[44]. Thus, this is a partial defence to, for example, the stricter liability under *Nettleship v Weston*[45]. Both legal systems provide remarkably similar outcomes with regard to defences - though English law takes a more nuanced, less maximalist approach in favour of a greater recognition of the role of fault - and both show that liability under both systems is not absolutely strict.

## V. Special liability regimes for motor vehicles, defective products and fire

French legislation enacted in 1985 delineated the manner in which compensation is given for 'victims of a traffic accident in which a motor vehicle is involved'[46]. The main purpose of the legislation was to improve the likelihood of compensation[47] through creating stricter liability[48]. The approach under the new law is simply to identify causation and harm, with the presence or absence of physical contact between the vehicle and the person harmed being decisive[49]. Under the 1985 law, a *gardien* cannot rely on *force majeure* against any victims[50] and *gardiens* involved in a crash caused by black ice, for example, will be liable; as will they be for an

42 s 1 Law Reform (Contributory Negligence) Act 1945.

43 *Barrett v MOD* [1995] 3 All ER 87.

44 *Froom v Butcher* [1976] QB 286.

45 [1971] 2 QB 691.

46 Loi 85-677 5 July 1985 arts 1-6.

47 Alain Bénabent, *Droit Civil : Les Obligations* (11th edn, 2007) 436.

48 Redmond-Cooper (n 10) 302.

49 François Terré, Philippe Simler and Yves Lequette, *Droit Civil: Les Obligations* (10th edn, 2009) 920.

50 Loi 1985, Article 2.

accident caused by an unforeseeable and unpreventable act of a third party. As regards the fault of other drivers, *any* fault on their part can reduce or extinguish a defendant's liability[51], again, on the *risque-créé* justification[52]. Thus, strict liability, while being difficult to circumvent, can be evaded under exceptional circumstances.

English traffic accident law relies on the principles of negligence and the Road Traffic Act 1988. It has long since distanced itself from a fault-based approach by imposing an objective standard of care. In *Nettleship v Weston*, a learner driver was held liable for damage caused in a road accident, despite the fact that it was arguably unfair to expect her to have attained the same level of competency as a qualified driver[53]. The English system therefore resolves road accidents very much in the same way as the French; i.e. by rendering defendants essentially strictly liable for damage they cause, with the objective standard of care in English law being practically equivalent to the more obviously strict liability of the French system.[54] Consent to risk is not a defence in either system[55].

With regard to defective products, the former, very strict liability imposed in France was somewhat weakened by the EC Product Liability Directive[56], which provides that a producer must compensate injury caused by a defect in a product he has put into circulation, 'when it does not provide the safety which a person is entitled to expect, taking all circumstances into account'[57]. This implies the possibility of strict liability; however, a number of defences are available and the range of recoverable harm is restricted. The

51 ibid Article 4.

52 Cooper (n 18) 300.

53 Nettleship (n 45).

54 Nils Jansen, 'Duties and rights in negligence: a comparative and historical perspective on the European law of extra contractual liability' (2004) OJLS 468.

55 s 149 Road Traffic Act 1988.

56 Dir. 1985/374/EEC.

57 Loi 98-389 19 May 1998.

development risks defence may avoid liability, where the state of technical knowledge available at the time did not enable the producer to discover a given defect[58]. English law also incorporates the development risks defence, and The Consumer Protection Act 1987 imposes an almost identical level of strict liability for defective products to the French system[59]. Liability for fire in English law comes under the Fire Prevention (Metropolis) Act 1774, which dictates that one is not liable for non-negligent fire, whereas French law takes a strict liability approach[60].

## VI. *Rylands v Fletcher* and hazardous substances

*Rylands v Fletcher*[61] established in English law a rule of strict liability for someone who brings on his land anything likely to do 'mischief' if it escapes. A defendant will be answerable for all the damage that is the consequence of its escape (provided that the storage of the thing is a 'non-natural user' of the land [62]), subject to reasonable foreseeability[63]. The hazardous quality of the thing is itself of no import: what matters is the likelihood of damage if it escapes. Thus the rule does not extend to damage caused by hazardous substances in general[64], contrary to the French system[65]. The Pearson Commission recommended that strict liability be extended by statute to all hazardous substances, though these proposals 'fell on stony ground'[66].

---

58 ibid.

59 Simon Taylor, 'The harmonisation of European product liability rules: French and English law' (1999) ICLQ 429.

60 *Gare de Bordeaux* Civ 16 Nov 1920.

61 (1868) LR 3 HL 330.

62 The rationale for this being that a 'non-natural user' brings with it increased risk, for which the defendant should be liable if damage is caused thereof.

63 *Cambridge Water Co v Eastern Counties Leather Plc* [1994] 1 All ER 53.

64 *Transco plc v Stockport MBC* [2004] 2 AC 1; *Read v Lyons & Co Ltd* [1945] KB 216.

65 Code Civil arts 1349-1351.

66 DK Allen, *Accident Compensation After Pearson* (Sweet & Maxwell 1979).

The defences to the rule in *Rylands v Fletcher*, such as statutory authority and 'act of God', demonstrate an approach that is not entirely strict, but which once again - notwithstanding the higher standard of care owed due to the increased risk inherent in the situation - retains a small element of fault.

*Risque-créé* is undoubtedly the underlying explanation for all of the above exceptions to putatively strict liability; it is thus that *force majeure*/unforeseeable acts and contributory negligence may still apply. Liability is still, in a sense, fault-based; one is liable if the 'fault' of creating risks through things under one's control engenders damage. But where damage occurs which cannot in any way be attributed to a thing under one's control, liability is generally avoided. This is most likely the explanation for the existence of the (restrictive and exceptional) defences in both English and French law to liability for things under one's control.

## VII. Conclusion

It is evident that the divergences in method as between the French and English legal systems' treatment of fault in establishing liability for things under one's control are generally more stylistic than substantive. The broad provision of art.1384.1[67] has been refined over the years to provide strict liability, and thus greater protection for victims, in areas where the courts have seen fit to do so. The English approach to such situations is to impose an objective standard of care that will render someone liable if it is not met, irrespective of any fault. Thus the two systems meet on a broadly similar level[68]. Both systems have moved away from fault-based liability[69], changing the threshold of liability according to prevailing social and economic normative judgments of responsibility. The graduation towards strict

67 *Code civil*.

68 Jansen (n 54) 465.

69 With the exception of non-negligent fire in English law.

liability varies with the situation. Where a victim is deemed to deserve greater protection, strict liability is imposed. Similarly, high-risk activities are legitimate only if they carry strict liability for damage resulting therefrom. Both systems achieve similar results. However, while both systems' overwhelming objective is victim compensation[70], the English approach takes a more nuanced view on the issue of fault, in contrast to the more maximalist French approach which provides near-absolute protection for victims. So long as defences such as *force majeure* and contributory negligence continue to exist, it would be erroneous to conclude that either system adopts strict liability as the fundamental principle underlying liability for things under one's control. A *morally-tainted* definition of fault is certainly almost obsolete; but it is liability for the 'fault' of creating certain risks through things under one's control which underlies both systems[71]; an idea which is still subject to fault-based defences.

70 Giliker (n 3) 582.
71 ibid 582.

## BIBLIOGRAPHY


### English Cases

*Anns v Merton LBC* [1978] AC 728
*Barrett v MOD* [1995] 3 All ER 87
*Blyth v Birmingham Waterworks* (1856) 11 Ex 781
*Bolton v Stone* [1941] AC 850
*Cambridge Water Co v Eastern Counties Leather Plc* [1994] 1 All ER 53
*Caparo Industries v Dickman* [1990] 2 AC 605
*Carslogie Steamship Co Ltd v Royal Norwegian Government* [1952] AC 252
*Darby v National Trust* [2001] PIQR 372
*Donoghue v Stevenson* [1932] UKHL 100
*Nettleship v Weston* [1971] 2 QB 691
*Paris v Stepney BC* [1951] AC 367
*Read v Lyons & Co Ltd* [1945] KB 216
*Roberts v Ramsbottom* [1980] 1 WLR 823
*Rylands v Fletcher* (1868) LR 3 HL 330
*Transco plc v Stockport MBC* [2004] 2 AC 1
*Ward v Tesco Stores* [1976] 1 WLR 810
*Watt v Hertfordshire CC* (1954) 1 WLR 835

### French Cases

Civ 16 juin 1896
Civ 16 novembre 1920
Civ 7 novembre 1922
Cp Req 30 mars 1897
Req 19 janvier 1914 s 1914 I 128
Ch réun 13 février 1930
Civ 9 juin 1933 DH 1393 449
Ch réun 2 mars 1941
Civ 19 février 1941
Req 22 janvier 1945, S 1945 1 57
Civ 29 mai 1964
Civ 18 décembre 1964, D 1965, 191
Civ (3) 4 février 1971 JCP 1971

Civ (2) 19 mars 1980, JCP 1980 IV 216 21 juillet 1982, D 1982, 449

## English Statutes

Consumer Protection Act 1987
Law Reform (Contributory Negligence) Act 1945
Road Traffic Act 1988

## French Statutes

Loi 85-677 5 July 1985
Loi 98-389 19 May 1998

## EU Directives

Dir 1985/374/EEC

## English Journals

Giliker P, 'Codifying tort law: lessons from the proposals for reform of the French Civil Code' (2008) *International & Comparative Law Quarterly* 561

Jansen N, 'Duties and rights in negligence: a comparative and historical perspective on the European law of extracontractual liability' (2004) *Oxford Journal of Legal Studies* 443

Redmond-Cooper R, 'No fault liability on the French roads' (1995) *Journal of Personal Injury Litigation*

Taylor S, 'The harmonisation of European product liability rules: French and English law' (1999) *International & Comparative Law Quarterly* 419

## English Textbooks

Allen DK, *Accident Compensation After Pearson* (Sweet & Maxwell 1979

Bell J, Boyron S, Whittaker S, *Principles of French Law* (2nd edn, OUP 2008)

Zweigert K and Kötz H, *An Introduction to Comparative Law* (3rd edn, OUP 1998)

## French Textbooks

Bénabent A, *Droit Civil: Les Obligations* (11th edn 2007)
Carbonnier J, *Droit Civil* Vol 4, *Les Obligations* (22nd edn, 2000)
Terré F, Simler P and Lequette Y, *Droit Civil: Les Obligations* (10th edn 2009)

# Sham Self-Employment Contracts: Taking a Liberty?

Fabian McNeilly

## Abstract

*Sham self-employment reduces employer liability, limits workers' rights and cuts tax revenues. This article considers the restrictions on contractual freedom in the context of employment contracts, focusing on sham self-employment. The parol evidence and signature rules are examined in detail, assessing that over time the strength of these contractual principles has been eroded by judicial decisions about the nature of employment contracts. I then turn to the public policy considerations of sham self-employment including the protection of vulnerable workers from economic duress and the collection of taxes. The need to balance contractual liberties with public protection leads me to a proposal for the introduction of a further stage in contractual relations. This would entail an explanation and summary of the terms by the dominant party in order to help address the unequal bargaining powers ubiquitous in employment relationships. Furthermore, I recommend that a contract of employment should be presumed into a work contract, so as to provide further safeguards for the public whilst not unduly restricting the sanctity of contractual freedom.*

## I. Introduction

'A contract of employment is...radically different from a contract to purchase a chocolate bar'[1] but there is much controversy over the extent to which contract orthodoxy should apply in the context of a contract of employment, particularly as regards 'sham' self-employment. This essay will examine what sham self-employment entails and how it interacts with established principles of contract law, such as the parol evidence rule and the signature rule.

---

1 Ewan McKendrick, *Contract Law* (8th edn Palgrave Macmillan, Kilbride 2009) 2.

The discussion will also consider the development of contractual principles in the context of employment and public policy. This will be done with a view to proposing some measures to maintain a delicate equilibrium between freedom of contract and the protection of employees.

## II. Sham Contracts Defined

In order to assess the optimum scope of contract law in employment status, it is first necessary to investigate what is meant by sham self-employment. Sham contracts were defined by Lord Diplock in *Snook v London & West Riding Investment Ltd*[2] as those contracts whose terms were 'different from the actual legal rights and obligations (if any) which the parties intend to create.'[3] That is to say, a sham contract exists where the written agreement does not accurately reflect the de facto agreement made between two (or more) parties. Therefore it can be seen that the concept of a sham contract is a well-established principle of contract law.

This principle was expressly extended in the sphere of sham self-employment where it was held that the terms must reflect the reality of the situation 'not only at the inception of the contract but...as time goes by.'[4] As a result, the present definition of a sham contract differs in employment law from traditional contract orthodoxy. As far as contract law is concerned, the parties must intend a contract to be a sham from the outset for it to be classified as such.[5] A further difference in definition is that in contract law, 'all the parties...[to a contract] must have a common intention'[6] to deceive the courts or a third party as to their true intentions, whereas in the context of employment it is

2 [1967] 2 QB 786.

3 (n 2) 803.

4 *Firthglow Ltd (Protectacoat) v Szilagyi* [2009] ICR 835, 846.

5 *Shalson v Russo* [2003] EWHC 1637.

6 (n 2) 802.

more often the case that the weaker party 'may be the victim of the deceit himself.'[7] This reflects the disparity between a commercial contract (where it is in the interests of both parties that the contract reflects the true agreement) and an employment contract (where there is greater inequality of bargaining power), which justifies the difference between the definitions.

## III. Parol Evidence Rule

Where an agreement between two parties has been committed to writing as a contract, it is a general presumption of contract law that the terms contained therein are the only terms to be considered in interpreting the contract[8] as it is 'intended by the parties to constitute the whole agreement.'[9] This literal approach is the traditional means by which the terms of a contract are interpreted. Its primary advantage is that the boundaries of investigation are clearly laid out and thus anything that lies outside of them can be dismissed without consideration. Whilst this leads to greater certainty in the contracting process, a number of concerns have been raised about it as an approach in a contemporary context. Firstly, employment contracts are typically drafted by employers with 'armies of lawyers,' [10] allowing them to exclude a number of terms without the knowledge of the other party. Secondly, the principle has been weakened by the growing number of exceptions to its application[11] and thirdly, 'the contents of documents may bear little relationship to the practice of a particular employment relationship.'[12] Therefore, it would seem inappropriate to

7 Anne Davies, 'Sensible Thinking About Sham Transactions' [2009] ILJ 318, 319.

8 *Jacobs v Batavia & General Plantations Trust Ltd* [1924] 1 Ch 287.

9 Alan Bogg, 'Sham Self-Employment in the Supreme Court' [2012] ILJ 41(3) 328, 334.

10 *Consistent Group v Kalwak* [2007] WL 1425696 [57].

11 McKendrick (n 1) 148.

12 Simon Deakin, 'Interpreting Employment Contracts: Judges, Employers, Workers' [2004] IJCLLIR 20 201, 217.

apply the rule strictly when 'a strong employer can easily impose a contract'[13] on unfavourable terms.

This judicial move towards a more purposive interpretive approach has largely been driven by the proliferation of sham terms in written documents, particularly in employment[14] where the 'relative bargaining power of the parties'[15] plays a pivotal role in contract negotiations. However, the Courts have at times been at pains to state that they do not 'seek to recast the contracts'[16] but rather discover 'what the actual legal obligations in the employment contract were.'[17] This includes bogus substitution clauses, often inserted to seek to avoid employment status being declared by the Courts.[18] In a number of cases there has been an express term in a contract to the effect that an individual need not perform the work themselves, in an attempt to circumvent the requirement of personal service which is necessary for a contract of employment.[19] Whilst cynical uses of these substitution clauses in an attempt to escape employer liability may be struck down,[20] the simple fact that a right to substitution was not exercised is not enough to render it a sham.[21]

This can be seen as the Courts attempting to tread a delicate line between established contract law principles on the one hand and the need to protect workers on the other. However, the relaxation of the parol evidence rule, with the increased adoption of the purposive approach of contract

13 Guy Davidov, 'Who is a Worker?' [2005] ILJ 34(1) 57, 67.

14 Gerard McMeel, 'The Principles and Policies of Contractual Construction', in A Burrows and E Peel (eds), *Contract Terms* (OUP, Oxford 2007) 27, 45.

15 *Autoclenz v Belcher* [2011] ICR 1157, 1168.

16 *Autoclenz v Belcher* [2009] EWCA Civ 1046 [106].

17 Spencer Keen, 'Things Are Seldom As They Seem' [2011] NLJ 161(7481) 1235, 1236.

18 *Glasgow City Council v MacFarlane* EAT/1277/99.

19 *Express & Echo Ltd v Tanton* [1999] ICR 693.

20 (n 18).

21 *Premier Groundworks Ltd v Jozsa* UKEAT/0494/08/DM.

interpretation is open to attack as 'words on a page provide order'[22] whereas permitting other evidence to be included creates uncertainty and undermines predictability.[23] As a result, whilst Courts should take account of 'the reality of the relationship,'[24] further erosion of the parol evidence principle would be detrimental to employers and workers alike. Nevertheless, it is clear that such a stance remains tenable only whilst the increasing formalism of employment relationships continues; for example, through 'the proliferation of standard form [employment] contracts.'[25]

## IV. Signature Rule

The signature rule holds that when a document has been signed, 'the party signing it is bound'[26] by the terms expressed therein. Whilst a contract of employment is often made orally, sham self-employment contracts are typically committed to writing in an attempt to shore up their legal weight. A major issue with a strict application of this rule is that often an individual will have little understanding of the document they are signing, particularly the implications of being classified as self-employed. This can be demonstrated by evidence from a survey that showed that the majority of homeworkers who were classed as self-employed did not realise the tax advantages of their position and as such were 'doubly disadvantaged.'[27] From this, it would seem unjust to permit this contract law principle to penetrate into the sphere of employment contracts beyond what could reasonably have been understood by the workers. However, that in itself

22 Peter Linzer, 'The Comfort of Certainty: Plain Meaning and the Parol Evidence Rule' [2002-2003] 71 FLR 799, 802.

23 Bogg (n 9).

24 Davidov (n 13) 64.

25 Hugh Collins, 'Legal Responses to the Standard Form Contract of Employment' [2007] ILJ 36(1) 2.

26 *L'Estrange v Graucob* [1934] 2 KB 394, 403.

27 Trade Union Congress Report, *Hard Work, Hidden Lives* <http://www.vulnerableworkers.org.uk/cove-report/full-report/index.htm> accessed 20 November 2012, 181.

could raise further problems, given the high proportion of sham self-employed migrant workers whose grasp of English is likely to be limited.[28] Furthermore, since the signature rule 'underpins the whole of commercial life,'[29] it 'must be presumed that the parties realised the importance of the written document.'[30]

Nonetheless, it would be remiss to dismiss the problem entirely. The principle issue lies in the lack of understanding of the terms of the contract. This allows unscrupulous companies to take advantage of vulnerable workers. As such, two potential paths present themselves for ensuring that the sanctity of contract is protected whilst also providing an adequate level of cover for workers. The first option would be to permit the use of the *non est factum* defence in a wider range of employment cases. This defence voids a signed contract where a signatory 'has not brought a consenting mind' [31] to the bargain due to lack of comprehension, particularly illiteracy. Whilst this defence has been given a narrow definition in cases such as *Saunders v Anglia Building Society,*[32] this has largely been due to the need to protect innocent third parties. However, sham employment contracts are typically bilateral agreements and therefore such objections are less persuasive. Nevertheless, this route would not be ideal as it would merely render a contract void and thus if this defence were to be expanded then there would need to be changes made so as to only void the written document and not the contract itself. Instead the Courts could use 'subsequent conduct evidence'[33] to infer the actual contract.

---

28 (n 27).

29 *Peekay Intermark Ltd & Anor v Australia and New Zealand Banking Group Ltd* [2006] 1 CLC 582, 598.

30 Bogg (n 9) 338.

31 McKendrick (n 1) 150.

32 [1971] AC 1004.

33 Bogg (n 9) 333.

Another, more appealing, option would be to maintain the contractual principles of offer and acceptance for employment contracts but with an additional requirement of explanation. For many workers there is a 'considerable disjuncture between what someone thinks their status is and what it actually is.'[34] Often, workers will be induced into signing up as self-employed but are 'never...told what that means.'[35] This could be remedied rather simply by requiring that, in the context of contracts relating to work, the dominant party be obliged to provide an explanation of the salient terms such that a reasonable layperson could understand them.

This could be, for example, a summary at the start of the written contract, clearly explaining the terms of the working relationship. Whilst there are many proponents of the view that 'parties should...be free to contract as they see fit,'[36] this proposal does not undermine the freedom to contract but merely enshrines a more equal understanding of the terms of the agreement, thereby ensuring protection for the vulnerable party. This summary would not vary the terms, though if there were disparity then the summary would take precedence and *contra preferentum* be applied. Moreover, this proposal is not such a divergence from contract law orthodoxy as might be assumed; it has long been established that 'the basic aim of contract law...is to deter people from behaving opportunistically toward their contracting parties.'[37] Therefore this proposal can be seen not as a revolutionary change but rather a natural progression of orthodox contractual principles in a contemporary context. There would remain an element of uncertainty as to

34 Deakin (n 12) 202.

35 (n 27) 152.

36 Samuel Engblom, 'Equal Treatment of Employees and Self-Employed Workers' [2001] 17 IJCLLIR 211, 225.

37 Richard Posner, *Economic Analysis of Law* (7th edn Aspen Publishers, New York 2007) 94.

which are the 'salient terms' but this is a concept that could be developed judicially.

## V. Public Policy

Despite the statement of Gibson LJ that 'public policy has nothing to say'[38] about sham self-employment relationships, there are 'huge implications'[39] in terms of the protection of vulnerable people, avoidance of legislative duties, and tax. These issues will be dealt with in turn, firstly in the context of unwanted sham contracts, forced on the worker by a party seeking to avoid employer liability, and secondly in terms of mutual shams, where both parties wish to avoid employee status being established.

### *A. Unwanted Shams*

The 'fundamental problem'[40] of resting employment rights on a contract is that it potentially excludes a large proportion of the workforce from statutory protection. Whilst the Employment Rights Act 1996 voids any clause in an employment contract which excludes or limits protection under the Act,[41] it does nothing to void a sham contract which is falsely set up as that of a self-employed contractor. As such, if contractual freedom were allowed to apply unfettered, many vulnerable people for whom Parliament had intended to provide protection would suffer a detriment. Indeed, they are likely to suffer not only financially and with less protection from unjust treatment, but also 'sites using bogus self-employment have a higher rate of injuries and fatalities.'[42] As a result there is a clear public interest here in ensuring that people are not only afforded the correct label in law but also in practice, so as to encourage further training

38 *Calder v Kitson, Vickers and Sons Ltd* [1988] ICR 232, 250.

39 Patricia Leighton and Michael Winn, 'Classifying Employment Relationships – More Sliding Doors or a Better Regulatory Framework?' [2011] ILJ 40(1) 5, 24.

40 Bob Hepple, 'Restructuring Employment Rights' [1986] ILJ 15(2) 69.

41 Employment Rights Act 1996, s 203(1)(a).

42 (n 27) 182.

and better safety provisions. In this context it can be seen that some limitations on orthodox contract principles are necessary to prevent them being a 'barrier to effective employment protection'[43].

Whilst there are those who argue that the law should not be about providing 'material justice,'[44] unconscionability is a well-established feature of contract law[45] and as such, it does not require a great divergence from established contractual principles (despite the claims of some academics who see freedom of contract and employment protection as almost mutually exclusive concepts[46]). Therefore, whilst it is necessary for the Courts to take a more interventionist approach to provide adequate cover for workers, this is not to say that orthodox contractual principles must be disregarded.

### *B. Mutual Shams*

Mutual shams are where, far from being taken advantage of by an unscrupulous employer, a would-be employee seeks to declare him or herself as self-employed. There are two primary motivations for this: tax and terms. An employer who will not be liable for a worker and is not limited by unfair dismissal laws for example, will be more willing to offer more favourable terms to a worker. Here it can be seen that there is a trade-off between rights and resources. By declaring themself to be self-employed, workers may be able to ensure a greater rate of return for their work; however, this short-term gain carries high risk potential, due to the potential costs of any injuries sustained and the lack of job security. Nonetheless, if a worker should wish to assume those risks, there is little persuasive reasoning

43 Douglas Brodie, 'Employees, Workers and the Self-Employed' [2005] ILJ 34(3) 253, 256.

44 Engblom (n 36) 217.

45 *Earl of Chesterfield v Janssen* (1751) 2 Ves Sen 125.

46 Davidov (n 13).

why they should not be free to do so, despite the 'constant erosion'[47] of the freedom to contract.

Where public policy is more concerned however is in the tax implications of sham self-employment. Whilst the Courts have generally been persuaded to intervene in a contract where everything but the label has suggested a contract of employment and there has been an element of unequal bargaining power, they have been less willing to do so when both parties have applied the label of self-employment willingly. In *Massey v Crown Life Insurance*,[48] where an employee had arranged a new contract so as to become self-employed for tax purposes, this label was held to be valid despite many indications that the claimant was working under a contract of employment. This judgment is troubling since it essentially permits the employer/employee relationship to be changed simply for the purpose of avoiding tax, a difficult position to justify. It would seem from a public policy point of view that whilst changing the employment status to redistribute the risks and rewards can be justified under freedom of contract, tax avoidance here seeks merely to harm a third party and therefore cannot be justified.

## VI. Conclusion

Whilst there are many who advocate a simpler relationship between contract and employment law,[49] such a position cannot be easily established. It is submitted that if an employment relationship were presumed into most contracts relating to work,[50] then the bulk of orthodox contractual principles could be applied with little restraint. This is because the Courts could consider the nature of the contract, rather than its written form. If the parties wished to

47 Simon Honeyball, 'Employment Law and the Primacy of Contract' [1989] 18(2) 97, 99.

48 [1978] 1 WLR 676.

49 (n 27).

50 International Labour Organisation Conference Report, 'The Employment Relationship' [2006] Report V(1) [27].

refute the *prima facie* employment relationship then the onus would be on them to do so. Therefore, contractual liberty could be preserved whilst also ensuring that any negative impact upon workers would have to be expressed in unambiguous, certain terms. However, due to the great inequality of bargaining power (and the economic duress to which this amounts) it is important that the effects of contract law principles are limited. The introduction of a requirement to explain the terms of an offer before it can be accepted would be a move towards ensuring the long-standing principles of contractual freedom are upheld whilst also maintaining an adequate level of protection for vulnerable members of the workforce.

## BIBLIOGRAPHY


### Books

Burrows A and E Peel (eds), *Contract Terms* (OUP 2007)

McKendrick E, *Contract Law* (8th Edn Palgrave Macmillan 2009)

Posner R, *Economic Analysis of Law* (7th Edn Aspen Publishers 2007)

### Case Law

*Autoclenz v Belcher* [2009] EWCA Civ 1046

*Autoclenz v Belcher* [2011] ICR 1157

*Calder v Kitson, Vickers and Sons Ltd* [1988] ICR 232

*Consistent Group v Kalwak* [2007] WL 1425696

*Earl of Chesterfield v Janssen* (1751) 2 Ves Sen 125

*Express & Echo Ltd v Tanton* [1999] ICR 693

*Firthglow Ltd (Protectacoat) v Szilagyi* [2009] ICR 835

*Glasgow City Council v MacFarlane* EAT/1277/99

*Jacobs v Batavia & General Plantations Trust Ltd* [1924] 1 Ch 287

*L'Estrange v Graucob* [1934] 2 KB 394

*Massey v Crown Life Insuranc* [1978] 1 WLR 676

*Peekay Intermark Ltd & Anor v Australia and New Zealand Banking Group Ltd* [2006] 1 CLC 582

*Premier Groundworks Ltd v Jozsa* UKEAT/0494/08/DM

*Saunders v Anglia Building Society* [1971] AC 1004

*Shalson v Russo* [2003] EWHC 1637

*Snook v London & West Riding Investment Ltd* [1967] 2 QB 786

### Journals

Bogg A, 'Sham Self-Employment in the Supreme Court' [2012] 41(3) ILJ 328

Brodie D, 'Employees, Workers and the Self-Employed' [2005] 34(3) ILJ 253

Collins H, 'Legal Responses to the Standard Form Contract of Employment' [2007] 36(1) ILJ 2

Davidov G, 'Who is a Worker?' [2005] 34(1) ILJ 57

Davies A, 'Sensible Thinking About Sham Transactions' [2009] ILJ 318
Deakin S, 'Interpreting Employment Contracts: Judges, Employers, Workers' [2004] 20 IJCLLIR 201
Engblom S, 'Equal Treatment of Employees and Self-Employed Workers' [2001] 17 IJCLLIR 211
Hepple B, 'Restructuring Employment Rights' [1986] 15(2) ILJ 69
Honeyball S, 'Employment Law and the Primacy of Contract' [1989] 18(2) ILJ 97
Keen S, 'Things Are Seldom As They Seem' [2011] 161(7481) NLJ 1235
Leighton P and M Winn, 'Classifying Employment Relationships – More Sliding Doors or a Better Regulatory Framework?' [2011] 40(1)ILJ 5
Linzer P, 'The Comfort of Certainty: Plain Meaning and the Parol Evidence Rule' [2002-2003] 71 FLR 799

## Reports

Trade Union Congress Report, *Hard Work, Hidden Lives* 'http://www.vulnerableworkers.org.uk/cove-report/full-report/index.htm' (Last accessed 20/11/12)
International Labour Organisation Conference Report, 'The Employment Relationship' [2006] Report V(1)

# The Bloody Code

Harriet Evans

## Abstract

*This Article focuses on the bloody code in England during the second half of the eighteenth century and assesses the extent to which its effectiveness depended upon the discretion of judges, jurors and prosecutors to mitigate and to nullify the law. Discretion was far reaching in 1750, playing a role in pre-trial proceedings, the trial itself and post trial procedure. This Article will also discuss other discretionary bodies such as the Justices of the Peace and the Grand Jury as well as the impact of transportation and the introduction of defence counsel.*

*Discretion was so prominent that this paper questions whether the bloody code could have been effective at all in its absence. It will be argued that whilst discretion is undoubtedly the most prominent factor in the effectiveness of the system other factors did contribute. Namely, its strength as an ideology, the position of society at the time and how a strict application of the statutes saw the law mitigate itself. It concludes that whilst there is evidence to suggest that the system could have been effective in the absence of discretion it is doubtful that it would have remained for so long had discretion not played such a large role.*

## I. Introduction

John Beattie has described the 18th century criminal justice system in England as one which 'was shot through with discretionary powers. Indeed it could hardly have worked had it not been.'[1] The aim of this essay is to discuss the Bloody Code in the second half of the 18th century and assess the extent to which its effectiveness depended upon the discretion of judges, jurors and prosecutors to mitigate and to nullify the law. This will lead to an examination of further areas of discretion within the system such as Justices of the Peace, the Grand Jury and Parliament. The final part of this essay will address whether the system could have been

1 John Beattie, *Crime and the Courts in England 1600-1800* (Princeton University Press 1986) 404.

effective in the absence of discretion, before concluding with a brief discussion on the appeals for reform that were simultaneously developing and gaining strength by discrediting the Bloody Code.

When dealing with historical issues there is, of course, the danger of projecting our own understanding backwards about the 'nature and workings of law itself.'[2] This essay will attempt to be sensitive to this fact and seek to interpret legal issues as contemporary agents understood the law to be.[3]

## II. The Bloody Code

Following the Glorious Revolution in 1688, the number of capital statutes in England and Wales grew from approximately 50 to 200 by 1820.[4] Almost all of these were for offences involving property. It was this vast number of offences, punishable by death, that led to the era being labelled as the Bloody Code by those who were arguing for reform.

Douglas Hay attributes the increase in capital statutes as a calculation by the ruling classes to manipulate the poor and maintain socio-political control: 'Again and again the voices of money and power declared the sacredness of property in terms hitherto reserved for human life.'[5] It has been estimated that approximately 35,000 people were condemned to death in England and Wales in 1770-1830 with about 7000 actually being killed.[6] The disparity

2 Stroud Francis Charles Milsom, *A Natural History of the Common Law* (Columbia University Press 2003).

3 Michael Lobban, 'Introduction' in Michael Lobban and Andrew Lewis (eds) *Law and History* (OUP 2003).

4 Leon Radzinowicz, *A History of English Criminal Law and its Administration from 1750*, Volume I (Stephens & Sons 1948) 4.

5 Douglas Hay, 'Property, Authority & the Criminal Law,' in Hay, Langbein et al, *Albion's Fatal Tree: Crime & Society in 18th Century England* (Peregrine Books 1975) 19. The actual effects of the increase in capital statutes may have been less significant than Hay suggests. See discussion of Emsley's ideas.

6 Vic Gatrell, *The Hanging Tree, Execution & The English People*, (OUP 1994) 7.

between the numbers condemned and the numbers executed may, to some extent, have been a result of the discretion exercised by judges, jurors and prosecutors.

## III. Discretion Exercised by Judges, Jurors and Prosecutors

Judges in the 18th century held extensive discretionary power and they exercised it to mitigate and to nullify the law. In the absence of defence counsel the judge would ensure fair play by questioning the witnesses and commenting on the evidence. Although there was no clearly developed law of evidence at this point, judges were beginning to examine evidence more closely. Judges would also recommend that the accused plead not guilty as this would at least let the jury hear certain mitigating factors, such as good character, which could mean the difference between life and death.

It was however in the post-trial procedures that the judge could exercise the most discretion. If the judge was unsure on a point of law he could reserve a case and suspend his verdict until he had gained the opinion of others. Should the point be found to be in favour of the accused then he would be pardoned at the next assizes. Following a capital conviction the judge would also reprieve some of the convicted or grant a conditional pardon. If the judge refused to grant a reprieve then the accused could petition to the King for mercy.[7]

---

7 Peter King, *Crime, Justice & Discretion in England* 1740-1820 (OUP 2000) 113. King has researched the frequency with which a particular factor was used in petitions for pardon in an attempt to determine what feature was the most successful in obtaining one. King criticised Hay for using only a small number of quotations from judge's reports to highlight the fact that he believed respectability was the most important factor in receiving a pardon. King therefore undertook a study of all factors mentioned between 1784-1787. The results in order of importance were: Good character, youth, circumstances of the crime, poverty of the culprit and finally respectability of the culprit. Beattie provides figures for the number of royal pardons for property offences in London in 1600-1800: 1139 people were sentenced, 703 of whom were pardoned. This is a pardon rate of 61.7%.

Juries were viewed by some as independent bodies who were the 'bulwark of English liberty.'[8] Langbein explains that 'whereas Hay has exaggerated the extent of prosecutorial discretion, he has underestimated the importance of jury discretion.'[9] Juries had great discretional ability. They could mitigate the law by finding the accused guilty of a lesser charge or by acquitting them. It is thought that jurors found not guilty verdicts or verdicts of 'not found' in favour of nearly 40% of the accused.[10] They could also find special verdicts where they 'decided the facts but left the court to determine whether those facts gave rise to criminal liability.'[11] Partial verdicts were an element of jury discretion that Blackstone called 'pious perjury.'[12] There are many of examples of verdicts where goods were valued at thirty-nine shillings,[13] in order to avoid the capital sanction given for thefts of goods over forty shillings. Beattie explained that the 'scale of undervaluation was frequently staggering.'[14] This is presumably because the jury 'thought about their verdicts at least to some extent in light of the punishment that would follow.'[15]

The prosecutors of crimes also played a large role in mitigating and nullifying the law. People from almost every class in the 18th century took others to court.[16] This, argues King, 'put a tremendous breadth of discretionary power in

---

8 John Hawles, *The Englishman's Right: A Dialogue Between a Barrister at Law and a Juryman* (1686). This text focuses in detail on jury independence.

9 John Langbein, 'Albion's Fatal Flaws,' *Past & Present* 98 (1983) 105.

10 King (n 7) 359.

11 John Langbein, *From Altercation to Adversary Trial* (OUP 2003) 329.

12 William Blackstone, *Commentaries on the Laws of England* (Cavendish Publishing 2001) 239.

13 Case of Alexander Duglass (1750) (theft from a specified place under 40s) Goods valued at 39s. As a result the punishment was transportation. Reference number: t17501017-9 <www.oldbaileyonline.org/static/crime.jsp> Accessed 11 November 2012.

14 Beattie (n 1) 424.

15 ibid 419.

16 King (n 7) 357.

the hands of the non-elite groups.' As a result a large majority of cases were settled within the community, before the issue could ever reach the courts.

Whilst developments to facilitate prosecution were improving – such as a growing rewards system, networks of thief-takers and help with prosecution expenses – there was also plenty of room for discretion in prosecutorial procedure. Prosecutors could decide 'what type of charge they wanted to bring without the interference of professional bureaucratically organized police forces.'[17] They could also weaken their evidence or downgrade their charge, which effectively gave them 'the equivalent to the jurors' partial verdict option.'[18] Some prosecutors chose not to turn up, 'contenting themselves with the fact the accused had often spent a considerable time in gaol awaiting trial.'[19]

## IV. Other Discretionary Bodies

It was not solely the judges, jurors and prosecutors who exercised discretion. There were plenty of participants in the criminal justice system who utilised this concept prior to the trial. Indeed, 'evidence suggests that the major participants in these earlier stages exercised wide and often almost untrammelled discretion.'[20]

The Justices of the Peace were a body in which discretion could be found at work. They tended to be people who were of some social standing and played a role in local governance. For minor offences, the Justice of the Peace could try the accused themselves but for more serious ones they would bind it over for trial by judge and jury. This was a procedure that was undoubtedly influenced by

17 ibid 356.

18 ibid 357.

19 ibid 356. According to Emsley in the Surrey assizes between 1771-1800 thirty-six men and women committed for trail in property cases were discharged due to a lack of prosecutor.

20 ibid 355.

discretion, not least because it often took place in an informal setting such as a local inn.

The Grand Jury's role was to consider if the indictment, drawn up by a clerk was a true bill and could be sent to trial or 'ignoramus,'[21] which meant that there was no case to be brought. They 'applied the law with discretion [when] deciding whether or not the prisoner should be sent to trial.'[22] Their decisions were influenced by many factors including the type of charge, who the accused was, the apparent state of crime and the need at that moment for examples to be made in order to deter potential offenders.[23]

We can question whether Parliament intended its statutes to be strictly enforced or whether it intended them to be applied with discretion. Radzinowicz argued that Parliament did intend for their legislation to be enforced and that judges 'increasingly vitiated that intention by extending pardons freely.'[24] Hay has strongly disagreed by saying that 'a conflict of such magnitude between Parliament and the judiciary would have disrupted 18th century politics and nothing of the sort happened.'[25] Paley, on the other hand, thought that 'the laws were never meant to be carried into indiscriminate execution...the legislature when it [established] its last and highest sanctions, [trusted] the benignity of the crown to relax their severity, as often as circumstances [appeared] to palliate the offence.'[26]

The introduction of transportation and varying lengths of imprisonment provided judges and juries with greater discretion when sentencing. In the 50 years after

---

21 John Baker, 'Criminal Courts & Procedure at Common Law 1550-1800' in James Cockburn (ed) *Crime in England 1550-1800* (Princeton University Press 1977) 18.

22 Beattie (n 1) 403.

23 ibid.

24 Hay (n 5) 23.

25 ibid.

26 William Paley, *Principles of moral and Political Philosophy,* (West and Richardson 1785).

1718, 30,000 were transported to North America. This provided an alternative which could leave death as an 'awful example to be visited upon by the worst few.'[27]

The introduction of defence counsel in the later part of the 18th century also allowed for further discretion within the system, though not perhaps in an obvious sense. The purpose of defence counsel was to simply cross-examine witnesses. What ensued however was a manipulation of 'cross-examination for the purpose of making a [defensive] argument.'[28] Langbein argues that the growing aversion to capital punishment was what contributed to contemporaries tolerating 'the truth impairing attributes of adversary procedure.'[29] Most trials in the 18th century were still lawyer-free however and, as a result, the accused still had to rely on the discretion of the judge and jury. The discretion of the lawyers in formulating arguments under the guise of cross-examination did, however, help a lucky few.

## V. An Effective System in the Absence of Discretion?

It is clear from the above argument that judges, jurors and prosecutors, along with other pre-trial bodies, acted with great discretion. Could the Bloody Code still be an effective system in the absence of such discretion? It can be argued that the system did not solely depend on these discretionary bodies to mitigate and to nullify the law:

> One of the key errors of many historians has been to take the 18th century Bloody Code at a face value based on modern perceptions of the

27 William Cornish, *Law & Society* (Sweet & Maxwell 1989) 694.

28 Langbein (n 11) 299. Langbein gives the example of the case of Gabriel Beaugrand and Louis Brunet OSB 1743 #256-7. The case involved murder by stabbing. The defence counsel, banned from arguing that the victim died accidentally from his own weapon, instead formed a question during cross examination: 'If a man had a sharp knife in his pocked might it not run into his body by accident?'

29 ibid 254. Langbein argues that this was a grave mistake and had the judges recognised the effect on the legal system they would never have allowed defence counsel in.

> law, thus they have assumed that the increase in capital statutes during the 18th century was a meaningful one...In reality the new capital legislation defined offences in a very narrow way and often made reference to a specific institution or piece of property only; as a consequence the number of prosecutions likely to follow the passing of a capital statute was tiny.[30]

Indeed, the law was, to a certain extent, mitigating itself by being so specific.

There were also limitations on who could exercise discretion; age and gender were the most obvious. Women were 'completely excluded from serving as judges, magistrates and jurors, and were much less likely to play a role as prosecutors, character witnesses or petitioners for pardon.'[31] In homicide cases too the presence of a coroner largely eliminated the room for discretion.[32] These facts do not detract from the wide discretion already exercised in the criminal justice system, but they do show that the role of discretion was marginally restricted.

It could also be argued that the Bloody Code was effective in the absence of discretion because of the position of society at the time. Mid-18th century England witnessed a dramatic transformation in society and economy due to the Industrial Revolution, and the population grew from 7 million in 1770 to nearly 14 million by 1830. The fear of disorder and social unrest was therefore running throughout this period – people only needed to look to France to see what could happen.[33] Perhaps the system was viewed as 'the price the English cheerfully paid for the liberty and prosperity.'[34]

---

30 Clive Emsley, *Crime & Society in England 1750-1900* (Longman 2005) 263.

31 King (n 7) 357.

32 Langbein (n 9) 103.

33 The French revolution, 1789.

34 Gatrell (n 6) 8.

It is possible to suggest that the Bloody Code was effective due to a combination of discretion and strict application of the law. If we believe Hay, the Bloody Code was effective due to its strength as an ideology. With no police force, physical force lay with the people and the ruling class used ideology to maintain authority. Hay talks of the characteristics of the criminal justice system as being majesty, justice and mercy. Majesty was seen in the twice-yearly visits of the High Court judges. Their visits 'had considerable psychic force...[and were an] elaborate manifestation of state power.'[35] Justice was seen to be shown 'when the ruling class acquitted men on technicalities...In short, its absurd formalism was part of its strength as ideology.'[36] Mercy was demonstrated through the act of pardoning. Considering this combination of technical acquittals and merciful pardons, it is unsurprising that it led to a system of justice which – when presented in this sense – resisted reform for so long.

## VI. Reform

It is interesting to note that both those arguing for reform and those loyal to the existing system 'shared a common description of the current process of law...They argued that judicial discretion was the operative principle of the system, where they differed so sharply was over the value to be attached to discretion.'[37] To its defenders 'the exercise of some degree of personal judgement in awarding punishment was necessary and desirable.' However, 'the Whig reformers challenged the uncertainty in operation of the law by this discretion and suggested that personal whim

35 Hay (n 5) 27.

36 ibid 33. There are numerous cases where men have been acquitted due to technical faults on the indictment or where the indictment does not match up to the evidence presented. As to the numbers of people who were acquitted in this way Beattie explains that these acquittals based on technicalities were often marked as 'not guilty' and as a result 'it is possible that a larger proportion than [we] realize of the not guilty verdicts were arrived at by these means.' 412.

37 Randall McGowen, 'The Image of Justice & Reform of the Criminal Law in early 19th century England, 32 *Buffalo L Rev,* 89 (1983) 110.

played too large a role in determining punishment.'[38] What had been created and sustained was effectively a 'lottery of justice.'[39]

Samuel Romilly was one of the first to propose to Parliament a mitigation of the law, in the early 19th century. He argued that 'the psychology of [reform] was sounder, [as] it represented the clear association of act and punishment.'[40] The subsequent Reform Act 1832 'gave new energy to independent abolitionist MPs,' to continue to push for reform of the criminal justice system.[41] The speed at which reform eventually ensued is indicative of the failings of the Bloody Code and the idea that the 'capital law had come to look randomly cruel and terminally silly.'[42]

## VII. Conclusion

It is clear that 'the great age of discretion was not necessarily the golden age of legitimation within the history of the English criminal law.'[43] The system contained a 'complex multidimensional set of decision-making processes,'[44] and at each stage it was clear that there was a 'continuous winnowing of the capital cohort, with the goal of leaving only the worst few for execution.'[45] The argument in this essay has been that the effectiveness of the Bloody Code relied on the discretion, not just of judges, jurors and prosecutors but also of other pre-trial bodies to mitigate and to nullify the law. Whilst there is evidence to suggest that the system could have been effective in the absence of discretion due to its strength as an ideology, the position of society at

38 ibid 91.
39 ibid 100.
40 ibid 118.
41 Gatrell (n 6) 22.
42 ibid.
43 King (n 7) 372.
44 ibid 356.
45 Langbein (n 11) 334.

the time and the fact that the law mitigated itself to some extent, it is doubtful that it would have remained for so long had discretion not played such a large role. It can be argued that the Bloody Code was not an effective system and this can be evidenced by the speed at which reform finally took hold. This essay has not addressed this point in detail. What this essay has attempted to show is that the Bloody Code, taken for what it actually was and not what it proposed to be – a system that contained a large amount of discretion and merciful pardoning instead of a strict application of the capital statutes – was an effective system in that it functioned in this way for so long. The system would undoubtedly have collapsed sooner had it not been for the discretion of judges, jurors and prosecutors, combined with other pre-trial bodies that acted with the knowledge that 'too much truth brought too much death.'[46]

46 ibid 334.

## BIBLIOGRAPHY

### Books

Baker J, 'Criminal Courts & Procedure at Common Law 1550-1800' in JS Cockburn (ed) *Crime in England 1550-1800* (Princeton University Press 1977)

Beattie J, *Crime and the Courts in England 1600-1800* (Princeton University Press 1986)

Blackstone W, *Commentaries on the Laws of England* (Cavendish Publishing 2001)

Cornish WR, *Law & Society* (Sweet & Maxwell 1989)

Emsley C, *Crime & Society in England 1750-1900* (3rd edn Longman 2005)

Gatrell V, *The Hanging Tree, Execution & The English People* (OUP 1994)

Hawles J, *The Englishman's Right: A Dialogue Between a Barrister at Law and a Juryman* (1686)

Hay D, 'Property, Authority & the Criminal Law,' in Hay, Langbein et al, *Albion's Fatal Tree: Crime & Society in 18th Century England* (Peregrine Books 1975)

King P, *Crime, Justice & Discretion in England* 1740-1820 (OUP 2000)

Langbein J, *From Altercation to Adversary Trial* (OUP 2003)

Lobban M, 'Introduction' in Lobban M and Lewis A (eds) *Law and History* (OUP 2003)

Milsom SFC, *A Natural History of the Common Law* (Columbia University Press 2003)

Paley W, *Principles of moral and Political Philosophy* (8th edn West and Richardson 1785)

Radzinowicz L, *A History of English Criminal Law and its Administration from 1750, Vol 1 The Movement for Reform* (Stephens and Sons 1948)

### Journal Articles

King P, 'Decision Makers and Decision Making in the English Criminal Law 1750-1800' (1984) The Historical Journal 27(1)

Langbein J, 'Albion's Fatal Flaws' (1983) Past & Present 98
McGowen R, 'The Image of Justice & Reform of the Criminal Law in early 19th century England' (1983) 32 Buffalo Law Review 89

## Cases

Gabriel Beaugrand and Louis Brunet OSB 1743 #256-7 (murder)
Alexander Duglass (theft under 40s) (1750) Reference: t17501017-9

## Websites

Robert Shoemaker and Tim Hitchcock 'Trial Procedures', 'Judges and Juries', 'Trial Verdicts' & 'Punishments' *Old Bailey Proceedings Online* 1674-1913: <www.oldbaileyonline.org/static/crime.jsp> Accessed 11th November 2012

# Proportionality - An Unattainable Ideal in the Criminal Justice System

Joel Goh

## Abstract

*In spite of its centrality in the criminal justice system, the principle of proportionality is poorly defined, and its role in judicial sentencing rests on shaky ground. The idea that criminal sanctions should be imposed only in proportion to those crimes to which they seek to respond is well recognised and ostensibly applied in most modern legal systems. However, by examining the role of proportionality in actual judicial sentencing, it is apparent that its application is highly problematic. Indeed, proportionality is founded on criminal punishment theories that are mired in complex and unresolved debates, offering little guidance to judges on the role of proportionality and the way it should be applied in sentencing. Moreover, proportionality competes with other sentencing goals, potentially giving rise to incompatibility when various objectives of criminal punishment are prescribed by sentencing guidelines. Further, it is crucial to note that crime and punishment are fundamentally disparate matters that do not in themselves possess any common benchmark for comparison vis-à-vis each other. Therefore, any proportionality that may exist between an offence and a sentence must necessarily be sought elsewhere - in social sentiments. Ultimately, the only meaningful and practical 'proportionality' that may exist in criminal punishment can only be the manifestation of society's opinions and moral assumptions. Consequently, the principle of proportionality cannot be an objective ideal to be attained but rather a goal to be continually strived for.*

## I. Introduction

Intrinsic in the concept of justice is the idea that where the criminal justice system imposes punishments, it should do so only in proportion to the crimes to which it seeks to respond. The principle of proportionality in criminal punishment is a fundamental aspect of most modern legal systems. However, it is ultimately an unattainable ideal and is, at best, a goal to be continually

strived for. This paper seeks to explain the role of proportionality in modern Western legal systems such as Canada and the United States, delve into the problems and difficulties posed by the principle of proportionality, and then explore how this principle may be understood in a more meaningful and practical way.

## II. The justice of criminal punishment

### *A. Scope of this paper*

The traditional theory of criminal punishment provides that the state imposes sanctions in response to the breaking of law.[1] This theory finds its basis in the ideas of the Social Contract through which free and rational individuals have collectively consented to relinquish certain rights in order to subsist peaceably in society.[2] Hence, the state alone, as the embodiment of the body politic, has the right to inflict punishment on its members, and to determine the sort of sanctions to be imposed for different crimes. Nevertheless, it has been argued that even Rousseau, one of the most influential writers on the Social Contract, was ambiguous with regards to the issue of how criminal punishment should be determined.[3] Subsequent thinkers have attempted to answer this question with the purposes of criminal justice such as those of deterrence, incapacitation, and rehabilitation.[4] While it is generally recognised that

1 See generally James Q Whitman, 'Between Self-Defense and Vengeance / Between Social Contract and Monopoly of Violence' (2004) 39 Tulsa L Rev 901, 913-917.

2 ibid.

3 For more on the debate over what Rousseau's ideas on punishment were, see Corey Brettschneider, 'Rights Within the Social Contract: Rousseau on Punishment' in Austin Sarat, Lawrence Douglas, Martha Merrill Umphrey (eds), *Law As Punishment / Law As Regulation* (Stanford University Press 2011).

4 Richard S Frase, 'Excessive Prison Sentences, Punishment Goals, and the Eighth Amendment: "Proportionality" Relative to What?' (2004) 89 Minn L Rev 571, 592. For a detailed account of how the purposes of criminal punishment have evolved, see Albert W Alschuler, 'The Changing Purposes of Criminal

criminal justice is concerned with such goals of punishment, the underlying issue of how these goals are achieved is shaped and restrained by the concept of proportionality. As such, proportionality is a fundamental principle in criminal sentencing, and the subject of much academic debate over its role in the concept of justice.[5]

## *B. The Proportionality Principle*

Much has been written about the concept of proportionality, which has been held to be the 'dominant principle driving the determination of sentences'. [6] Proportionality is considered to be so important in criminal sentencing because it 'accords with principles of fundamental justice and with the purpose of sentence - to maintain respect for the law and a safe society by imposing just sanctions'.[7] It 'embodies, or seems to embody, notions of justice. People have a sense that punishments scaled to the gravity of offences are fairer than punishments that are not. Departures from proportionality - though perhaps eventually justifiable - at least stand in need of defense'.[8]

In seeking to impose what is a just and fair punishment for criminal offences, the mantra 'the punishment must fit the crime' has been the prevailing sentiment, that the severity of the penalty should be

---

Punishment: A Retrospective on the past Century and Some Thoughts about the Next' (2003) 70 U Chicago L Rev 1.

5 See eg Franklin E Zimring, Gordon Hawkins and Sam Kamin, *Punishment and Democracy: Three Strikes and You're Out in California* (OUP 2001) 190; Margaret Jane Radin, 'The Jurisprudence of Death: Evolving Standards for the Cruel and Unusual Punishments Clause' (1978) 126 U Pa L Rev 989, 1043-1056; Richard G Singer, 'Sending Men to Prison: Constitutional Aspects of the Burden of Proof and the Doctrine of the Least Drastic Alternative as Applied to Sentencing Determinations' (1972) 58 Cornell L Rev 51, 56 and 72-89, cited in Frase (n 4) 596.

6 *R v Arcand* [2010] AJ No 1383 (Alta CA) 55 [*R v Arcand*].

7 ibid 52.

8 Andrew von Hirsch, 'Proportionality in the Philosophy of Punishment' (1992) 16 Crime and Justice 55, 56.

proportionate to the gravity of the offence committed.[9] The proportionality principle has long been an intrinsic aspect of criminal justice and is considered at sentencing in different ways. For instance, in jurisdictions like the United States and Canada, concepts such as 'gross disproportionality' have been developed from the prohibition of excessive 'cruel and unusual punishments' as enshrined in Section 12 of the Canadian Charter of Rights and Freedoms and in the Eighth Amendment to the United States Constitution. Section 12 of the Canadian Charter prescribes that '[e]veryone has the right not to be subjected to any cruel and unusual treatment or punishment',[10] and the relevant section of the Eighth Amendment to the United States Constitution provides that '[e]xcessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted'.[11] Further, proportionality at judicial sentencing has been specifically identified in judicial guidelines such as the Canadian Criminal Code. For example, Section 718.1 of the Code provides that '[a] sentence must be proportionate to the gravity of the offence and the degree of responsibility of the offender'.[12]

Nevertheless, despite this strong recognition of the importance of proportionality in criminal justice, 'the law with respect to proportionality in sentencing is confused, and what the law can be discerned rests on weak foundations'.[13] As a result, the application of the proportionality principle in judicial cases has been criticised. For instance, the decisions

---

9 Andrew von Hirsch, 'Proportionality in the Philosophy of Punishment: From "Why Punish?" to "How Much?"' (1990) 1 Criminal Law Forum 259, 262; Ronen Perry, 'The Role of Retributive Justice in the Common Law of Torts: A Descriptive Theory' (2006) 73 Tenn L Rev 177.

10 Canadian Charter of Rights and Freedoms, Part I of the Constitution Act, 1982 being Schedule B to the Canada Act 1982 (UK), 1982, c 11, s 12 [Charter].

11 US Const amend VIII.

12 Canadian *Criminal Code*, RS C 1985, c C-46, s 718 1.

13 Steven Grossman, 'Proportionality in Non-Capital Sentencing: The Supreme Court's Tortured Approach to Cruel and Unusual Punishment' (1994) 84 Ken L Rev 107, 107-108.

of the United States Supreme Court on gross disproportionality based on Eighth Amendment infringements have been considered to be significantly flawed,[14] in particular because of the lack of 'a constitutional standard consistent with accepted philosophical justifications of punishment and embodying principles'.[15] Indeed, there are many underlying problems inherent in the attempt to apply the proportionality principle to sentencing, posing several difficulties to the criminal justice system.

## III. Difficulties posed by the proportionality principle

There are several problems arising from the concept of proportionality, and four particular issues shall be considered in this section: (a) The vague definitions and theories of proportionality in the law, (b) the irreconcilability of other sentencing goals with the proportionality principle, (c) the inherently different natures of crime and punishment, and (d) the underlying character of the proportionality principle as a manifestation of mere opinions and sentiments.

### *A. Conflicting theories and poor definition in the law*

Despite the obvious importance of the proportionality principle in criminal sentencing, the concept of proportionality itself is poorly defined in the law and the theories concerning it are the subject of much unresolved debate. This vague definition is a glaring gap in the criminal justice system. For instance, although the Canadian Criminal Code provides that sentences 'must be proportionate' to the severity of the crime and the culpability of the criminal,[16] it does not proceed to elaborate on what 'proportionate' might mean with respect to gravity of offence and degree of responsibility, or how such a 'proportionate' sentence may be

14 For a thorough discussion on the 'series of flawed opinions from the Supreme Court' in 'all of the modern holdings of the Court in this area', see ibid.

15 Grossman (n 13) 108.

16 Canadian *Criminal Code*, RS C 1985, c C-46, s 718 1.

determined. Similarly, although the United States Supreme Court clearly professes to apply the proportionality principle in criminal sentencing, it has been observed that through its judicial decisions, it 'has never made clear what it means by proportionality in the context of prison sentences.'[17]

It is possible that proportionality is assumed to be so self-evident a principle that it does not necessitate elaborate expositions and definitions of its precise meaning and operation. However, to hold such a view would be to overlook the large amount of ongoing debate over the different theories of proportionality. It is more likely, then, that the reason for this lack of clarity concerning the concept of proportionality is that there is a lack of consensus over what the ideal form of proportionality is, what the purposes of punishment (which proportionality is meant to be a means to fulfil) are, and how to derive both of these. Consequently, the ideal form of proportionality and its role in punishment have been the subject of much academic discussion, and several theories have emerged, including that of retributive proportionality, utilitarian proportionality, and the concerns of ordinal and cardinal proportionality.

Retributive proportionality concerns the history of the offender and considers proportionality as a means to the punishment goal of retribution by measuring a sentence according to the offender's blameworthiness.[18] As expressed by Immanuel Kant, one of its supporters,

> Juridical punishment can never be administered merely as a means for promoting another good either with regard to the criminal himself or to civil society, but must in all cases be imposed only because the individual on whom it is inflicted has committed a crime. For one man ought never to be dealt with merely as a means subservient to the purpose of another...Against such treatment his inborn personality has a right

17 Frase (n 4) 588.
18 See eg ibid 590-592.

> to protect him, even though he may be condemned to lose his civil personality. He must first be found guilty and punishable before there can be any thought of drawing from his punishment any benefit for himself or his fellow-citizens. The penal law is a categorical imperative; and woe to him who creeps through the serpent-windings of utilitarianism to discover some advantage that may discharge him from the justice of punishment, or even from the due measure of it, according to the Pharisaic maxim: 'It is better that one man should die than the whole people should perish.' For if justice and righteousness perish, human life would no longer have any value in the world.[19]

Retributive proportionality is manifested in two forms. Firstly, 'defining retributivism' determines the punishment as precisely as possible to the severity of the offence, leaving little room for other punishment purposes. The purpose of retribution thus informs the sentencing judge to formulate a punishment which is proportionate to this intended end result. Secondly, 'limiting retributivism' allows other sentencing goals to be considered, merely placing retributive outer limits on the range of possible sentences. This way, the sentencing judge formulates a punishment in order to meet the various goals of punishment, such as social deterrence and denunciation, but then reins in the sentence to conform to the principle of proportionality.

In contrast, utilitarian proportionality is prospective rather than retrospective, with proportionality measured against sentencing goals which concern the future rather than the past, such as deterrence, rehabilitation, and cost to society.[20] There are two aspects of utilitarian proportionality.[21] The first is in terms of 'ends

---

19 Pincoffs 1966 at 2-3, cited in von Hirsch (n 8) at 60.

20 See eg Frase (n 4) 592-596. See also Michael Cavadino and James Dignan, *The Penal System: An Introduction*, (2nd edn, Sage 1997) 39 (on the debate between retributive and utilitarian proportionality).

21 Frase (n 4) 592-597.

proportionality', which concerns whether the costs of pursuing the goals of the criminal sentence outweigh the benefits to be derived from it (to both society and the individual offender). The other aspect of utilitarian proportionality is 'means proportionality', which assesses whether alternative less costly sanctions are available for achieving the same intended benefit.

As the 18th Century philosopher Cesare Beccaria argued, sanctions should be proportional to the gravity of the offences, as measured by the harm done to society.[22] Similarly, Jeremy Bentham asserted that punishments should have a utilitarian function and so must be proportional to the gravity of the crime in order to maximise efficiency in public resource allocation because 'the greater an offence is, the greater reason there is to hazard a severe punishment for the chance of preventing it'.[23] He further explained that 'punishment itself is an evil and should be used as sparingly as possible' and that a form of punishment should not be used if 'the same end may be obtained by means more mild'.[24] Since punishment harms and dissatisfies those upon whom it is inflicted, it can only be justified insofar as it produces a net amount of other benefits or satisfaction exceeding the harm. As the concept of utility is wholly consequentialist, the moral concept of 'just deserts' cannot be the reason for punishment. Instead, punishment is only justified inasmuch as its beneficial effects, for instance in deterrence, exceed the harm it produces.

H.L.A. Hart sought to reconcile the two competing ideas of retributive and utilitarian proportionality, suggesting that while 'we can agree that the reason for having a penal system at all is the general betterment of society...we can at the same time maintain with consistency that punishment should only be handed out to those who deserve it, and only

22 Cesare Beccaria, *On Crimes and Punishments* 62-66, as cited in ibid 593.
23 Jeremy Bentham, *The Theory of Legislation* 326, as cited in ibid 593.
24 ibid.

to the extent of their guilt.'[25] This synthesis of utilitarianism and retributivism has had significant and current influence on many criminal justice systems.[26]

These debates [27] are useful in answering the questions of *how* proportionality should be applied to criminal punishment and *why* it should be applied in a particular way, *viz.* the fulfilling of the purposes of punishment. However, there is no easy resolution to these debates, and much of the differences between the competing theories stem from a deeper divergence in opinions concerning the criminal justice system. They 'differ from one another largely in the emphasis they give the principle of proportionality - that is, the requirement that sanctions be proportionate in their severity to the seriousness of offenses'. [28] More crucially, however, these debates are focused on the *application* of proportionality, and do not answer the more fundamental questions regarding the basis for the concept of proportionality and what it really means, *viz.* proportional as to *what*. It seems as if proportionality is assumed to be an intrinsic good in and of itself, without a need for deeper analysis of issues such as what it really is, how it is derived, and its appropriateness as a sentencing principle. Consequently, these questions concerning the essence of what proportionality, at its root, is remain ambiguous and unanswered, and this is the first difficulty concerning the proportionality principle.

---

25 Morris J Fish, 'An Eye for an Eye: Proportionality as a Moral Principle of Punishment' (2008) 28 OJLS 57, 66.

26 ibid 67.

27 For a more detailed exposition on what Jeremy Bentham, Immanuel Kant, and HLA Hart wrote, respectively, on penal utilitarianism, retributive sanctions, and a reconciliation of both, see von Hirsch (n 8) 57-63.

28 ibid 55-56.

### *B. Inconsistencies between proportionality and the objectives of punishment*

Secondly, there is difficulty in reconciling the various goals of punishment with the proportionality principle. Logically, where two different forces direct a criminal sanction, a judge deciding the sentence needs to choose between one and the other in determining the appropriate sentence. Even if we accept the premise that proportionality is an inherent good in the sentencing process, the disparate goals of punishment necessarily lead to different penalties from that produced through applying the proportionality principle. Several policy objectives of criminal punishment seem to demand sentences decidedly *disproportionate* to merely what the severity of the crime and the culpability of the offender would attract. Such a statement is made with the acceptance of the premise that a 'proportionate' sentence can be objectively determined from the severity of a crime and the culpability of an offender. As will be explained later in this paper, such a premise is flawed but is what drives sentencing regimes in the criminal justice system today.

For instance, in seeking to expressly 'denounce' a crime, a sentence will often need to exceed what is simply 'proportional' to the offence because there would be no discernible denouncement if a 'denouncing sentence' were exactly the same as a 'proportionate sentence'. Similarly, the objective of 'separating offenders from society, where necessary',[29] implies that a criminal should be incarcerated for a period likely longer than what is merely proportionate to his offence. Such a dilemma is illustrated in Title 18 of the United States Code which provides[30] that the purposes of a sentence should be 'to reflect the seriousness of the offense, to promote respect for the law, and to provide just punishment for the offense',[31] as well as 'to protect the public

29 Canadian *Criminal Code*, RS C 1985, c C-46, s 718(c).

30 18 USC § 3553.

31 18 USC § 3553 (a)(2)(A).

from further crimes of the defendant'.[32] A sentencing judge, then, taking into consideration the full set of sentencing goals, is faced with the question of how to reconcile all the different sanctions that each of these goals would necessitate. It is almost certain that at least in some cases, the punishment prescribed by one sentencing goal will conflict with that of another, compelling the judge to choose one at the expense of the other. This inadvertently compromises the requirements laid down by sentencing guidelines such as Title 18 of the United States Code. Even if there is assumed to be a range of 'proportionate' sentences for each crime within which judges may exercise discretion and take into consideration the other goals of sentencing (i.e. through 'limiting retributivism'), there will inevitably be cases where proportionality and policy objectives contradict in the scale of the punishment to be prescribed. Although there admittedly will be much overlap between what is a 'proportionate' sentence and what is a 'deterring' or 'incapacitating' sentence, there will also be instances where they differ. Where proportionality prescribes one form of punishment while other policy objectives requires a different and irreconcilable one, the sentencing judge will have to choose one or the other, and cannot fulfil both.

Compounding this problem, there remains considerable disagreement over the different justifications for punishment and, by extension, between the various sentencing goals. For example, John Kleinig describes the contention concerning whether criminal punishment should be utilitarian or morally informed, a manifestation of the larger debate underlying utilitarian and retributivist proportionality.[33] Punishment is undeniably for the public good, but what is disputed is whether this public good consists in punishing for certain utilitarian goals or for moral concerns of what is 'right', either of which leads to a

32 18 USC § 3553 (a)(2)(C).

33 John Kleinig, 'R S Peters on Punishment' (1972) 20 Brit J of Edu Studies 259, 265-266.

consideration of 'proportionality' differing from the other. Similarly, the competing ideas of rehabilitative and retributive punishment disagree with regards to how punishment should consider the offender: either the evaluation of blameworthiness is a pointless exercise and so punishment should only be meted out for the purpose of rehabilitating the individual, or the punishment should seek to inflict upon the offender a sentence which manifestly reflects the gravity of his or her personal culpability.[34]

If the evaluation of blameworthiness is recognised as a means of retributive punishment, then proportionality will rightly find its place in assessing the wrongfulness of conduct. It has been argued that the concept of proportionality 'only has meaning in relation to retributive sentencing goals and that a proportionality requirement makes no sense if the Court is not going to require states to adopt a retributive theory'.[35] If, however, as Jeremy Bentham argues, this evaluation of blameworthiness is pointless, and that punishment should instead seek to rehabilitate the offender to change his or her ways and to deter potential offenders in society, then the proportionality principle takes on a fundamentally different role, *viz.* one of assessing the *utility* of the penalty. It is these unsettled disputes over the underlying dynamics of criminal sentencing which lead to fundamental uncertainty over how to sentence. Again, either idea will result in a disparate conception of the 'proportionality' to be applied in formulating the criminal sanction.

Thus, the principle of proportionality is founded on vague definitions and unsettled debates over the purposes of punishment that determine the relevance of the principle in the first place. Consequently, if even the very basis of criminal sentencing - *why* sentence, and *how* to sentence - are at the centre of such current and open debate, it is difficult

34 von Hirsch (n 8) 64.
35 Frase (n 4) 588.

for sentencing judges to reconcile all these theories in order to satisfy each of them. Indeed, '[t]he practice of punishment...rests on a plurality of values, not on some one value to the exclusion of all others'.[36] As such, a judge under a legal system which purports to dispense punishment in accordance with a range of sentencing goals such as deterrence and denunciation (for instance, in the Canadian Criminal Code) will, at certain points of irreconcilability, have to decide to either mete out a sentence based on proportionality contrary to other policy goals (i.e. 'defining retributivism') or choose other goals contrary to proportionality.

Additionally, legislative involvement in sentencing, such as through the prescription of mandatory minimum prison terms, elevates these problems by reducing the scope of judicial discretion in applying the principle of proportionality in criminal sentencing. For example, some jurisdictions require a mandatory minimum sentence for certain crimes, which the legislature presumably deems to be 'proportionate' to the nature of those crimes but which deprives the judiciary of a wide discretion in determining each individual case on their facts. Because of this, the Supreme Court of Canada in ***R** v Smith*[37] held that the mandatory minimum of a seven-year prison sentence for the importation of drugs was a violation of the right against cruel and unusual punishment as enshrined in Section 12 of the Canadian Charter of Rights and Freedoms[38].

Therefore, it is clear that proportionality is, in certain cases, necessarily a defining principle of the judicial sentencing process and may thus be irreconcilable with other sentencing goals. As such, its application in criminal punishment conflicts with the requirement that judges take

---

36 Hugo Adam Bedau and Erin Kelly, 'Theory of Punishment' (Stanford Encyclopedia of Philosophy, 19 February 2010) <http://plato.stanford.edu/entries/punishment/#2> accessed 12 April 2012.

37 *R v Smith (Edward Dewey)* [1987] 1 SCR 1045.

38 Charter (n 10) s 12.

into account other sentencing objectives and legislative prescriptions on judicial sentencing.

### *C. Meeting crime with punishment - comparing wholly different matters*

Also, crime and punishment are inherently separate concepts of entirely different natures, making it impossible to simply compare the two on a scale of 'proportionality' against each other on their own. Thus, they require a preceding separate *a priori* judgement on their values from which ideas of 'proportionality' can then be scaled.

The definition of crime has been the subject of much intense debate,[39] and it is not the ambition of this paper to produce a definitive resolution to it. What it seeks to highlight, however, is the fact that the nature of crime is fundamentally different from the nature of punishment. Descriptively, crime is 'the point of conflict between the individual and society' [40] because it 'is fundamentally a violation of conduct norms which contain sanctions, no matter whether found in the criminal law of a modern state or merely in the working rules of special social groups.'[41] However, the nature of crime is immensely complicated, and involves several approaches in understanding it. One of these is the economic approach which considers most crimes in general to be the generation of losses which can almost never be repaid perfectly. [42] Although an admittedly simplistic portrayal of crime which may not fit in absolutely every case, the economic approach fits in the general case, and is but one of several approaches to understanding the nature of crime. For instance, theft is the generation of a loss

---

39 See eg William M Ivins, 'What is Crime?' (1911) 1 Reform of the Criminal Law and Procedure 531.

40 ibid 531.

41 Walter C Reckless, *Criminal Behavior* (McGraw-Hill 1940) 10.

42 For a detailed exposition of this economic characterization of crime, see Gary S Becker, 'Crime and Punishment: An Economic Approach' in Gary S Becker and William M Landes (eds), *Essays in the Economics of Crime and Punishment* (UMI 1974).

of personal property; defamation is the loss of good reputation; rape is the loss of dignity (amongst other things); and homicide is the generation of a loss of life.

Even within this simplistic depiction of crime as the creation of losses, it would be impossible to repay the loss generated by most kinds of crimes, such as the loss of dignity, loss of a bodily function, or loss of life. Moreover, even for crimes where it may be possible for an offender to repay the loss (for example, in cases of theft or fraud), save for minor offences where community service or compensation orders may be meted out, criminal punishment typically does not seek the restitution of a victim, requiring a separate civil suit for that purpose to be filed instead. While restitutive justice may sometimes be considered to be a goal of the criminal sentence, the purposes of punishment are diverse and generally include other objects such as deterrence, retribution, incapacitation, and denunciation which may take precedence over restitution. Moreover, even where restitution is considered, it is often not the sole aspect of the criminal sentence, but merely a part of it, usually meted out with a supplemental punishment in addition to the compensation.[43]

Moreover, as noted earlier, this economic approach is but one portrayal of crime, and there are several other methods to understanding the intricate nature of crime which are beyond the scope of this paper. These include considerations of the moral wrongfulness of crime, the social stigma of criminal offences, and the philosophy of wrongdoing, all of which contribute to a fuller understanding of the complex nature of crime. From the complexity of the nature of crime then, three conclusions may be drawn. Firstly, that it is difficult to characterise crimes and reduce them to something measurable. Secondly, it is even more difficult to find a common benchmark (or benchmarks) to

---

43 Lorenn Walker and Leslie Hayashi, 'Pono Kaulike: Reducing Violence with Restorative Justice and Solution-Focused Approaches' (2009) 73:1 Fed Probation J 23.

holistically measure crimes against each other, whether in terms of severity of losses, social stigma, moral wrongfulness, or any other yardstick. Thirdly, it is as a consequence virtually impossible to meaningfully consider the 'proportionality' of a crime in terms of a particular form of punishment just by considering crime and punishment without the separate attachment of social values or moral assumptions.

Clearly, the nature of punishment is fundamentally different from that of crime. Punishment, according to the British philosopher Richard Stanley Peters, is 'the authoritative imposition of something regarded as unpleasant on someone who has committed a breach of rules'[44] and while criminal punishment is meted out in many different ways, the majority of sanctions take the form of either fines or jail sentences.[45] The punishing element of monetary fines is the deprivation of a sum of money, which is essentially the generation of a monetary loss for the offender. Because of this, fines are capable of being the only type of punishment potentially suitable for the concept of proportionality to be considered in sentencing in and of itself, where a *proportionate* financial loss is retributively inflicted on an offender as a punishment for having inflicted a financial loss. Because it is possible in those circumstances to mathematically calculate the monetary loss suffered by the victim, it is possible to formulate and impose an equal monetary loss on the offender, thus creating a meaningfully proportionate sanction.[46] However, monetary fines are but a small segment of criminal punishment in most legal systems; the form of punishment which is the subject of most debates concerning the principle of proportionality is incarceration.

---

44 Kleinig (n 33) 259 and 267.

45 For further discussion on the forms of punishment, see ibid 267-269.

46 Even where such 'proportionality' may be formulated, it should be noted that the victim's losses in terms of factors such as time, opportunities, and legal costs may at best be *estimated* by the sentencing judge, and ultimately render the punishment and the crime at least different to some degree.

The purposes of punishment through imprisonment are manifold and include incapacitation, retribution, deterrence, rehabilitation, and denunciation.[47] Amongst these, there is dispute over which goals should be considered or ignored, and how much weight or precedence each of them should carry. For instance, Morris Fish argues that retribution should have 'little or no role to play' in punishment, and that the purpose of punishment should instead be other utilitarian goals.[48] With regards to the punishing element of incarceration, however, incarceration is essentially the infliction of pain on the offender - the infliction of psychological and emotional 'loss' through the deprivation of one's liberty, normalcy, privacy, and often (whether intended or not), through the poor and unsafe conditions of prisons, the deprivation of dignity.[49] Indeed, '[a]t the very least, prison is painful, and incarcerated persons often suffer long-term consequences from having been subjected to pain, deprivation, and extremely atypical patterns and norms of living and interacting with others'.[50] Moreover, '[f]or some prisoners, incarceration is so stark and psychologically painful that it represents a form of traumatic stress severe enough to produce post-traumatic stress reactions once released'.[51] In addition to the pain inflicted upon the offender being imprisoned, incarceration also harms the family and children of the sanctioned offender, resulting in a punishing element which far exceeds the *prima facie* sentencing goal and range.[52] Incarceration, as the

47 Richard S Frase, 'Punishment Purposes' (2005) 58 Stan L Rev 67, 70; Craig Haney, 'The Psychological Impact of Incarceration: Implications for Post-Prison Adjustment' (2001) US Department of Health and Human Services working papers prepared for the 'From Prison to Home' Conference (January 30-31, 2002) 3.

48 Fish (n 25) 65.

49 Haney (n 47) 4-6.

50 ibid 4-5.

51 ibid 11.

52 For more discussion on this topic, see Joyce A Arditti, Jennifer Lambert-Shute and Karen Joest, 'Saturday Morning at the Jail: Implications of Incarceration for Families and Children' (2003) 52 Family Relations 195; Joyce A Arditti, 'Families

infliction of profound psychological (and in many cases, physical) pain through severe deprivations of action and association, has a destructive effect on an offender's private and family life.[53] It also impacts future career prospects,[54] and leads to other significant post-incarceration consequences on communities[55] and the offender's health (either through long-term incarceration or through infectious diseases).[56]

As such, when compared to the crimes which offenders are being punished for, the penalty of imprisonment (together with its far-ranging consequences) is too different to be meaningfully measured for 'proportionality'. The nature of crime and the nature of punishments (primarily incarceration) are so disparate that there is no meaningful way to compare the two on any scale on their own. Where one is the generation of losses on the victim of a crime which in most cases cannot be repaid, the other is the infliction of pain on the offender. The two are of

---

and Incarceration: An Ecological Approach' (2005) 86 J Contemporary Social Services 251; Justin Brooks and Kimberly Bahna, '"It's a Family Affair" - The Incarceration of the American Family: Confronting Legal and Social Issues' (1993) 28 USF L Rev 271; Jeremy Travis and Michelle Waul, *Prisoners Once Removed* (Urban Institute 2003) 189-225.

53 See (n 52).

54 Amanda Geller, Irwin Garfinkel and Bruce Western, 'The Effects of Incarceration on Employment and Wages - An Analysis of the Fragile Families Survey' (2006) Center for Research on Child Wellbeing Working Paper 2006-01-FF, <http://www.saferfoundation.org/files/documents/Princeton-Effect%20of%20Incarceration%20on%20Employment%20and%20Wages.pdf> accessed 23 April 2013; Bruce Western, 'The Impact of Incarceration on Wage Mobility and Inequality' (2002) 67 American Sociological Rev 526.

55 For a detailed exposition on the effect of incarceration on communities, see Dina R Rose, Todd R Clear and Judith A Ryder, 'Addressing the Unintended Consequences of Incarceration Through Community-Oriented Services at the Neighborhood Level' (2001) 5(3) Corrections Management Quarterly 62; Joan Petersilia, 'When Prisoners Return to Communities: Political, Economic, and Social Consequences' (2000) 65 Fed Probation 3.

56 For a detailed exposition on the effect of incarceration on the imprisoned individual's health, see Jason Schnittker and Andrea John, 'Enduring Stigma: The Long-Term Effects of Incarceration on Health' (2007) 48(2) J Health and Social Behavior 115; Michael Massoglia, 'Incarceration as Exposure: The Prison, Infectious Disease, and Other Stress-Related Illnesses' (2008) 49 J Health and Social Behavior 56.

completely different natures and it is impossible to weigh one against another without a prior conception of what the 'value' of losses in terms of emotional and physical pain are, a conception which cannot be based on the distinct natures of crime and punishment on their own, but which must find its basis on some other principle.

Even 'proportionality' based on the *lex talionis*, in which the principle of 'an eye-for-an-eye' prescribes an identical loss to be meted out as punishment for a loss inflicted by the offender, has been severely criticised. Apart from being a clearly primitive and barbaric form of punishment based on retaliation, the strict literal interpretation of the *lex talionis* has been described as 'overlooking its historical significance and moral relevance' such as that of preventing mob justice and vengeful violence.[57] Modern criminal sanctions no longer call for strict mirror punishments such as the amputation of an arm for causing the loss of another person's arm; implicitly recognising that criminal justice of this sort no longer has any currency in modern civilised society. Furthermore, as H.L.A. Hart observed,[58] mirror punishments are impossible in many instances anyway - the crime of theft cannot be punished by a theft, nor can defamation be recompensed by defamation. Because crime and punishment are of such fundamentally different natures, it is impossible to find an appropriate punishment that 'fits' any crime based on proportionality alone, and it is impossible and meaningless to claim that a punishment is, on its own, 'proportionate' to a crime without an extra and external benchmark to measure it against. What is retained from the *lex talionis*, however, is the fundamental underlying concept of proportionality. Nevertheless, the question which remains to be asked is whether 'proportionality' has any meaning if it is not to mirror a crime. Indeed, if *lex talionis* punishments are to be

57 Fish (n 25) 57.
58 HLA Hart, Punishment and Responsibility (OUP 1968) 161.

rejected, wherein lies the concept of 'proportionality'? It is difficult to see how any sanction can be designed to be 'proportionate' to a crime if it does not strive to be a clear mirror of that crime it is meant to punish.

Here, the theories of ordinal and cardinal proportionality offer some insight.[59] The former is concerned with how offenders of crimes of comparable gravity should be punished with sentences of comparable severity, *viz.* that similar crimes should attract similar penalties. Ordinal proportionality, then, is a matter of how different crimes may be measured against each other. The question which is left open, however, is how does one determine that a maiming, for instance, is 'comparable' with a rape, or the crime of defamation with the crime of theft? Fundamentally, the problem of how to compare different crimes remains unresolved. Cardinal proportionality offers a nuanced difference in approach. It is concerned instead with the overall severity levels anchoring the penalty scheme, so that the severity of punishments for the whole range of crimes in the criminal code should be determined in proportion to each other. Within the theory of cardinal proportionality, however, there is also much discussion over how to find anchoring points within the penal system so as to determine these calibrations.[60] As such, although both ordinal and cardinal proportionality may be useful in helping to formulate a concept of 'proportionality' that is meaningful in criminal sentencing, their utility only arises *after* there has first been an understanding of the underlying nature of proportionality as a reflection of social values. Only then can these comparisons and calibrations be measured and anchored.

---

59 von Hirsch (n 9) 282-283.

60 For a more detailed discussion on ordinal and cardinal proportionality, see von Hirsch (n 8) 75-84.

### *D. Meeting crime with punishment - proportionality as a reflection of sentiments*

Ultimately, 'proportionality' is a reflection of moral assumptions, opinions, estimates, and, often, the product of conscious or unconscious prejudices and preconceived notions such as racial stereotypes and other perceived correlations between members of a certain class and certain types of crime.[61] As it is impossible to mathematically calculate the value of a crime in terms of a criminal sentence, proportionality can at best be a measure of what is *perceived* to be the values attached to the losses of crime, and the values attached to the pains inflicted by punishment. There is no immediately discernible common benchmark between the gravity of crimes and the severity punishments on their own, so they can only be measured in proportion to each other insofar as they have been scaled according to the values attached to them by society or by the judiciary. As such, it is not crime and punishment themselves which are considered in proportion to each other, but the *values* attached to them which are used to make these comparisons. Therefore, it is possible to strive towards proportionality only after placing the spectrum of crimes and punishments on this scale of social values, from which they may then be compared. This is the only meaningful understanding of what proportionality involves when it is said to be applied in judicial sentencing. 'Proportionality' can only strive to be as proportionate as possible in reflecting these values, and its application can come through two theoretical steps.

Firstly, the different crimes to be sanctioned within the criminal justice system should be measured in proportion to each other according to public sentiment (either through the legislature which prescribes sentencing guidelines or by

61 See eg Steven E Barkan and Steven F Cohn, 'Racial Prejudice and Support for the Death Penalty by Whites' (1994) 31 J Research in Crime and Delinquency 202; Thorsten Sellin, 'Race Prejudice in the Administration of Justice' (1935) 41 American J Sociology 212; Philip A Currya and Tilman Klumpp, 'Crime, Punishment, and Prejudice' (2009) 93 J Public Economics 73.

the judiciary which forms case law), and then correspondingly set on a scale of varying degrees of severity. In the same way, the different punishments available as criminal sanctions should be set on a scale of proportionality against each other. This is in keeping with the theory of ordinal proportionality, in order to facilitate the conceptualisation of 'similar crimes' and 'similar punishments', where otherwise, objectively on their own, crimes can only differ amongst themselves just as different punishments amongst themselves, and neither can be compared with the other on grounds of similarity because no common benchmark exists. This benchmark must therefore be found not in crime and punishment themselves, but in the *opinions* which society harbours towards them. Indeed, because different crimes are of different natures - since a murder cannot on its own be compared as a measure of similarity to a rape, for instance - they can at best be compared based on society's valuation of their harm or repulsiveness. A rape is so different from murder, and the loss of one's dignity so disparate from the loss of one's life, that it is impossible to judge from the character of a crime itself to objectively say that the loss of a life is necessarily worse than the loss of one's dignity as a person, the remnant of which may be a life of pain and shame. As such, it is the values of each society, reflected in their respective criminal codes, that produce 'proportion' between different offences and sanctions. This proportionality does not exist on it own but is ultimately a reflection of each society's moral assumptions, estimates, opinions, and sentiments. Just as individual members of society harbour each their own value systems and moral assumptions concerning crime and punishment, contributing to general social sentiments toward the concept of justice, so too legislators and judges whose endeavours to achieve just and fair laws and judgements through the principle of proportionality reflect not only their personal value systems, but also that of general society.

In the same way, all available sanctions in the criminal justice system must also be set upon a scale, so that

the severity of each punishment is weighed against other sanctions. Again, such an endeavour will necessarily be done through the consideration of social values in order to determine, for instance, how the severity of a particular monetary fine compares in proportion to an incarceration sentence. How does a $100,000 fine weigh against a five-year imprisonment term? The proportionality scale of punishments, like that of crimes, will therefore be a scale of the opinions and values that society attaches to them. Therefore, the product of these efforts will be two different scales of proportionality: one scale of the various criminal offences weighed in proportion to each other based on their attached societal values, and another scale of all the punitive sanctions available, weighed against the prevailing sentiments of society to plot them along a proportional range.

Secondly, these two scales - of crimes and of punishments - must be anchored against each other so that there may be points of intersection between the two, from which other offences and sentences may then be meaningfully compared in proportion with each other. This happens either through case law, or through legislation prescribing that a particular crime should attract a particular punishment (or range of punishments), and from which other identified offences and sanctions in the criminal code are then scaled accordingly. On their own, the proportionality scale of crimes do not relate with the proportionality scale of punishments, and in order to compare the two, there needs to be a value judgement of how a crime may measure against a punishment, such that the two may be considered in 'proportion' to each other. For instance, what should be the appropriate punishment meted out for the crime of rape? The crime does not, on its own, prescribe the 'proportionate' punishment it should attract, but social opinions and sentiments may demand a punishment which is, in accordance with moral assumptions, 'proportionate' to that crime. Indeed, '[a]ccording to the principle of proportionality, punishment is supposed to comport with the seriousness of the crime. There does not,

however, seem to be any precise way of fixing the deserved amount of punishment. Armed robbery is a serious crime, but it is not apparent whether its punishment should be two years' confinement, three years' confinement, or some milder or some more severe sanction.'[62] Therefore, this anchoring of the scale of crimes against the scale of punishments ultimately depends on moral assumptions and displays a symbolic valuation of societal sentiments.[63]

It is clear, therefore, that the concept of proportionality can only be understood meaningfully if it is acknowledged to be the reflection of a society's opinions, values, and moral assumptions. There cannot be proportionality between two things of disparate natures, and in order to compare crime and punishment, one must compare the *sentiments* that people hold towards them. This is the true 'proportionality' which the criminal justice system strives towards. Necessarily, these opinions will be strongly debated and the myriad values of society will undoubtedly wrestle with each other to be applied in the law, but this is the natural exercise of common public policy. For example, the issue over whether the death penalty is a proportionate criminal sanction for certain crimes is an old and still hotly disputed current debate, epitomising the sort of struggles determining proportionality in criminal punishment. There are differing views over whether execution is proportionate to the crimes it is used for, based not on the nature of execution nor of the nature of those crimes alone - since it is impossible to come to any objective conclusion about how a murder, for example, on its own is decidedly either proportionate or disproportionate to the termination of an offender's life through lethal injection - but rather, is based on what society *perceives* the evil of murder to be, and the associated values they attach to the sanctity of human lives, as well as the state's role and responsibilities in these

62 von Hirsch (n 9) 283.
63 ibid 283-284.

matters. Similarly, other punitive sanctions such as monetary fines and incarceration are weighed in 'proportion' against crimes, based on social sentiments and moral assumptions attached to them. Because proportionality is not an objective truth to be discovered from the natures of crime and punishment on their own, but rather is the manifestation of subjective human sentiments toward the evils of crime and the utility of punitive sanctions, the best that the criminal justice system can do is only to strive ever closer to a 'proportionality' which reflects the norms of the society it is meant to serve.

## IV. Applying the proportionality principle

Having thus acknowledged that the principle of proportionality is really the reflection of ever-changing social sentiments and moral values rather than an objective conclusion to be derived from a comparison of crimes and punishments on their own, it is clear that proportionality can only ever be strived towards as an ideal, rather than attained completely. The practical application of the proportionality principle therefore raises several issues.

Firstly, given that proportionality in criminal sentencing is a reflection of sentiments, legislators and judges have a large discretion in determining which punishments are 'proportional' to different crimes, giving rise to potentially arbitrary results in legislation and judgements. Although the social sentiments and moral assumptions that attach values to crimes and punishments will undoubtedly be restrained by good reasoning and logical explanations in Parliament and courtrooms, because opinions and sentiments are so fluid and subjective, there remains a large potential for abuse. After all, how does a judge determine if the crime of defamation should attract a monetary fine of $5000 or $7000? How does the legislature assess the values that society attaches to the incarceration sentences of five years and ten years? While the legislature and judiciary will undoubtedly take into consideration all factors that are possible to be assessed, ultimately, however, these are

estimates at best, and will require the input of norms and values which can be callously arbitrary and unreflecting of the prevailing social sentiments.

Clearly, the most difficult aspect of applying proportionality in criminal justice is in determining what is 'proportional' in the first place, *viz.*, deciding which punishments are considered to correspond to which crimes. There are no easy answers to these questions, which is why judges hear cases individually to decide on the scale of proportionality, taking into consideration all the facts and the social values attached to those facts, just as Parliament debates with the resources it is endowed with in order to determine the best estimates it is able to find. Hence, proportionality is an ideal which is continually strived towards, through which the law endeavours to come as close as possible to reflecting the evolving values of society.

It is through this that the principle of proportionality is able to concurrently set boundaries to limit discretion in criminal sentencing, since it requires judges to take into account the prevailing social sentiments when sentencing. Herein lies the utility of sentencing codes which require judges to impose only proportionate sentences for crimes, not because there exist punishments which naturally correspond with crimes on their own, but because the law needs to reflect social norms.[64] It is through the consideration of what values are attached to crime and punishment, and the moral assumptions underlying public opinion, that judges may mete out sanctions that fulfil the purpose of the law to 'maintain respect for the law and a safe society by imposing just sanctions'.[65] As case law develops in particular areas of crime, with each judge establishing a precedence striving ever closer to the values of prevailing social sentiments, a range of proportionality emerges from which sentencing judges cannot depart without evident

64 See eg Canadian *Criminal Code*, RS C 1985, c C-46, s 718 1.
65 *R v Arcand* (n 6) at 52.

changes in public opinion. This is the meaningful application of proportionality, that judicial discretion is restricted because judges must impose sentences which are proportionate, and this proportionality is established through the consideration of social norms. Thus, proportionality is an ideal and guide for judges, to restrict arbitrary discretion in sentencing, to aid in reflecting prevailing societal opinion towards criminal justice, and to uphold the values which are attached to them by imposing sanctions that are in keeping with these moral assumptions. Indeed, it is only by so doing that the criminal justice system is able to reconcile the proportionality principle with other goals of punishment such as denunciation, deterrence, and rehabilitation, since these are the very sort of concerns which shape and define the social sentiments and values that society attaches to crime and punishment.

As such, proportionality is an enterprise which seeks to come closer and closer to encapsulating and reflecting all of these myriad concerns - concerns over what society opines about crime and punishment and the values they attach to them, concerns about achieving the other goals of punishment, and concerns over limiting judicial discretion so as to reflect the prevailing societal sentiments towards criminal justice. In the application of the proportionality principle, therefore, judges strive towards coming ever closer to the goal of satisfying all of these concerns, so that crime and punishment, although of disparate natures that cannot meaningfully be compared against each other, may be placed on a scale from which they *can* be measured against each other. It is on this scale of proportionality, formulated through the social values and moral assumptions attached to criminal justice, that the meaningful and useful concept of proportionality as an ideal can be found.

## V. Conclusion

Proportionality in criminal justice is derived not from merely considering crime and punishment on their own, but through taking into account the social sentiments towards

them, as well as the values attached to crimes and punishments. The application of the proportionality principle, then, is not an objective measurement to be made of criminal offences and sanctions, but is a comparison of the moral assumption that society harbours towards them. Therefore, proportionality can be reached by first scaling crimes and punishments according to these social values, and then by anchoring these two scales against each other, from which calibrations and meaningful comparisons can then be made, and a practical application of proportionality may then be derived. As such, proportionality is never truly *attained*, since it is not an objective truth to be discovered from the observation of criminal offences and punishments, but is an enterprise of striving towards the goal of representing the wide-ranging and evolving values of society.

## BIBLIOGRAPHY


### Jurisprudence

*R v Arcand*, [2010] AJ No 1383 (Alta CA) 55

*R v Smith (Edward Dewey)*, [1987] 1 SCR 1045

### Legislation

Canadian Charter of Rights and Freedoms, Part I of the Constitution Act, 1982 being Schedule B to the Canada Act 1982 (UK), 1982, c 11

Canadian Criminal Code, RS C 1985, c C-46

### Articles

Alschuler A W, 'The Changing Purposes of Criminal Punishment: A Retrospective on the past Century and Some Thoughts about the Next' (2003) 70 U Chicago L Rev 1

Arditti J A, 'Families and Incarceration: An Ecological Approach' (2005) 86 The Journal of Contemporary Social Services 251

——, Lambert-Shute J and Joest K, 'Saturday Morning at the Jail: Implications of Incarceration for Families and Children' (2003) 52 Family Relations 195

Barkan S E and Cohn S F, 'Racial Prejudice and Support for the Death Penalty by Whites' (1994) 31 Journal of Research in Crime and Delinquency 202

Brooks J and Bahna K, '"It's a Family Affair" - The Incarceration of the American Family: Confronting Legal and Social Issues' [1993] USF L Rev 271

Currya P A and Klumpp T, 'Crime, Punishment, and Prejudice' (2009) 93 Journal of Public Economics 73

Fish M J, 'An Eye for an Eye: Proportionality as a Moral Principle of Punishment' (2008) 28 Oxford J Legal Stud 57, 66

Frase R S, 'Excessive Prison Sentences, Punishment Goals, and the Eighth Amendment: "Proportionality" Relative to What?' (2004) 89 Minn L Rev 571

——, 'Punishment Purposes' (2005) 58 Stan L Rev 67

Grossman S, 'Proportionality in Non-Capital Sentencing: The Supreme Court's Tortured Approach to Cruel and Unusual Punishment' (1994) 84 Ken L Rev 107

Ivins W M, 'What is Crime?' (1911) 1 Reform of the Criminal Law and Procedure 531

Kleinig J, 'R. S. Peters on Punishment' (1972) 20 British Journal of Educational Studies 259

Massoglia M, 'Incarceration as Exposure: The Prison, Infectious Disease, and Other Stress-Related Illnesses' (2008) 49 Journal of Health and Social Behavior 56

Perry R, 'The Role of Retributive Justice in the Common Law of Torts: A Descriptive Theory' (2006) 73 Tenn L Rev 177

Petersilia J, 'When Prisoners Return to Communities: Political, Economic, and Social Consequences' (2000) 65 Fed Probation 3

Radin M J, 'The Jurisprudence of Death: Evolving Standards for the Cruel and Unusual Punishments Clause' (1978) 126 U Pa L Rev 989

Rose D R, Clear T R and Ryder J A, 'Addressing the Unintended Consequences of Incarceration Through Community-Oriented Services at the Neighborhood Level' (2001) 5(3) Corrections Management Quarterly 62

Schnittker J and John A, 'Enduring Stigma: The Long-Term Effects of Incarceration on Health' (2007) 48(2) Journal of Helath and Social Behavior 115

Sellin T, 'Race Prejudice in the Administration of Justice' (1935) 41 American Journal of Sociology 212

Singer R G, 'Sending Men to Prison: Constitutional Aspects of the Burden of Proof and the Doctrine of the Least Drastic Alternative as Applied to Sentencing Determinations' (1972) 58 Cornell L Rev 51

von Hirsch A, 'Proportionality in the Philosophy of Punishment: From 'Why Punish?' to 'How Much?'' (1990) 1 Criminal Law Forum 259

——, 'Proportionality in the Philosophy of Punishment' (1992) 16 Crime and Justice 55

Walker L and Hayashi L, 'Pono Kaulike: Reducing Violence with Restorative Justice and Solution-Focused Approaches' (2009) 73(1) Federal Probation Journal 23

Western B, 'The Impact of Incarceration on Wage Mobility and Inequality' (2002) 67 American Sociological Review 526

Whitman J Q, 'Between Self-Defense and Vengeance / Between Social Contract and Monopoly of Violence' (2004) 39 Tulsa Law Review 901

## Books

Becker G S, 'Crime and Punishment: An Economic Approach' in Gary S Becker and William M Landes (eds), *Essays in the Economics of Crime and Punishment* (UMI 1974)

Brettschneider C, 'Rights Within the Social Contract: Rousseau on Punishment' in Austin Sarat A, Douglas L, Umphrey M M (eds), *Law As Punishment / Law As Regulation* (Stanford University Press 2011)

Cavadino M and Dignan J, *The Penal System: An Introduction*, (2nd edn, Sage 1997)

Reckless W C, *Criminal Behavior* (McGraw-Hill 1940)

Hart HLA, *Punishment and Responsibility* (Oxford University Press 1968)

Travis J and Waul M, *Prisoners Once Removed* (Urban Institute 2003)

Zimring F E, Hawkins G and Kamin S, Punishment and Democracy: Three Strikes and You're Out in California (Oxford University Press 2001)

## Other Material

Bedau H A and Kelly E, 'Theory of Punishment' (Stanford Encyclopedia of Philosophy, 19 February 2010) <http://plato.stanford.edu/entries/punishment/#2> accessed 12 April 2012

Geller A, Garfinkel I and Western B, 'The Effects of Incarceration on Employment and Wages - An Analysis of the Fragile Families Survey' (2006) Center for Research

on Child Wellbeing Working Paper 2006-01-FF, <http://www.saferfoundation.org/files/documents/Princeton-Effect%20of%20Incarceration%20on%20Employment%20and%20Wages.pdf> accessed 23 April 2013

Haney G, 'The Psychological Impact of Incarceration: Implications for Post-Prison Adjustment' (2001) US Department of Health and Human Services working papers prepared for the 'From Prison to Home' Conference (January 30-31, 2002) 3

# MARPOL 73/78: The Challenges of Regulating Vessel-Source Oil Pollution

Mark Szepes

## Abstract

*Merchant shipping is the major method of international transportation for all types of goods, including oil. Shipping has been a major cause of degradation to the marine environment due to operational and accidental discharge of oil which accounted for an estimated 2 million tons of oil entering the world's oceans in the 1980s. To put this into perspective, the grounding of Exxon Valdez in 1989 resulted in a discharge of 35,000 tons of crude oil which is estimated to have killed 250,000 sea birds, 2800 otters, 300 seals and 13 orca whales. It also required over $3.5 billion in clean-up costs.*

*MARPOL is the international convention that has been brought into effect to protect the oceans of the world. Annex I was specifically created to prevent and reduce oceanic oil discharges.*

*This article examines the challenges that are faced by such an ambitious international regulation that combines International, Environmental and Maritime Law. Many of these challenges are connected to unique jurisdictional gaps and overlaps. In most cases there is more than one jurisdiction which may take action in response to a suspected violation. However, in many cases, states tend to defer responsibility for reasons such as the costs involved in taking action.*

*The overall success of MARPOL will be measured by the impact it has had in achieving its objective. The conclusion will be reached that MARPOL is a legitimate international regime that has made significant progress in achieving its objectives, but still has some way to go.*

## I. Introduction

International trade would be impossible without the marine shipping industry. Merchant ships around the world transport the majority of the products considered essential to international trade as well as everyday life. Those products include manufactured goods, food products, raw materials

and the principle source of global energy, oil.[1] Shipping as the major method of transportation of goods around the world has had a costly impact on the environment but it is only in the last 50 years that this impact has been acknowledged.[2] The significant sources of degradation have been identified as both operational and accidental vessel-source oil pollution from the continually increasing international merchant fleets.

Prior to the Second World War, the accepted practice for managing shipboard waste was to, as it were, 'throw it into Davy Jones' locker'.[3] This practice and lack of concern for the ocean environments encouraged the pollution of the sea. It is the realization of damage being done to the oceans that has led to the development and implementation of international law to eliminate marine pollution. The focus of this article is on a significant aspect of marine pollution, that being, vessel-source oil pollution.

The International Convention for the Prevention of Pollution from Ships 1973 as amended by the Protocol of 1978, which is more commonly known as MARPOL 73/78, is the most ambitious attempt on a global level to prevent marine pollution from operational activities and accidents.[4] Every vessel at sea, regardless of size or purpose generates oily waste. This waste is generated by the operation of the vessel, and additionally through the transportation of oil. MARPOL was created as an organic regulation with an expectation that it would expand over time and include additional environmental aspects. This expansion has occurred: at present there are six annexes, each dealing with a different type of pollution from ships. It is MARPOL

1 Patricia Birnie, Alan Boyle, Catherine Redgwell, *International Law and the Environment* (3rd Edition, OUP 2009) 398.

2 Nickie Butt, 'The Impact of Cruise Ship Generated Waste on Home Ports and Ports of Call: A Study of Southampton' (2007) 31 Marine Policy 591.

3 ibid.

4 Manfred Nauke, Geoffrey L Holland, 'The Role and Development of Global Marine Conventions: Two Case Histories' (1992) 25 Marine Pollution Bulletin 74.

Annex I that deals with the significant issue of oil pollution, and for this reason this article will focus on Annex I.[5]

In the late 1980s, when MARPOL had only recently come into effect, it was estimated that vessels released 2 million tons of oil into the sea.[6] By 2007, it was estimated that this figure had been reduced to 450,000 tons of oil entering the marine environment annually.[7] This decline indicates a significant decrease in a major cause of marine degradation and the position that this should be credited to MARPOL will be demonstrated. MARPOL is not intended to totally eliminate all oil discharges into the sea, however, the point has not been reached where those discharges have reached a level that has no more of an impact on the environment than that of naturally occurring oil releases.

The location and concentration of vessel related discharges can have a catastrophic impact on the marine and coastal environment. This impact is demonstrated by the fact that a single discharge of 35,000 tons of crude oil, a result of the grounding of the *Exxon Valdez* in 1989[8], is estimated to have killed 250,000 sea birds, 2800 sea otters, 300 harbour seals, and 13 orca whales, as well as shutting down the commercial Alaskan salmon fishery and requiring over $3.5 Billion USD in clean-up costs.[9] Without the MARPOL Annex I discharge standards, tanker vessels would be discharging up to 2500 tons of crude oil for each voyage they make; which could possibly amount up to 10 million tons per year.[10]

---

5 Butt (n 2) 594.

6 Andrew Griffin, 'MARPOL 73/78 and Vessel Pollution: A Glass Half Full or Half Empty?' (1994) 2 Indiana J Global L Studies 489.

7 Birnie (n 1) 381.

8 Ronald B Mitchell, *Intentional Oil Pollution at Sea: Environmental Policy and Treaty Compliance* (The MIT Press 1994) 82.

9 John M Weber, Robert E Crew, 'Deterrence Theory and Marine Oil Spills: Do Coast Guard civil Penalties Deter Pollution?' (2000) 58 J Environmental Management 161.

10 Mitchell (n 8).

In exploring the issues of vessel-source oil pollution, there will be an examination of the international regulations which have brought about the significant reduction in vessel-source oil discharge. Additionally, attention will be drawn to the difficulties faced in further achieving the mandate of Annex I. One cause of these difficulties is due to MARPOL being a hybrid of international, environmental and maritime law. This unique composition produces significant challenges, specifically those associated with jurisdictional and operational success.

The opening section of this article will outline the historical context which led to the adoption of MARPOL. This is the starting point for evaluating the success of the convention in achieving its ambitious goals. The subsequent sections will focus on the jurisdictional challenges encountered in enforcing MARPOL, as well as the issues connected to the operation of key requirements within the regulation. The conclusion will be made that supports the position that MARPOL should be viewed as a legitimate international regime. It is worth briefly noting at this point that a 'flag state' is the state to which a ship is flagged and registered, a 'coastal state' is a state which has territorial waters due to its location bordering an ocean or sea and a 'port state' is the state where a ship calls into port for any purpose.

## II. Development Of Maritime Pollution Regulations

### *A. Early Developments in Vessel-Source Pollution Regulation*

Development of international law with the aim of regulating vessel-source pollution beyond the territorial three nautical mile limit occurred in the early 20th century. This development took place as a result of significant political pressure from both the United Kingdom and the United

States.[11] This pressure led the two draft conventions: namely, the 1926 Washington Draft Convention and the League of Nations Draft Convention.

Although they were drafted, the conventions were never adopted formally. The outbreak of the Second World War resulted in the suspension of any action in relation to vessel-source pollution control.[12] In the post-war period, and on account of the rapid growth of the global economy and the enhanced demand for energy resources, attention returned once again to protection of the marine environment from shipping-related pollution. In 1948, the United Nations took the first post-war steps to address the issue of vessel-source pollution of the marine environment by holding an international maritime conference in Geneva.

This ultimately led to the establishment of the Inter-Governmental Maritime Consultative Organization (IMCO). This organization would eventually come to be known as the International Maritime Organization (IMO) and this transformation took place through the process of amendments to the conventions of the IMCO in 1982.[13] The progression of the new IMCO from establishment to becoming operational was protracted, as the IMCO did not become operational until 1958.[14]

During the development stages of the IMCO between 1948- 1958, the UK began to acknowledge the need for immediate action in the area of vessel-source marine pollution. This was the result of growing public concern with regard to oil discharges from ships, and the impact it had on

---

11 Alan Khee-Jin Tan, *Vessel Source Marine Pollution: The Law and Politics of International Regulation* (Cambridge University Press 2006) 107.

12 ibid 109.

13 International Maritime Organization, *MARPOL 73/78 Consolidated Edition, 2002* (IMO 2002).

14 Rebecca Becker, 'MARPOL 73/78: An Overview in International Environmental Enforcement' (1997) 10 Georgetown Int Environmental L Rev 626.

the marine environment.[15] The UK's dedication to taking action in this area was demonstrated by the creation of the Committee on the Prevention of Pollution of the Sea by Oil, which was chaired by Lord Faulkner, to explore potential global measures to harmonize regulatory action regarding oil discharges.[16]

Following the report of the Faulkner Committee in 1953, a diplomatic conference was called in London in May 1954 with the intention of negotiating an international convention on this subject.[17] The London Conference is held to have been a success, as it was the birthplace of the first multilateral agreement on oil pollution control. This agreement became known as the International Convention for the Prevention of Pollution of the Sea by Oil (OILPOL), which came into force on the 26th of July 1958.

### *B. OILPOL: The Birth of Multilateral Marine Pollution Agreements*

Essentially, OILPOL prohibited the release of oily waste into the sea within a 50 nautical-mile coastal zone. The prohibition predominately targeted oil tankers, whilst non-tanker commercial vessels were largely unaffected. In truth, the restriction on tankers was limited. When operating outside of the coastal zones, within the majority of the world's oceans, tanker crews were generally free to discharge oily waste.[18]

In addition to being limited in scope[19], OILPOL lacked sufficient enforcement controls for coastal and port

---

15 Ronald B Mitchell, 'Regime Design Matters: Intentional Oil Pollution and Treaty Compliance' (1994) 48 International Organization Foundation 431.

16 R Michael M'Gonigle, Mark W Zacher, *Pollution, Politics and International Law: Tankers at Sea* (University of California Press 1979) 84.

17 Tan (n 11) 110.

18 ibid 111.

19 The lack of significant prohibition regulating oil discharges was due to the active dispute among nations at the 1954 conference as to whether there were harmful impacts on the marine environment from oil discharges (Mitchell 1994: 84).

states. Responsibility was passed to a vessel's flag state once it had been informed of an alleged violation. The flag state was to investigate the matter, and if it determined there was sufficient evidence to initiate proceedings it could elect to do so.[20] Due to the limited ability of coastal and port states to monitor oily discharge, and a general reluctance by flag states to prosecute alleged offending vessels, OILPOL was not as effective in dealing with oil pollution as had been the intention of the UK as the leading party to the London Conference.[21]

The events surrounding the *Torrey Canyon,* which in on March 1967 ran aground near the Isles of Scilly and released its cargo of 120,000 tons of crude oil, probably had the largest impact on changing marine pollution regulations. Being the largest oil spill ever recorded up to that time,[22] it drew attention to the fact that vessel-source oil pollution was a serious problem that needed to be addressed. Although accidental pollution, such as the *Torrey Canyon,* was often more visible to the public, it was actually operational pollution that resulted in a much more consistent and significant source of oily discharge.[23]

In an effort to reduce the amount of operational discharge at sea and to pre-empt regulation, oil companies led by Shell Oil established the practice known as Load On Top[24] (LOT).[25] LOT reduced oily discharge, however,

---

20 International Convention for the Prevention of Pollution of the Sea by Oil, 1954 Article X.

21 M'Gonigle and Zacher (n 16) 89.

22 Tan (n 11) 120.

23 International Maritime Organization, 'Brief History of IMO' <http://www.imo.org/about/historyofimo/Pages/Default.aspx> Accessed 1 May 2012.

24 Jeff B Curtis, 'Vessel-Source Oil Pollution and MARPOL 73/78: An International Success Story?' (1984) 15 Environmental L 689.

25 LOT allowed for ballast water to be taken on after oil had been offloaded in port, during the return journey separation would occur and oily sludge would settle on top of the water which could be discharged without the significant release of oil. A new cargo of oil would be loaded on top of the remaining oil without the need to discharge oily sludge into the sea (ibid 690).

there still remained significant technical shortcomings within LOT and operational pollution continued to occur.[26] The official requirement for LOT and a modification of discharge standards were brought into effect through a 1969 amendment to OILPOL. This amendment did not make any adjustments to the compliance and enforcement measures of the convention.[27] When the 1969 amendment was in the process of being brought into force, the maritime nations which had initially supported OILPOL came to the agreement that it was no longer adequately suited to fulfil its mandate.[28] Thus, the shortcomings of OILPOL were the catalyst that brought MARPOL into existence.

### *C. From OILPOL to MARPOL 73*

Following the *Torrey Canyon* disaster, the United States was forced to respond to public pressure, and it did so in a drastic manner. The response by the United States President Nixon's administration was to create the Environmental Protection Agency (EPA) in 1970. The mandate of the newly established EPA was to protect the natural environment of the US.[29] The US objected to the poor state of international regulation on vessel-source pollution, and lobbied for improvements.

As a result of US influence, and with the support of a number of other maritime states, reform began to take shape in a manner that would significantly impact the issue of vessel-source pollution.[30] The 1973 International Conference on Marine Pollution in London was attended by 71 states representing both the developed and developing world. It was the International Conference on Marine Pollution that was the birthplace of the International

26M'Gonigle (n 16) 102.
27 Tan (n 11) 121.
28 IMO (n 23).
29 M'Gonigle (n 16) 107.
30 ibid 109.

Convention for the Prevention of Pollution from Ships (MARPOL 73).[31]

MARPOL 73 was adopted by the International Conference on Marine Pollution. This conference was convened by the IMCO largely as a result of US determination to drive change. Although MARPOL 73 was adopted, it was unable to meet the double ratification requirements [32] for several years after it had been negotiated.[33]

Due to the inability to ratify, combined with recognition that MARPOL 73 was necessary, the Convention was modified by the Protocol of 1978. The result of this modification was the creation of the regulation known as MARPOL 73/78. [34] MARPOL 73/78 (MARPOL) successfully met the double ratification threshold and came into effect in October 1983, with the mandate of eliminating international pollution of the marine environment. [35] MARPOL superseded OILPOL, which had been the previous regulation relevant to dealing internationally with marine pollution from oil[36].

## *D. International Regulation and the Position of the IMO*

Marine pollution is a concept which crosses national boundaries, and is governed by International, Regional and Domestic Laws. This has resulted in overlaps of applicable

---

31 ibid 112.

32 The requirement of ratification contains a double threshold which must be achieved. The double threshold being at minimum 15 State that account for at least 50% of the gross tonnages of the international merchant shipping fleet (IMO 2002: Article 15(1)).

33 Elizabeth R DeSombre, Global Environmental Institutions (Routledge 2006) 74.

34 IMO (n 13) iii.

35 John McEldowney, Sharron McEldowney, *Environmental Law* (Pearson Education Limited 2010) 35.

36 MARPOL also applies to any technical aspects of pollution from all types of ships, something that was far beyond the mandate of OILPOL (Tan 2006: 129).

laws and regulations as well as jurisdictions. The United Nations has played an important role in codifying the various treaties and conventions regulating marine pollution. This challenging process commenced with the adoption of the United Nations Convention on the Law of the Sea 1982 (UNCLOS). The ultimate objective of UNCLOS was to create a single consolidated legal instrument that eliminated contradictions and overlap, and to ensure that all gaps in international law were filled.[37]

UNCLOS commenced the process of creating a climate of clarity in relation to governance, and establishing where authority lies in connection to different aspects of the law of the sea. The most significant contribution in this area is in relation to the enhancement of jurisdictional zones for coastal states.[38] UNCLOS reinforces the role of the IMO and the regulations which were created by it. This is done via the designation of certain specific functions to the "competent international organization", a reference that has been accepted to mean the IMO.[39] UNCLOS has accepted and endorsed the IMO regulations through references to the "generally accepted international rules", those being interpreted as MARPOL and SOLAS (Safety of Life at Sea).[40]

The IMO is the international body responsible for setting maritime vessel safety regulations and marine pollution standards. The IMO is a body of the United Nations and is composed of members from over 150 nations.[41] All states which are members of the UN may join the IMO. Any state that is not a member of the UN has the ability to join the IMO provided that the candidate state

37 DeSombre (n 33) 80.

38 ibid.

39 ibid 83.

40 ibid.

41 Claudia Copeland, *Cruise Ship Pollution: Background, Laws and Regulation, and Key Issues* (Congressional Research Service Report for Congress, 2008) CRS-7.

receives endorsement from at least two-thirds of the existing members of the IMO.[42]

The structure of the IMO in relation to decision-making is straightforward. The IMO Assembly, composed of all member states, is the primary decision making body and is mandated to meet every second year. The IMO Council is the body which coordinates the business of the IMO when the Assembly is not in session. The Council is a more manageable group made up of 32 of the member states and the members of Council are elected by the Assembly to serve a two year term and during this term Council must meet at least twice per year.

The Council is not empowered to make recommendations on behalf of the IMO to national governments in areas related to maritime safety and prevention of pollution, as this function is restricted to that of the Assembly.[43] The IMO contains two significant committees which are open to all IMO members, as well as non-members who are parties to the SOLAS and MARPOL conventions. These committees are the Maritime Safety Committee (MSC), which deals with all matters related to maritime safety, and the Marine Environment Protection Committee (MEPC), which deals with all matters related to prevention and control of pollution from ships, and specifically the adoption and enforcement of conventions and regulations related to pollution.[44]

## *E. The International Convention for the Prevention of Pollution from Ships*

The MARPOL convention, as noted, contains six annexes which provide the technical substance on the

42 This requirement is contained in Articles 6 and 8 of the Convention of the IMCO.

43 International Maritime Organization, 'Structure of IMO' <http://www.imo.org/About/Pages/Structure.aspx> Accessed 1 May 2012.

44 ibid.

international standards for protection of the environment from discharge of waste by ships.

For the MARPOL convention to be held as binding, ratification must occur by member states. The requirement is that the number of states which ratify each annex must represent at minimum 50% of global shipping gross tonnage, and be at least 15 states in total. This is known as the double threshold and has not been modified since the adoption of the original MARPOL regulation.[45] All six of the annexes have been ratified as of 2005. Once a state has become a signatory to MARPOL it is that state's responsibility to create and enact domestic legislation which will implement the convention rules. This includes the compulsory annexes (Annex I and II) and the voluntary annexes (Annexes III to VI) to which the country has agreed. The domestic legislation must recognize the related legislation of other MARPOL signatory states and agree to comply with that legislation.[46] Ships that are flagged under a state which is a signatory to MARPOL are subject to the convention regardless of where they sail or operate. It is the duty of the flag state to be responsible for the vessels which are registered under their flag.[47]

There is a very high level of acceptance of IMO negotiated agreements; this is likely due to the fact that the majority of the major shipping states participated in the conventions where the agreements were created, and states are more likely to accept an agreement if they took part in the process of creating them. This level of acceptance is demonstrated by the fact that the states to which 98% of the world's merchant tonnage is registered to, have accepted and become parties to MARPOL.[48]

---

45 Copeland (n 41) CRS-8.
46 ibid CRS-8.
47 ibid CRS-7.
48 DeSombre (n 33) 74.

An overview of the significant components of MARPOL is necessary in order to understand the manner in which the convention is meant to operate. Once this overview has been completed, it is possible to examine the issues that are faced in achieving the MARPOL objectives.

## *F. MARPOL Overview*

The International Convention for the Prevention of Pollution from Ships is laid out in a manner which allows for amendments and additions to take place within the Annexes, and not require a major overhaul of the entire convention. MARPOL 1973 is organized into 20 Articles. It is these articles that lay out what the convention parties have agreed upon. Preceding the articles is the preamble which recognizes that there is a need to preserve the marine environment, and that deliberate, negligent and accidental release of oil from ships is a major source of pollution which results in damage to the environment.

With recognition of the key issues taking place, the intention of MARPOL 73 is stated to be the complete elimination of intentional and accidental pollution of the marine environment from oil and all other harmful substances. Finally, it is held that the best method of achieving this is through the establishment of rules.[49] The 20 articles are laid out over 13 pages and provide a framework for MARPOL 73. These articles include: the general obligations under the convention (Art 1), seven key definitions for clarification purposes (Art 2), application (Art 3), violation (Art 4), certificates and special rules on inspecting ships (Art 5), detection of violations and enforcement of the convention (Art 6), undue delay to ships (Art 7), reports on incidents involving harmful substances (Art 8), other treaties and interpretation (Art 9), settlement of disputes (Art 10), communication of information (Art 11), casualties to ships (Art 12), signature, ratification, acceptance,

49 IMO (n 13) 3.

approval and accession (Art 13), optional annexes (Art 14), entry into force (Art 15), amendments (Art 16), promotion of technical co-operation (Art 17), denunciation (Art 18), deposit and registration (Art 19), and languages (Art 20).

Following MARPOL 73 is the Protocol of 1978. In the preamble to this protocol there is an outline of the reasons for its addition. It is recognized that the International Convention for the Prevention of Pollution from Ships can make a significant contribution to the protection of the marine environment, and there is the need to implement the regulations contained within Annex I in order to achieve the prevention of pollution by oil. However, there was a need to defer the implementation of Annex II due to a number of technical problems.

The Protocol of 1978 is very brief and laid out over five pages and nine articles. The main objective of the protocol as set out in Article I is to give effect to MARPOL 73, including Annex I. Article II contains the main structural amendment to MARPOL 73, that being the delay of the implementation of Annex II for a period of 3 years. This period may be extended by approval of two-thirds of the parties to MARPOL 73 who are members of the MEPC. Article III provides an amendment to 11(1)(b) in regards to communication of MARPOL 73. Article IV provides a revised procedure for: signature, ratification, acceptance, approval and accession. Article V provides the ratification requirements and holds that once ratified the protocol will come into force 12 months from the date of ratification. This, in essence, provides an extra year to the three year delay of the implementation of Annex II. Articles VI-IX set out respectively the procedure for amendments, denunciation, the depository and languages, and their relation to MARPOL 73.

The Protocol of 1978 relating to the International Convention for the Prevention of Pollution from Ships 1973, once ratified, would establish the International Convention for the Prevention of Pollution from Ships 73/78 (MARPOL). There also exist two additional protocols,

Protocol I 'Provisions concerning reports on incidents involving harmful substances' and Protocol II 'Arbitration'. These protocols supplement Articles 8 and 10 of the MARPOL convention by providing additional details and requirements.

With the historical development of MARPOL having been examined, as well as an explanation of the structure of the regulation, it is possible to now focus on the challenges faced by MARPOL as an international law regulating marine pollution.

## III. MARPOL and Jurisdiction

This section will initially focus on the impact of having multiple jurisdictions associated with the regulation of vessel-source oil pollution. There are overlaps between the jurisdiction of flag, port and coastal states, and due consideration will be given to the challenges that result from this overlap and how MARPOL and UNCLOS operate in practice in this area.

### *A. The Issue of Jurisdiction*

A ship is viewed as quite a unique subject of the law, due to its ability to be subject to more than one system of law; international, national and customary maritime systems of law may all apply simultaneously.[50] It is this unique nature which creates numerous legal discussions, one of which focuses on the question of where jurisdiction over ships rests in relation to MARPOL.

A significant weakness of MARPOL is that the regulation is one that is voluntarily accepted by shipping states. States are responsible for implementing domestic legislation which gives effect to the rules agreed upon in the regulation; and this is also the case with enforcement. Although the IMO exists as the body responsible for

50 Daniel Patrick O'Connell, *The International Law of the Sea: Volume II* (Clarendon Press 1984) 747.

MARPOL, there are no powers vested within the IMO for implementation and enforcement of MARPOL. It is the flag, port, and coastal states, which are the relevant parties in the context of implementation and enforcement of MARPOL. Based on the categorization of the state, their jurisdiction and powers are regulated by customary maritime law as well as UNCLOS, which is a codification of customary international maritime law, currently ratified by 162 nations[51].

## *B. Development of Jurisdiction to enforce MARPOL*

The nationality of a ship is the starting point for determining where jurisdiction lies in relation to that ship and its crew. Historically, there were four connecting factors which were held to be relevant in identifying the nationality of a ship: the nationality of the ship-owning country, the state to which the ship was registered, the nationality of crew members, and finally the nationality of the master of the vessel. [52] The modern position has simplified the determination and provides that the nationality of the vessel is that of the state whose flag it flies under.[53]

Under the customary law of the sea, and affirmed by the Permanent Court of International Justice in the *Lotus Case* 1927, the flag state is the only one which has jurisdiction to enforce any regulations over ships while on the high (international) seas. Only once a ship voluntarily enters a port, may states other than the flag state attempt to enforce a regulation.[54] It is with the flag state that the majority of obligation lies.

A main source of criticism is flag states having the jurisdiction and the responsibility for enforcement over ships

---

51 United Nations, 'Oceans and Law of the Sea' <http://www.un.org/Depts/los/reference_files/chronological_lists_of_ratifications.htm> Accessed 1 May 2012.

52 O'Connell (n 50) 752.

53 ibid.

54 Birnie (n 1) 401.

flagged to the state as this causes a reduction in the efficiency of MARPOL.[55] The underlying reason for this is the lack of incentive for flag states to impose and enforce the pollution control rules diligently. UNCLOS Article 217(1) requires that the flag state ensures that the ships registered under its flag comply with all international rules and standards. Yet there does not exist, in any capacity, a means of review of flag state enforcement.

In addition, there are no penalties for flag states who fail to fulfil their MARPOL obligations.[56] While it may be a port or coastal state which detects a violation of MARPOL outside of their own territorial waters, those states are obligated to report the violation to the flag state who is then responsible for bringing proceedings against the offending ship.[57] Flag states tend to be averse to fulfilling the responsibility of prosecuting ships, and this dereliction of responsibility is owed largely to the advent of the flag of convenience.

### *C. Flag States*

The flag of convenience (FoC) is a practice which provides a significant impediment on the achievement of the MARPOL Annex I objectives. It is suggested that flag states lack incentive to enforce and are not subject to penalties for not doing so, which is the main cause for their failure to fulfil MARPOL responsibilities.[58]

The advent of the FoC has allowed the majority of the world's shipping tonnage to be registered with nations that ships would otherwise have no connection to, and

55 IMO (n 11) 203.

56 ibid.

57 Rebecca Becker, 'MARPOL 73/78: An Overview in International Environmental Enforcement' (1997) 10 Georgetown Int Environmental L Rev 631.

58 Richard J Payne, 'Flags of Convenience and Oil Pollution: A Threat to National Security?' (1980) 3 Houston J Int L 72.

possibly never even visit their ports.[59] Ship owners do not even have to visit the flag state to complete registration.[60] The Geneva Convention on the High Seas; Article 5: Section 1 proclaims that 'it is for each state to set the requirements for the granting of ships to fly its flag'.

The flag under which a ship is registered has a significant impact upon the operational costs of that ship. It is for this reason that shipping companies favour the FoC in the same way that multinational corporations base their manufacturing in developing nations, as it allows costs to be reduced and profits to be increased.[61] If a shipping company were interested in registering their vessel, for example, under the flag of the United States, that ship must be constructed in the US.[62] Labour costs, for example, under a Liberian FoC are estimated to be about 25% of that under the US flag. In addition, regulations for taxes and working hours are strictly enforced under a US flag.[63] The most important factor in relation to MARPOL is that the FoC state has in most cases, little power or interest to fulfil their commitments under international law. This allows shipping companies to operate pretty much as they wish on the high seas.[64]

In the majority of cases, port and coastal states detect MARPOL discharge violations. The standard procedure upon detection, as noted above, is to inform the flag state of the violation. Once reported by the port or coastal state, responsibility for prosecuting the vessel shifts to the flag state, which in the majority of cases is reluctant to prosecute.[65] In a study conducted by a Dutch environmental organization, related to violations within the North Sea, it was found that of

59 Mongolia for example, which is landlocked, allows ships to register under its flag.

60 Payne (n 58) 70.

61 ibid 69.

62 In most cases doubling the cost of construction.

63 Payne (n 58) 71.

64 ibid 72.

65 Becker (n 57) 631.

the violations reported to flag states,[66] only in 17% of the cases did the flag state investigate the matter via the process prescribed by the IMO and the MEPC. Only 6% of the total reported violations actually resulted in convictions and fines. Although fines were levied, they were generally held to be insignificant and unlikely to impact future conduct.[67] This aspect will be discussed further in more detail later.

Ultimately, it is the object of FoC states to benefit from the revenue generated by registering ships under their flag. This means that relying on the flag state to enforce MARPOL and take significant action in many instances will prove to be fruitless due to this arrangement.

### *D. Port States*

The jurisdiction of port states has improved since the introduction of MARPOL and was significantly improved further with the adoption of UNCLOS. Historically, port states only had jurisdiction to deal with violations which occurred within their territorial sea[68] or internal waters. If violations occurred outside of this area, then the port state could only inspect the ships documentation once it voluntarily entered into its port. If evidence of a violation was found, this had to be reported to the flag state.[69]

Since the adoption of UNCLOS III, the jurisdiction of port states has been enhanced under Article 218. Port states are able to prosecute foreign flagged ships for violations of internationally accepted regulations[70] that have occurred in international waters. If the violation has occurred within the jurisdiction of another state's coastal or Exclusive Economic Zone (EEZ) waters, the port state is only able to

66 Violations were a mix of FoC and non-FoC.

67 Ton Ijlstra, 'Enforcement of MARPOL: Deficient or Impossible?' (1989) Volume 20 Marine Pollution Bulletin 596.

68 The Territorial Sea extends 12 nautical miles from the baseline.

69 Tan (n 11) 217.

70 MARPOL is considered to be an internationally accepted regulation within Article 218 UNCLOS.

take action upon request by that state or the flag state (UNCLOS Article 218(2)).

Article 218 is tempered by Article 228 which provides flag states with the power of pre-emption in relation to violations that have occurred outside the territorial sea of the prosecuting state, and this must occur within 6 months of the start of proceedings. This pre-emption allows flag states to take action, yet it does not require a judgement against the alleged violator (UNCLOS III, Article 228). MARPOL Article 5(2) provides port states with the power to prevent unseaworthy ships from sailing until repairs have been made. This enhanced position of port state jurisdiction is an important improvement in achieving a greater level of compliance with MARPOL.

It must also be noted, that there remain a number of limitations in relation to the jurisdiction of port states. The port state is not obligated to take action and prosecute when informed of a violation by a coastal state. Once informed of a suspected violation, the port state is able to then report the violation to the flag state and avoid the cost involved in bringing proceedings against the violator, since the state prosecuting an alleged violation bears the cost of the legal proceedings.[71] Due to the significant financial costs involved in legal actions against alleged violators, it is common practice for a port state to choose to exercise the option of reporting to the flag state, rather than initiating proceedings under its enhanced powers.[72]

In an extensive study conducted by another Dutch environmental organization, it was found that in a period of seven years where 1335 violations were reported by port states to the IMO, 1077 were referred to the flag states for action, with only 238 being dealt with by the port state.[73] The existence of the enhanced port state jurisdiction only

71 Becker (n 57) 633.

72 ibid 632.

73 ibid.

resulted in slightly over 17% of cases reported by port states to the IMO to have resulted in judicial action by the port states within the study period. The violations reported to the IMO and referred to flag states resulted in only 5% of the alleged violations having any type of hearing or trial and just under 0.1% resulted in disciplinary action.[74]

It is evident that although there has been an enhancement in the jurisdiction of port states, the application of this increased optional power is not significant. In addition, there is no mandatory requirement for the port state to take any action. It is suggested that the reluctance of port states to act may be based on having to incur the cost of action against violators, as well as the impact it may have on the commerce of its ports if it gains a reputation for taking strict action against violators.[75]

Additionally, there are logistical challenges in high traffic ports that deal with thousands of vessels per year.[76] Port states are empowered to inspect ships to ensure that the flag state has issued an International Oil Pollution Prevention Certificate (IOPP). If a ship is in possession of an IOPP, then the port state, under MARPOL, must treat the certificate as if it had been issued by the inspecting port state (Annex I, Article V). The port state may only disregard the IOPP where there is clear evidence that the condition of the ship or its equipment does not correspond with that of the IOPP, and intervention is warranted.[77] The standard intervention is that the port state will not allow the ship to sail until repairs have taken place. It is important for a port state to have a clear understanding of what is significant enough to warrant intervention through the prevention of departure from a port, as opposed to being overzealous and causing

74 ibid 633.

75 Tan (n 11) 220.

76 ibid.

77 Andrew Griffin, 'MARPOL 73/78 and Vessel Pollution: A Glass Half Full or Half Empty?' (1994) 2 Indiana Journal of Global Legal Studies 501.

undue delay, which could result in the port state being liable to the ship owner.[78]

### *E. Coastal States*

The coastal state has traditionally been viewed as the innocent bystander who, through no fault of its own but by virtue of its geographical location, was significantly impacted by all types of oil discharges due to shipping.[79] The *Torrey Canyon* disaster was the event which focused attention on the deficiency of the customary maritime laws to address the impact of events which take place outside the territorial sea of a coastal state, yet still having a significant impact on that state.[80]

The central point of this issue has been the customary right of unimpeded free usage of the high seas, under Article 2 of the High Seas Convention 1958. Only within the 12 nautical mile territorial sea of a coastal state and within the internal waters, was there freedom to enforce national legislation (UNCLOS III Article 2). This legislation, the High Seas Convention may be more stringent than that of the commonly accepted international regulation, due to the right of national sovereignty.[81]

There are however, limitations within UNCLOS on this matter, one of which is that the coastal state cannot interfere with the right of innocent passage or international navigation. Coastal state jurisdiction is also excluded from matters related to the construction and infrastructure of vessels, areas where supremacy is given to MARPOL and SOLAS (UNCLOS III, Article 21(2); 211(4), Articles 17-19; 24-25).

Through the French and British reaction to *Torrey Canyon*, came the UNCLOS adoption of the rights of coastal

78 Birnie (n 1) 406.
79 M'Gonigle (n 16) 143.
80 ibid 146.
81 Birnie (n 1) 414.

states to the EEZ under Article 56. This is an area which extends beyond that of the territorial sea to a distance of 200 nautical miles from the baseline. The coastal states are given jurisdiction in relation to protection of the marine environment in this area (UNCLOS III, Article 56).

Birnie highlights that the EEZ is not an automatic jurisdiction but instead the coastal state must claim the jurisdiction in order to have it. To assume jurisdiction over matters of pollution, the accepted practice is to legislate on the matter domestically, and at that point jurisdiction will be assumed over the EEZ.[82] UNCLOS Article 211(5) only permits the enactment of laws over the EEZ that are in conformity with internationally accepted regulations, which in this case is MARPOL. This is stricter compared to UNCLOS Article 2, which applies only to territorial waters. There is one exception to Article 211 (5) and that is in application to the arctic and ice covered areas within the EEZ, which has been exercised by Canada.[83]

UNCLOS has extended the jurisdiction of coastal states to allow them to bring proceedings against ships which have violated MARPOL outside their territorial sea, provided that it is within the EEZ and only after they have entered that state's port (UNCLOS Article 220(1)). If a substantial violation which causes a significant threat of pollution has taken place within the EEZ, the coastal state is permitted to make a physical inspection of the vessel if it has refused to give required information or if the information provided is unreliable (UNCLOS Article 220(5)).

It is only in the situation described within Article 220(5) that the coastal state may arrest and detain a vessel that has entered into its territorial waters (UNCLOS Article 220(6)). If the violation is not deemed serious enough to warrant the above actions or entry into the territorial waters does not occur, then the coastal state must present the

82 Birnie (n 1) 418.
83 ibid 419.

evidence to either a port state visited by the vessel or to the flag state in the hope that the evidence will be sufficient for a prosecution to take place.

The difficulty with coastal state jurisdiction is that the state is limited to act only on significant violations occurring within their territorial waters, or upon those which have occurred in the EEZ where the offending vessel enters territorial waters. This is not a significant enforcement power due to its limitations, and if the coastal state is not also a port state, it is powerless to take judicial action directly against offenders.

Instead, it is required to rely on port or flag state prosecutions, something that has a minuscule chance of being successful. Although the EEZ is available, there are only a limited number of states who have legislated on EEZ pollution enforcement, and those who have legislated have done so in general terms or have not conformed with the requirements of UNCLOS.[84]

## IV. MARPOL And The Operational Challenges

The focus in this section is on the operational challenges faced in putting MARPOL regulations into practice. These challenges are largely dictated by financial complications. Foremost is the question of whether punitive fines have a deterrent impact on discharge violations. Consideration will also be given to the additional issues of the costs of enforcement by coastal states, as well as implementation by port states of reception facilities for oily waste from vessels.

### *A. Operational Issues in Executing MARPOL*

Discharge standards under the auspices of OILPOL failed to have an impact on operational oil discharges due generally to the lack of reliable monitoring equipment and

84 Tan (n 11) 214.

surveillance capability.[85] Compliance thus relied on the good faith and honesty of a ship's crew.[86]

The equipment requirements under MARPOL in the construction of new vessels and the retrofitting of older ones are suggested to have ensured more effective compliance with the regulations by ship owners than the discharge standards.[87] These equipment standards are able to be enforced by developed port states, as they have the resources, incentives and authority to ensure compliance.

Tan suggests that detention of vessels or denial of entry into ports for a blatant equipment or construction violation has a much greater deterrence on ship owners than does the possibility of a judicial fine for discharge violations.[88] This view on the effectiveness of the deterrence is based on the significant financial impact that detention or denial of entry will have immediately on ship owners.

### *B. Challenges of Enforcement and Deterrence: US Example*

The impact of financial penalties as a method of deterrence of illegal oil discharge from ships has been studied in the US. This study should be viewed as a realistic evaluation of the impact of these penalties, due to the level of enforcement exercised on behalf of the US. The United States Coast Guard (USCG) is responsible for prevention of damage to the marine environment through operational and accidental discharges and the USCG enforces the domestic legislation that implements MARPOL requirements in US territorial waters.[89]

---

85 M'Gonigle (n 16) 262.

86 ibid.

87 Tan (n 11) 237.

88 ibid 238.

89 Kishore Gawande, Alok K Bohara, 'Agency Problems in Law Enforcement: Theory and Application to the US Coast Guard' (2005) 51 Management Science 1595.

With approximately 30% of the annual operating budget dedicated to the Marine Inspections Program, the USCG is provided with approximately $3 Billion USD to fulfil this objective.[90] It is due to this substantial allocation of resources that the USCG should be considered a strong example of MARPOL enforcement and will be examined below.

Research conducted by analysing a number of studies beginning in the early 1980s through the late 1990s has found that the use of fines as deterrence in the US has only impacted operational discharges on a small scale. This is due to the fact that when issuing fines of $10,000 USD and under, only a limited amount of resources are required by the USCG and these cases can be disposed of rather quickly.[91]

In cases of larger discharges, it is suggested that there is a pattern of under penalization. This is due to the significant resources required for the USCG to issue fines in excess of $10,000 USD, or for these cases to be heard as a judicial civil penalty case. A case will be heard as a judicial civil penalty case where aggravating factors may be considered in order to assess a greater penalty than listed as standard for the discharge, but within the statutory maximum. In addition, the Oil Pollution Act 1990 imposes limits on the liability courts may impose on violators.[92]

In the period leading up to 2000, it was found that cases involving oil discharges regularly took more than a year to settle, and that the average penalty imposed by the USCG once a case had been settled was $3.96 USD per litre of oil discharged.[93] Additionally, there is reluctance on the part of the US to detain vessels for discharge violations that are

---

90 ibid.

91 ibid 1601.

92 ibid.

93 John M Weber, Robert E Crew, 'Deterrence Theory and Marine Oil Spills: Do Coast Guard civil Penalties Deter Pollution?' (2000) 58 Journal of Environmental Management 165.

considered minor (those below 5000 gallons). This lack of detention results in difficulty in the collection of fines for minor violations, especially if the ship does not return to a US port.[94]

The US example demonstrates how MARPOL enforcement even in a developed state with significant resources, legislation and motivation to enforce, is unable to overcome the fundamental difficulties connected to using financial penalties to deter illegal discharges.

### *C. Technological Difficulties of Monitoring Discharges and Collecting Reliable Evidence*

An early factor identified as a probable challenge for MARPOL was that of monitoring discharges from older ships which have not been the subject of strict construction requirements. The US National Academy of Sciences highlighted a lack of efficient monitoring as a deficiency in MARPOL. This shortcoming is viewed as a primary contributor to the difficulty in identifying the sources of oil discharges.[95] Tests conducted by the EEC demonstrated that discharges which fall within both the MARPOL special area and standard regulations were not detectable by the standard remote sensing equipment in use for the first decade of MARPOL.

It was concluded from those tests that discharges which were detectable were always above that of the standard limit, and observation by this method should be clear evidence of a violation of MARPOL.[96] Although it is possible to identify the existence of a discharge above that which is permitted, without the ability to take a sample it is near impossible to demonstrate to a court the exact discharge amount and it is thus unlikely for monitoring equipment

94 Mitchell (n 15) 451.
95 Nauke (n 4) 77.
96 Ijlstra (n 67) 597.

alone to provide sufficient evidence to court of a specific measurable violation.[97]

The difficulty of detecting discharge violations and the high cost involved in the collection of evidence through aerial surveillance, as well as the need to develop new technologies that are able to provide sufficient evidence which meets the evidentiary standards required by the judiciary are a significant factor which results in many nations being unwilling or unable to implement.[98]

The suggestion of Tan that equipment standards are the more reliable method of compliance is one which carries weight. Detention and barring from ports carries a very significant and immediate impact on ship owners. Total avoidance of compliance would result in a vessel being unable to trade in the majority of ports and thus significantly reduce its economic worth.[99] The ability for states to detain vessels creates significant and immediate financial penalties on ship owners due to "the cost of delays and lost trade.

There is no direct cost to the port state detaining a vessel. Detention can take place without the need for the drawn out process involving judicial hearings, which take place when taking action due to discharge violations. The evidentiary problems faced in dealing with discharge violations will also not arise.[100]

Although only a handful of MARPOL states have detained vessels by exercising this power, it does not mean that the equipment regime is not successful. The low number of detentions is considered evidence of a high degree of compliance. [101] Although compliance of equipment standards does not impose a cost on port states in relation to enforcement as the ship owners absorb

97 Becker (n 57) 637.
98 Mitchell (n 15) 454 .
99 ibid 451.
100 ibid 452.
101 ibid.

compliance costs, port states as members of MARPOL are required to have discharge reception facilities. This aspect of reception facilities as part of the equipment regime demonstrates yet another operational difficulty in putting MARPOL into practice.

### *D. Reception Facilities: An Unfulfilled Obligation*

It is stipulated by MARPOL that there should be reception facilities available in ports for wastes that cannot be discharged while at sea (Annex I, Regulation 12). The regulation placed the requirement for reception facilities to be available and operational one year from the entry into force of MARPOL Annex I. However, ten years on from the ratified deadline, many states still had not constructed reception facilities due to the cost involved as well as the fact that there is a lack of repercussions for non-compliance.[102]

In developing countries where the level of compliance with is lowest, the estimated cost to construct a reception facility starts at $500 million USD. This would require a level of investment which is impossible for these states to justify.[103] A survey assembled by the MEPC in 1990 based on reported findings from MARPOL member states to the IMO, found that of the 993 ports which were surveyed there were 104 which did not have any receptions facilities.[104]

Within the MARPOL special areas, where discharges are not permitted at all, and there is an increased need for receptions facilities, the survey found that 5.9% of the ports reported did not have the reception facilities required.[105] This has resulted in flag states electing not to enforce discharge standards for their ships operating in special area waters where the nations whose ports they called

102 Griffin (n 77) 505.
103 Tan (n 11) 267.
104 Mitchell (n 8) 196.
105 ibid 203.

upon had not complied with the reception facility requirement.[106]

The European Union, whose member states have all adopted MARPOL, has taken significant action in ensuring that reception facilities required within the convention are in place. This has occurred through EU Directive 2000/59/EC and due to the supremacy of EU law on the member states, they are obligated to comply with the directive and establish the necessary port reception facilities for the types of ships regularly calling at their ports.[107] Olson has noted that refinery terminals where vessels take on their cargo of oil are generally equipped with the necessary equipment to process oily water on a large scale.[108]

The presence of these facilities does not guarantee that they will be used, and this is due to the fact that some ports charge exorbitant reception fees or that the use of the facilities would cause a significant delay.[109] For these reasons, ships may be willing to take the risk of dumping oily water or waste into the sea due to the fact that even if they are caught there is a low chance of sanction, and in the limited case that a sanction is imposed it would be unlikely to have a major effect on the operation of the ship or its owner.[110]

It is evident that there is a circular relationship between the different operational aspects within MARPOL. This is demonstrated by the fact that it is necessary for the shipping industry to comply with construction and equipment requirements in order to operate in a manner where it is not necessary to discharge oil into the sea. There is a need for the MARPOL port states to ensure that reception facilities are in place and that they are operated in a manner which is not prohibitive financially in relation to the cost of use and

106 ibid 204.

107 Butt (n 2) 592.

108 Philip Olson, 'Handling of Waste in Ports' (1994) Volume 29 Marine Pollution Bulletin 289.

109 Tan (n 11) 256.

110 ibid 263.

the efficiency of that use. There must also be incentive to use the facilities, and significant deterrence to ensure that crews do not view the option of illegal discharge as more favourable due to insufficient enforcement and punishment. Thus the operational challenges are interconnected. As such, there is no single solution and it is not possible to point to a single shortcoming as the reason for unsatisfactory compliance with MARPOL.

The lack of available sufficient financial resources tends to be an underlying theme for unsatisfactory compliance with MARPOL. This is true for port, flag and coastal states, and a potential solution would be based on providing subsidization to the financially challenged states which would assist in providing the resources that result in improved compliance.

This potential solution creates its own issues, such as where would the funding come from, who would provide it, how would a state qualify, and is it possible to ensure the funding is spent on MARPOL related expenditures? Funding could also be provided to states as a reward for compliance, however once again the question arises as to how the states would find the financial resources necessary for compliance in the first place.

As these two simple examinations of possible solutions show, there is no single solution that can be applied without creating a host of new issues. This finding also supports the position that there is not a single shortcoming that can be addressed and result in total MARPOL Annex I compliance.

## V. Conclusion – Final Thoughts

There are a number of different factors that can be judged in determining whether MARPOL has been able to successfully achieve its Annex I objectives. Evidence presented here has proven that since the adoption of MARPOL discharge, construction and equipment regulations, there has been a decline in the amount of oil entering the sea from ships.

What remains to be seen is whether the reduction in vessel-source oil pollution is a direct result of MARPOL as an international regime driving change through compliance and behavioural changes. Breitmeier suggests that international regimes are established as a method of dealing with urgent transnational problems that occur in both the social and natural world.[111] In the case of vessel source oil pollution, the problems occur in one and affect the other. The success of MARPOL as an international regime thus can be judged on the questions of whether it is able to cause the individuals, companies and states involved to act positively towards alleviating the issue of vessel-source oil pollution.[112]

Action and compliance should not be driven by forced obligation but instead through improved knowledge and education about the issues and the problems that arise because of it. This knowledge will result in a willingness to be part of the solution rather than part of the problem.[113] The design of MARPOL as an international regime voluntarily adopted by maritime states suggests that a driving force behind acceptance should be viewed as improved understanding of the issue of vessel-source oil pollution, and the impact of discharges on coastal states.

The issues identified in the above sections, and related to compliance difficulties should be viewed not as an unwillingness of signatories to comply with MARPOL, but as the obstacles to compliance.

The most significant obstacle for states is that of financial resource availability. This is demonstrated, as discussed previously, through the difficulties faced by states in monitoring discharge violations at sea, the inability of flag states to take judicial action against offenders, as well as the reporting by coastal and port states back to flag states of alleged violations. In the USCG example considered above,

111 Helmut Breitmeier, *The Legitimacy of International Regimes* (Ashgate Publishing Limited 2008) 19.

112 Mitchell (n 15) 425.

113 Breitmeier (n 111).

it was found that actions taken are dictated by the cost of the resources required. Additionally, the underlying issue connected to the lack of total compliance with regard to reception facilities is driven by financial resources of both states and shipping companies.

It is significant to note that although financial factors impact compliance, the shipping companies have complied with equipment and construction standards more so than they have done with discharge standards, even though they are the far more expensive aspect with which to comply. The compliance of shipping companies with the construction and equipment standards allows for the prevention of violations by removing the possibility of them occurring, rather than working within a system of deterrence.[114]

Ultimately, there is no single solution to the challenges that exist in relation to the elimination of discharge violations. Discharge violations are not a black and white issue like that of construction and equipment, where either you are in compliance or you are not. Discharge violations follow more closely to the adversarial criminal law process, where there are many steps to get from the act at issue taking place, all the way to it being proven and then a punishment being imposed.

It is argued that international rules reflect the interests of the most powerful states. In shipping however, the most powerful states are those who have a significant number of vessels flagged to it, as well as the coastal states that export the majority of the world's crude oil, rather than the powerful western states who have been the victims of oil pollution.[115]

MARPOL, therefore, represents a legitimate international regime and though it has faced compliance challenges, there does appear to be the intention of the signatories to comply with MARPOL and to do their part to

114 Mitchell (n 15) 428.
115 ibid.

protect the marine environment from vessel-source oil pollution. MARPOL has not yet fully achieved this objective, yet it should be viewed as a successful international regime, for the reasons above, which has made a significant difference by empowering states to protect the marine environment and by putting in place a framework by which both global shipping and the marine environment can prosper without one suffering for the benefit of the other.

## BIBLIOGRAPHY

### Books

Birnie P, Boyle A and Redgwell C, *International Law and the Environment* (3rd Edition, OUP 2009)

Breitmeier H, *The Legitimacy of International Regimes* (Ashgate Publishing Limited 2008)

DeSombre ER, *Global Environmental Institutions* (Routledge 2006)

McEldowney J and McEldowney S, *Environmental Law* (Pearson Education Limited 2010)

—, *Environmental Law & Regulation* (Blackstone Press 2001)

International Maritime Organization, *MARPOL 73/78 Consolidated Edition, 2002* (IMO 2002)

M'Gonigle RM and Zacher MW, *Pollution, Politics and International Law: Tankers at Sea* (University of California Press 1979)

Mitchell RB, *Intentional Oil Pollution at Sea: Environmental Policy and Treaty Compliance* (The MIT Press 1994)

O'Connell DP, *The International Law of the Sea: Volume II* (Clarendon Press 1984)

Tan AK, *Vessel Source Marine Pollution: The Law and Politics of International Regulation* (Cambridge University Press 2006)

### Academic Journals

Ayorinde AA, 'Inconsistencies Between OPA '90 and MARPOL 73'78: What is the Effect on Legal Rights and Obligations of the United States and Other Parties to MARPOL 73/78?' (1994) 25 Journal of Maritime Law and Commerce 55

Becker R, 'MARPOL 73/78: An Overview in International Environmental Enforcement' (1997) 10 Georgetown International Environmental Law Review 625

Butt N, 'The Impact of Cruise Ship Generated Waste on Home Ports and Ports of Call: A Study of Southampton' (2007) 31 Marine Policy 591

Copeland C, *Cruise Ship Pollution: Background, Laws and Regulation, and Key Issues* (Congressional Research Service Report for Congress, 2008)

Curtis JB, 'Vessel-Source Oil Pollution and MARPOL 73/78: An International Success Story?' (1984) 15 Environmental Law 679

Dzidzornu DM, Tsamenyi BM, 'Enhancing International Control of Vessel-Source Oil Pollution Under the Law of the Sea Convention, 1982: A Reassessment' (1990) 10 University of Tasmania Law Review 269

Gawande K and Bohara AK, 'Agency Problems in Law Enforcement: Theory and Application to the US Coast Guard' (2005) 51 Management Science 1593

Griffin A, 'MARPOL 73/78 and Vessel Pollution: A Glass Half Full or Half Empty?' (1994) 2 Indiana Journal of Global Legal Studies 489

Ijlstra T, 'Enforcement of MARPOL: Deficient or Impossible?' (1989) Volume 20Marine Pollution Bulletin 596

Karim MS, 'Implementation of the MARPOL Convention in Developing Countries' (2010) 79 Nordic Journal of International Law 303

Kullenberg G, 'Approaches to Addressing the Problems of Pollution of the Marine Environment: An Overview' (1999) 42 Ocean and Coastal Management 999

Mitchell RB, 'Regime Design Matters: Intentional Oil Pollution and Treaty Compliance' (1994) 48 International Organization Foundation 425

Nauke M and Holland GL, 'The Role and Development of Global Marine Conventions: Two Case Histories' (1992) Volume 25 Marine Pollution Bulletin 74

Olson PH, 'Handling of Waste in Ports' (1994) Volume 29 Marine Pollution Bulletin 284

Payne RJ, 'Flags of Convenience and Oil Pollution: A Threat to National Security?' (1980) 3 Houston Journal of International Law 67

Peet G, 'The MARPOL Convention: Implementation and Effectiveness' (1992) 7 International Journal of Estuarine & Coastal Law 277

Weber JM, Crew RE, 'Deterrence Theory and Marine Oil Spills: Do Coast Guard civil Penalties Deter Pollution?' (2000) 58 Journal of Environmental Management 161

Wiswall FL, 'The Nature and Future of Maritime Law' (2004) 1 WMU Journal of Maritime Affairs 1

Xu J, 'Theoretical Framework of Economic Analysis of Law Governing Marine Pollution' (2006) 5 WMU Journal of Maritime Affairs 75

## Legislation

European Union Directive 2000/59/EC

Geneva Convention on the High Seas 1958

International Convention for the Prevention of Pollution of the Sea by Oil, 1954

International Convention for the Prevention of Pollution by Ships 1973 As Amended by the Protocol of 1978

United Nations Convention on the Law of the Sea III 1982

# Constitutional Reform and the Contribution of the Political Parties since the Beginning of the 20th Century

Richard Jones

## Abstract

*Some of the most significant reforms to the British constitution have occurred since the turn of the 20th Century, either through political and economic necessity, or through an unpressured desire to improve the system of fundamental laws on which the governing of the UK is based. This article delves into the various constitutional reforms brought about by the different political parties (Labour, The Conservatives, Liberals and Liberal Democrats) since 1900, discussing which have been the most significant. It must be stressed that this article focuses on the degree of impact that the reforms had, rather than their merits and whether they were beneficial for the country. The importance of this article has been to try and decipher which political party has been the most influential in shaping the constitution in recent times. In terms of methodology, the issue is tackled party by party, rather than chronologically, focusing mainly on their key reforms, and omitting some of the more minor ones. After reviewing relevant literature and documents such as books, academic articles, legislation and reports, I concluded that despite the importance of the New Labour changes, the single most significant constitutional reform in the period discussed was the Conservatives joining the EEC. The implication of the conclusions formed is that, ironically, the most significant constitutional reforms can be brought about by the most unlikely party, due to the pressures of the time.*

## I. Introduction

The primary focus of this article is to assess the roles of the key British political parties in constitutional reform, from the beginning of the 20th Century to present day. There will be a particular focus on critically evaluating which party (or parties) has crafted the most significant constitutional reform(s). It must be emphasised that this article will concentrate purely on the *significance* of constitutional changes, and not the merits or limitations of

the changes. I will not be delving into the benefits of Labour's 1999 House of Lords reform, for example, just the impact it had.

Defining constitutional reform can be quite difficult and consequently there is a lot of potential material to discuss, some of which may only be mentioned briefly, and some may not be mentioned at all due to restricted space. (Specific examples of developments that I will not mention include the Regency Act 1937, the Freedom of Information Act 2000, and signing the UN Charter.) Nevertheless, one definition of constitutional reform is: the introduction of legislation to modify 'the rules and practices that determine the composition and functions of the organs of central and local government in a state.'[1]

When analysing the main constitutional reforms across the period, they will be analysed party-by-party, dedicating a section to the Conservatives, Labour, and the Liberals (including the Liberal Democrats[2]). Without doubt, all three parties have brought about, or influenced, extremely significant reforms, but we must try and deduce the *most* significant. On initial reflection, the most noteworthy reforms in the 20th and 21st Centuries were perhaps the UK joining the European Economic Community (EEC) under Heath's Conservative government, and some of New Labour's constitutional reforms such as the Human Rights Act. However, in reality, it might be slightly optimistic to try and achieve a decisive conclusion on which party has played the most important role in constitutional reform, either through one event, or several.

---

1 Jonathan Law and Elizabeth Martin, *Oxford Dictionary of Law* (7th edn OUP 2009) 124.

2 Whilst they have a different title, they contain very similar values. The Liberals have also had much more influence in the Liberal Democrat direction than the Social Democrats.

## II. The Conservative Party

As described perfectly by Charmley, the traditional philosophy of the Conservatives is 'to conserve; it is the party of status quo.'[3] The Conservatives will only typically reform the constitution when necessary, and will usually not devise ambitious proposals, unlike the Liberals. However, despite being traditionally averse to constitutional change, the Conservatives over the past century have passed some highly significant pieces of constitutional legislation. Johnson writes of how the Conservatives have found themselves at times in the 'unusual role of protagonist of constitutional reform,'[4] suggesting they have played a significant role in reform somewhat unintentionally; with the exception of their relatively recent commitment to an elected House of Lords and Bill of Rights. One must concur, it does seem that any constitutional reform engineered by the Conservatives has occurred because of the circumstances of the time, rather than the party actively seeking reforms that are not completely necessary for national stability, but nonetheless beneficial (as the Liberal Democrats might). Even joining the EEC was for economic benefits, rather than a party desire for constitutional reform.

### *A. Joining the European Economic Community*

Nevertheless, an extremely important Conservative reform was the European Communities Act 1972, making Britain a member of the EEC, now the European Union (EU).[5] This was a momentous constitutional change. Britain had failed on two previous attempts to join the EEC, once in 1961-3 under Macmillan (Conservative), and once under

---

3 John Charmley, *A History of Conservative Politics, 1900-1996* (Macmillan Press 1996) 1.

4 Nevil Johnson, 'Constitutional Reform: Some Dilemmas for a Conservative Philosophy' in Zig Layton-Henry (ed), *Conservative Party Politics* (Macmillan Press 1980) 126.

5 For more information on Britain's history and membership in the EU see Anthony Bradley and Keith Ewing, *Constitutional and Administrative Law* (15th edn Pearson Education 2011) 117-143.

Wilson (Labour) in 1967, and was finally successful under Edward Heath, joining on 1 January 1973.

Community membership meant the UK was no longer in control of its own entity, having to answer to a more superior force, which completely reorganised the structure and hierarchy of our constitution. Lyon recognises the event as a 'major constitutional change,'[6] describing the 1972 Act as 'a piece of legislation which in the years since has caused enormous controversy and exercised a great many judicial and academic minds.'[7] This illustrates the sheer magnitude of the Act, being recognised as a key moment in constitutional history, attracting much debate. Lyon can be strongly agreed with; Britain's entry into the EEC is immediately recognisable as one of the landmark constitutional developments of recent times.

However, the reason for joining was not constitutional. The Conservative government (as well as Labour) were more interested in the economic and trading benefits of the EEC. So perhaps they do not deserve endless praise for this reform. Nonetheless, whether the constitutional impact was the intentional focus or not, it was still a remarkable development in constitutional law.

Regardless of the positives or negatives, joining the EEC had an incredibly significant impact on Parliamentary sovereignty, with the 1972 Act binding future Parliaments. Since 1973, the British Parliament has had to respect European Regulations, implement Directives, and ensure that domestic legislation does not conflict with European law, all because of one piece of legislation passed in 1972. A Diceyan view of Parliamentary sovereignty is that 'no person is recognised by the law of England as having a right to

---

6 Ann Lyon, *Constitutional History of the UK* (Cavendish Publishing 2003) 417; see also David Feldman, 'None, One or Several? Perspectives on the UK's Constitution(s)' [2005] CLJ 329, 345.

7 ibid 418.

override or set aside the legislation of Parliament.'[8] The Conservatives clearly undermined this fundamental principle in 1972 by giving such a right to the European institutions, demonstrating the sheer significance of the change.

William Wade accurately describes the effect the 1972 Act had on Parliamentary sovereignty as a 'constitutional revolution,'[9] highlighting its importance as a milestone in the history of British law. In relation to the Merchant Shipping Act 1988 and *Factortame* (mentioned later), Wade states that:

> The Parliament of 1972 had succeeded in binding the Parliament of 1988 and restricting its sovereignty, something that was supposed to be constitutionally impossible. It is obvious that sovereignty belongs to the Parliament of the day and that if it could be fettered by earlier legislation, the Parliament of the day would cease to be sovereign.[10]

Here, Wade suggests the 1988 Parliament had a key constitutional right taken away from them, exemplifying the significance of the 1972 Act, producing a restrictive knock-on effect for future Parliaments. Wade's opinion can be firmly endorsed, as this hindrance on legislative powers created a stranglehold over all future Parliaments, something which other constitutional reforms usually do not.

However, some academics argue, albeit rather weakly, that British EU membership is merely 'contingent upon'[11] the 1972 Act, and the restrictive effects of the Act are easily reversible, as it can be repealed like any other statute. Bradley believes this 'profound change in the operation of Parliamentary sovereignty is not necessarily permanent,

---

8 AV Dicey in Jeffrey Jowell & Dawn Oliver (eds), *The Changing Constitution* (OUP 2011) 53.

9 William Wade, 'Sovereignty – Revolution or Evolution?' [1996] LQR 568.

10 ibid 568.

11 F Nigel Forman, *Constitutional Change in the United Kingdom* (Routledge 2002) 351.

because the duty of British courts to apply EU law would not exist as a matter of UK law, but for the continued operation of the ECA 1972.'[12] It can be inferred that EU membership is not embedded in UK law, and any European obligations could easily be removed by repealing the 1972 Act. But in reality, I believe the 1972 Act is no ordinary statute, and 'was not subject to implied repeal.' [13] There is almost an unspoken understanding that the Act will not be revoked, as joining/leaving the EU is not something that can be constantly altered depending on the government of the day. It would also be extremely difficult to obtain the support of the majority in Parliament, as most moderate politicians believe that leaving the EU would be catastrophic. So any arguments devaluing the significance of the 1972 Act can be seen as flawed, as repealing the Act would be much easier said than done.

Another way the 1972 Act was constitutionally important was through creating the doctrine of Supremacy, ensuring European law has primacy over UK law. As Lord Denning stated, 'whenever there is any inconsistency, Community law has priority.'[14] He also stated that 'priority is given by our own law. It is given by the European Communities Act 1972 itself,' implying that the piece of Conservative legislation was the sole cause of EU law supremacy in the UK, highlighting the significance of the Act. Furthermore, Loveland believes EC membership has 'markedly affected traditional constitutional understandings,' resulting in a 'profound restructuring of the relationship between the courts, the executive and Parliament and the

12 Anthony Bradley, 'The Sovereignty of Parliament – Form or Substance?' in Jeffrey Jowell & Dawn Oliver (eds), *The Changing Constitution* (OUP 2011) 56; See also Trevor Allan, 'Parliamentary Sovereignty: Law, Politics and Revolution' [1997] LQR 443, 450.

13 Paul Craig, 'Britain in the European Union' in Jeffrey Jowell & Dawn Oliver (eds), *The Changing Constitution* (OUP 2011) 117.

14*Macarthys Ltd v Smith* [1981] QB 180, 200 (Lord Denning MR).

electorate,' [15] indicating that repercussions were felt in institutions other than just Parliament. This is an important point made by Loveland, as the 1972 Act affected the courts just as much as Parliament, as the judiciary have to oversee the enforcement of EU law supremacy. The constitutional impact of joining the EEC was undoubtedly widespread.

A key example of Community law supremacy created by the Conservatives was in *Factortame II*.[16] When the Merchant Shipping Act 1988 was found to be incompatible with EC law, the European law was given priority, and the British law subordinated,[17] meaning the 1988 Act was disapplied. This case provided solid confirmation of the significant and lasting constitutional impact of the 1972 Act.

Despite it being the Conservatives who made the final push for a successful application into the EEC, it must be noted that Wilson's Labour government made considerable efforts to join, with the 1967 application arguably only failing because of France and Charles de Gaulle's unreasonable veto. Therefore, joining Europe was not just a Conservative initiative; Labour also had a strong desire to bring about the same reform, meaning that the Conservatives perhaps do not deserve full credit. By the time of the UK's third application, de Gaulle was no longer the French President, and France was much more willing to welcome Britain into the EEC. In that sense, it could be argued, Heath was extremely lucky. Nevertheless, this does not draw attention from the fact that 'accession to the Community has proved by far the most significant constitutional innovation undertaken by any government in the 20th Century,'[18] as stated by Loveland. Concurring with

---

15 Ian Loveland, 'Britain and Europe' in Vernon Bogdanor (ed), *The British Constitution in the Twentieth Century* (OUP 2004).

16 *R v Secretary of State for Transport, Ex p Factortame Ltd (No 2)* [1991] 1 AC 603.

17 Wade (n 9) 568.

18 Loveland (n 15) 663.

Loveland, I believe the 1972 Act has been the *single* most important constitutional reform since the beginning of the 20th Century, as it provided for a considerable transformation of our political and legal system, ensuring that the British executive, judiciary and legislative now have an even greater power they must adhere to. Therefore, the Conservatives are strong contenders when considering which party has engineered the most significant constitutional change.

The Conservatives were also responsible for further European integration with Thatcher signing the Single European Act (SEA) 1986, and Major signing the Maastricht Treaty in 1992. Maastricht in particular was rather historic, creating the Euro currency,[19] and the pillar structure of the EU which meant further harmonisation in foreign/security policy, and justice/home affairs. The Conservatives felt compelled to sign these treaties to keep up with the developments of the EU. Evans writes of how 'the process of Europeanization has continued to mature as a structural response to the imperatives of the SEA (1986), and the Maastricht Treaty (1992),'[20] suggesting the treaties signed by the Conservatives had to have had a lasting effect on Britain's constitution and integration with the EU. Therefore, these tweaks in EU membership were of obvious importance.

### *B. The Abdication Act 1936*

Another Conservative constitutional statute was the Abdication Act 1936, taken as necessary action for Edward VIII's abdication. Baldwin's government passed the Act rather reluctantly, granting the King his wish to step-down from the throne to marry divorcee Wallis Simpson. Whilst it was a significant constitutional event at the time, it did not have a lasting effect for the future, and only brought about

19Although the UK opted out.

20 Mark Evans, *Constitution-making and the Labour Party* (Palgrave Macmillan 2003) 320.

reform in relation to the monarchy (rather than the executive/legislative/judiciary), arguably a mere symbolic aspect of our constitution. The statute did not even have a lasting impact on the monarchy, merely replacing one king with another. Moreover, the political parties were united in relation to the abdication crisis, so this event should not contribute too greatly towards any reputation the Conservatives have in reforming the British constitution.

### *C. Direct Rule of Northern Ireland*

The Northern Ireland Constitution Act 1973 was another reform not owing any particular merit towards the Conservatives, despite being a Conservative statute. The Act allowed for the direct rule of Northern Ireland from Westminster with the IRA/loyalist violence peaking between 1970 and 1972, and the Stormont government being unable to contain the security situation. The 1973 Act was merely a reactive piece of constitutional legislation that would have been passed out of necessity, regardless of who was in power.

### *D. House of Lords Peerage Reforms*

The Conservatives do however deserve credit for their House of Lords reforms in the shape of the Life Peerages Act 1958, which allowed for the creation of life peerages,[21] and for women to sit in the House;[22] and the Peerage Act 1963 which allowed females to inherit peerages,[23] and allowed heirs to hereditary peerages to disclaim their peerage.[24] Following the Parliament Acts, the 1958 Act in particular took the first big step in attempting to alter the composition of the House, laying the foundations for further reform in 1999. The aim of the 1958 Act was to reduce the number of part-time hereditary peers, introducing

21 The Life Peerages Act 1958 s1(1).
22 ibid s1(3).
23 The Peerage Act 1963 s 6.
24 ibid s1(1).

the more effective Life peers who specialize in specific political fields, and to achieve a fairer representation of Labour in the Lords, as the Conservatives accepted this would have to be addressed at some point.

Walters writes of how the 1958 Act meant that the 'hereditary mould [was] finally broken,'[25] implying the statute was key in modernising the House, which can be agreed with, as granting peers seats based on merit and ability, rather than through a genetic link (as had been the case for centuries), could only be seen as positive step forward, and thus, a significant constitutional reform by the Conservatives. Bogdanor makes the important point that the 1958 Act allowed the admission 'not only of party politicians but also of experts from all walks of life [...which] enabled the Lords to discover a new and valuable role for itself,'[26] suggesting a new era for the House had been created. In my view, to have experts in particular fields voicing their opinion in the Lords, rather than just hereditary peers, was a brave and crucial step forward, turning it into the modern-day institution that can scrutinise legislation more commendably. This was a key turning point for the Lords, which the Conservatives were responsible for.

However, Blackburn and Plant do negate the significance of the reform slightly when writing,

> this ostensibly modernising measure was in fact deeply reactionary: it served both to prolong the enfeeblement of the second chamber by deflecting rising criticism of the continuing appointment of hereditary peers, and to strengthen the premier's powers of political patronage.[27]

---

25 Rhodri Walters, 'The House of Lords' in Vernon Bogdanor (ed), *The British Constitution in the Twentieth Century* (OUP 2003) 198.

26 Vernon Bogdanor, *The New British Constitution* (Hart 2009) 155.

27 Robert Blackburn and Raymond Plant, *Constitutional Reform: The Labour Government's Constitutional Reform Agenda* (Addison Wesley Longman Ltd 1999) 24.

This implies the reform was perhaps for the gain of the government, avoiding the more substantial reform that was needed: *removal* of hereditary peers, as later achieved by Labour. Nevertheless, regardless of motives, this was still a significant reform. In my opinion, the 1958 Act, along with the European Communities Act, are the key reforms that must be considered when analysing the Conservatives' reform efforts, both of which were ground-breaking.

In relation to the 1963 Act however, the Conservatives should not receive all the credit, as the main reason it came into being was because of Labour's Tony Benn, who was protesting of his disqualification from the Commons. For this reason, a key influence in passing the statute was pressure applied by Labour, meaning the Conservatives cannot be given too much praise for its existence.

## III. The Labour Party

Labour have traditionally been more open to general reform than the Conservatives, but have only been proactive in *constitutional* reform quite recently,[28] having been rather ambivalent in the past.[29] For most of the 20th Century, Marquand believes that Labour saw 'constitutional arrangements [... as] frivolous diversions from the serious business of social and economic transformation,'[30] implying that they were relatively content with the existing format of the constitution. They only acquired an 'ostensibly greater commitment to constitutional reform since the Policy Review of the late 1980s,'[31] indicating that their position in the political wilderness forced them to rethink their attitude

28 Evans (n 20) 15.

29 Peter Dorey, *The Labour Party and Constitutional Reform: A History of Constitutional Conservatism* (Palgrave Macmillan 2008) 347.

30 David Marquand, 'Half-Way to Citizenship? The Labour Party and Constitutional Reform' in Martin J Smith and Joanna Spear (eds) *The Changing Labour Party* (Routledge 1992) 45.

31Dorey (n 29) 3.

towards constitutional reform. One can argue that the forced change was necessary in the modernisation of Labour, recognising the need to introduce exciting new policies to captivate the electorate.

The major constitutional reforms since New Labour came to power in 1997 'reshaped the UK's uncodified constitutional arrangements,'[32] and are the most significant *group* of constitutional reforms the UK has seen in a long time, and they have achieved in the shortest period of time possible.

### *A. The Parliament Act 1949*

However, one historic reform prior to this was the Parliament Act 1949, which built upon the 1911 Parliament Act in reducing the powers of the House of Lords, decreasing the time in which they can delay Bills from two years to one.[33] The primary motive to introduce the 1949 Act was to further cripple the Lords' powers, in the fear they would delay Labour's nationalisation programme, which Attlee wanted to complete within the life of the 1945 Parliament. Whilst the 1949 Act was not revolutionary in itself (unlike the 1911 Act, which passed the 1949 Act), it did cause a lot of constitutional debate and controversy, so it *is* significant in that sense. For example, it was used to pass the Hunting Act 2004, and the validity of both Acts were challenged in *Jackson v Attorney General*,[34] indicating that the Act had a great impact. The ruling that the 1949 Act was valid demonstrates its significance, as Attlee's government (along with Asquith's) have successfully bound future Parliaments, and the manner in which Bills are passed. It also demonstrates that the Commons are free to take action without being restricted by the Lords. Forsyth controversially suggests that the Commons could even alter s2(1) Parliament

32 Mark Ryan, 'The House of Lords and the Shaping of the Supreme Court' [2005] NILQ 135.

33 Section 1.

34 [2005] 3 WLR 733.

Act 1911 to remove the restriction on extending the life of Parliament, similar to the way the 1911 Act was altered by the 1949 Act.[35] It is extremely unlikely this would happen in reality, but if it did, the 1949 Act could be seen to have formed a highly significant precedent to follow.

Nonetheless, I must stress that the 1911 Act was far more revolutionary, taking the initial step. The 1949 Act merely made the Commons' stranglehold over the Lords slightly tighter. So as far as the Parliament Acts[36] go, the Liberals deserve much more acclaim.

### *B. European Convention on Human Rights*

Labour also ratified the European Convention on Human Rights in 1951, meaning the British legal system had to respect Convention rights to a certain extent, having a substantial impact on the constitution. Labour ratified the agreement because they had to acknowledge it at least on some level, as they were opting out of fully incorporating it into UK law. It was only in 1998, when the ECHR was finally implemented into British law by Blair, that Convention rights had a significant impact on the constitution. So Attlee's government do not deserve as much credit as New Labour in constitutionally recognising human rights, with the events of 1998 being far more significant than those in 1951.

### *C. New Labour's Reforms*

In turning our attention to New Labour, it must be stressed immediately that their collection of constitutional reforms were without doubt of pioneering importance. Labour introduced an entire shopping list of constitutional

35 Christopher Forsyth, 'The Definition of Parliament after Jackson: can the life of Parliament be Extended under the Parliament Acts 1911 and 1949?' [2011] IJCL 132, 143.

36 For more information on the Parliament Acts see Owen Hood Phillips, Paul Jackson and Patricia Leopold, *Constitutional and Administrative Law* (8th edn Sweet & Maxwell 2001) 168-172.

reforms in 1997, wanting to fulfil manifesto promises; promises that theoretically appealed to the masses by providing radical change in democracy (although realistically much electorate support was won through simpler factors, such as Blair's charisma). They wanted to contrast the lethargic constitutional policies of the Conservatives, by creating a reinvigorated constitution more representative of modern society.

However, the Liberal Democrats deserve some substantial credit for the reforms due to their input in the Labour-Liberal Democrat Joint Consultative Committee on Constitutional Reform; producing many shared ideas,[37] later implemented by Blair's government. This exemplifies how the Liberals often construct ambitious proposals for reform, but simply lack the means to implement them alone. It is extremely important that the Liberal Democrats are still recognised for their ideas and influence.

### *D. The Good Friday Agreement*

One of Blair's finest political and constitutional achievements was the Good Friday Peace Agreement and subsequent Northern Ireland devolution in 1998, [38] effectively resolving years of disagreement and violence, removing Westminster's direct rule that had existed since 1973. The aim was to achieve sustainable democracy in Northern Ireland, where opposing sides could cooperate and share power. Forman writes of how in 1997 there were 'new opportunities for resolving the Northern Ireland problem – opportunities which Tony Blair seized with both hands,'[39] resulting in the Good Friday Agreement, signed in an

37 For more details on the Labour-Liberal Committee see Roy Douglas, *Liberals: A History of the Liberal and Liberal Democrat Parties* (Hambledon & London 2005) 306-307; and Labour-Liberal Constitutional Committee, 'Report of the Joint Consultative Committee on Constitutional Reform' (1997).

38 For more detail on the devolution process in Northern Ireland see Colin Knox, *Devolution and the Governance of Northern Ireland* (Manchester University Press 2010) 1-46.

39 Forman (n 11) 70.

'atmosphere of exhaustion and euphoria.' This conveys a sense of initiative on behalf of Labour, taking brave and positive steps in a difficult constitutional area. A sense of 'euphoria' portrays the agreement as a momentous occasion, which it was. However, whilst Blair does deserve much credit, we must not forget that John Major also played a crucial role in the build up to a peace agreement, meaning this cannot be labelled an outright Labour achievement. Nevertheless, I believe Blair still made an outstanding contribution in this pivotal constitutional development, playing a vital role in negotiations between the two sides.

Despite the Northern Ireland Act 1998 including a seemingly significant provision, allowing Northern Ireland to leave the UK with the 'consent of the majority,'[40] this consent principle that would allow the Northern Irish to leave through a referendum was actually present in section 1 of the Northern Ireland Constitution Act 1973, drafted by the Conservatives, and to an extent in s1(2) of the Ireland Act 1949. So the 1998 Act was not so revolutionary in this aspect. Moreover, Northern Ireland had already experienced a devolved government between 1922 and 1972 anyway, so the 1998 Act again provided for nothing new, yet it was something very different to what the 1998 population were accustomed to. The agreement also arguably focused more on fixing political relations, than the constitutional element of devolution. Nonetheless, it still resolved a 30 year disturbance of peace, which should not be discredited.

## *E. Devolution*

Another key constitutional reform imposed by New Labour was general devolution, [41] creating the Scottish Parliament,[42] Welsh Assembly[43] and London Assembly,[44] as

40 s 1(1).

41 For more detail see Dorey (n 29) 203-347.

42 Scotland Act 1998.

43 Government of Wales Act 1998.

well as the Northern Ireland Assembly. Labour were keen to recognise the various national identities and cultures within the UK, awarding them an appropriate amount of independence. Forming such institutions was hugely significant, as it was the first time the whole of the UK was not directly ruled by Westminster since 1707. History was truly being made by Labour. Despite not being able to pass laws in some specific areas, such as foreign affairs, devolved institutions were given considerable legislative freedom, for example in education and health-care. Granting such competence was a remarkable forfeit of some of Westminster's powers, and a vital step towards Blair's vision of a 'more democratic, decentralised and plural state.'[45] Plus, even though the Welsh Assembly was not granted primary legislation powers at first, their ability has gradually enhanced following the Government of Wales Act 2006,[46] and the 2011 referendum.[47]

Bogdanor writes of how each of the home nations, as part of 'the new constitution,'[48] now have their 'own identity and institutions – a multi-national state rather than [...] a homogeneous British nation containing a variety of people.'[49] To address such an error that had gone unrecognised for several hundred years was an important historic achievement, ensuring that Welsh, Irish and Scottish values are properly represented in Britain, fixing an 'outmoded constitution,'[50] as O'Neill puts it. However, I would not go as far as stating it to be 'the biggest

---

44 Greater London Authority Act 1999 – due to space restrictions, I cannot discuss this further.

45 Tony Blair's speech to the Welsh Assembly, October 2001 – Forman (n 11) 39.

46 Creating Assembly Measures, s 97.

47 Creating Acts of Assembly, s 107 Government of Wales Act 2006.

48 Bogdanor (n 26) 89.

49 ibid 116.

50 Michael O'Neill, *Devolution and British Politics* (Pearson Education Ltd 2004) 171.

constitutional change since 1707,'[51] as Berhard Bort believes. That label should most probably be awarded to joining the EEC.

The use of referendums when trying to achieve devolution was also of great constitutional significance. As explained by Deacon, 'There was the possibility that a future Conservative government would abolish the devolved bodies in Scotland and Wales if they were not endorsed by referendum.'[52] This implies that Labour went one step further, safeguarding devolution through referendums, which in a sense embedded these new institutions into our constitution. Binding future Parliaments in such a manner contributed considerably towards the sheer enormity of this Labour reform.

Despite the huge significance of devolution, it would have been much more historic had Labour created federalism similar to in the US, or perhaps granted Scotland or Wales independence. On top of this, Westminster still maintains overall power, and with s28(7) Scotland Act, can override any decision made by Scottish Parliament. This means, according to Leyland, 'the supreme law-making capacity of Westminster remains intact,'[53] demonstrating that the subsection was included to deliberately ensure that devolution was not too significant, and did not cross a certain threshold. This vitiates devolution's significance, perhaps demonstrating that Labour were not quite as bold and brave as it initially appeared. Having said that, section 28(7) could possibly be seen as more of a technicality, having never been used without permission from the Scottish government.

Leyland also believes a weakness in devolution is that 'both the purse strings and sovereignty remain in the hands

---

51 Eberhard Bort, 'The New Institutions: an Interim Assessment' in Michael O'Neill (ed), *Devolution and British Politics* (Pearson Education Ltd 2004) 295.

52 Russell Deacon, *Devolution in Britain Today* (2nd edn Manchester University Press 2006) 4.

53 Peter Leyland, 'Devolution, the British Constitution and the Distribution of Power' [2002] NILQ 408, 413.

of Westminster,'[54] indicating that the UK government has held onto ultimate control in many ways. Additionally, Batey points out that there has been a substantial continuation of Westminster legislating in Scotland, writing that 'it was widely assumed that Westminster would cease to legislate in the devolved areas. The evidence shows this has not happened.'[55] She believes there are still many statutes passed that have UK-wide effect, and perhaps should not have, such as the Criminal Justice and Court Services Act 2000. This is an important point she makes, similar to those by Leyland, but I feel devolution must be assessed in relation to what it *did* do, rather than failed to do. Labour could have provided devolved bodies with more powers, yes, but what *was* achieved was extremely significant, regardless of any powers held back by Westminster.

A further important point made by Leyland is that 'devolution has been a dynamic process which has triggered further important constitutional changes.'[56] One possible and very significant ramification of devolution, particularly in Scotland, is that it could eventually trigger federalism or maybe even complete independence. Having being given some freedom, it is possible the Scottish will now want more and more, and devolution may have been the first substantial step towards an independent Scotland. Without the 1998 Act, it would not have been possible for the SNP and Alex Salmond to hold a referendum on independence in 2014, suggesting that Labour's actions may have had wider, more significant implications, than initially thought. Such an implication would be contrary to Blair's intentions, as stated in the White Paper *Scotland's Parliament*, which dedicates its

54 ibid 412; see also Peter Leyland, 'The Multifaceted Constitutional Dynamics of UK Devolution' [2011] IJCL 251, 254.

55 Andrea Batey, 'Scotland's other Parliament: Westminster Legislation about Devolved matters in Scotland since Devolution' [2002] PL 501, 522.

56 Leyland, 'Multifaceted Constitutional Dynamics' (n 54) 251.

focus towards 'legislative devolution,'[57] stressing that any 'policy of independence being implemented in the near future' is unlikely. Nevertheless, this unintended consequence may be a possibility, even if only a slight possibility. Scottish independence would be an exceptionally important moment in constitutional history if it took place, but it is uncertain which party would be responsible. It could be Labour for the trigger of devolution, the Coalition government for allowing it to happen, or purely the SNP for their determination and persistence. But there is no doubt that Labour's devolution would have played a vital role.[58]

### *F. The Human Rights Act*

Furthermore, another immensely important constitutional reform engineered by Labour was the Human Rights Act (HRA) 1998.[59] It was exceptionally significant because it finally incorporated the ECHR into British law,[60] triggering an institutional focus on rights, freedoms and liberties that had never been felt before in our constitution; something Labour were enthusiastic to fully recognise and consolidate. It had a 'momentous'[61] impact on the way government and Parliament can legislate, having to ensure laws are compatible with the Convention. There was also an influx of human rights cases appearing before the judiciary, encouraging citizens to protect liberties that they previously only had limited protection for, or who would have faced the daunting task of taking their case to Strasbourg. Human rights now find their way into cases and legal argument when

---

57 The Constitution Unit, *Scotland's Parliament: Fundamentals for a New Scotland Act* (Cmd 3658, 1997) para 3.

58Iain Jamieson, 'Playing Politics with the Law? Scottish Parliament's Power to Legislate for a Referendum on Independence' [2012] SLT 61.

59 For further details on the 1998 Act and how it can be used, see Dawn Oliver, *Constitutional Reform in the UK* (OUP 2003) 112-119.

60 See The Home Department, *Rights Brought Home: The Human Rights Bill* (Cmd 3782, 1997) paras 1, 14-1, 19.

61 Lammy Betten, *The Human Rights Act 1998: What it Means* (Kluwer Law International 1999) 12.

they previously would not have. The HRA also encouraged public bodies to abide by the convention.[62] This is the most important, and therefore significant, constitutional reform Labour has introduced, primarily because it has affected most, if not all, areas of law. Any attempt at a British Bill of Rights[63] by the Conservatives will be a mere modification of what was achieved by Labour, who took the ambitious and more important first step.

Bogdanor seemingly shares the same opinion on the HRA's significance, labelling it 'the key to our liberties'[64] and 'the cornerstone of the new constitution.' It can be inferred from this that the Act forms the foundations of our 21st Century legal system, for which Bogdanor can be strongly agreed with.[65] Bellamy also makes an important point about s.3 of the HRA, writing that 'read as convention compatible goes against the view that no Act of Parliament can bind later Parliaments.'[66] This implies that Parliamentary sovereignty was undermined by Labour when passing the HRA, so it can be deemed very significant indeed. I also agree with Starmer who believes section 3 can be a 'radical tool,'[67] as interpreting a statute as far as possible in line with Convention rights could alter a case's outcome completely; viewing an Act in an almost entirely different way from its 'natural meaning,'[68] as s3 provides judges with much discretion.

---

62 s 6.

63Commission on a Bill of Rights, 'Do we need a UK Bill of Rights?' (2011).

64 Bogdanor (n 26) 88.

65 Roger Smith similarly labels the HRA 'a pillar of an evolving constitution' – Roger Smith, 'Human Rights and the UK Constitution: can Parliament Legislate "Irrespective of the Human Rights Act"?' [2006] LIM 274, 280.

66 Richard Bellamy, 'Political Constitutionalism and the Human Rights Act' [2011] IJCL 86, 102.

67 Keir Starmer and Francesca Klug, 'Incorporation through the "Front Door": the First Year of the Human Rights Act' [2001] PL 654, 664.

68 Philip Sales and Richard Ekins, 'Consistent Interpretation of the Human Rights Act 1998' [2011] LQR 217, 222.

Despite the HRA's unquestionable importance, it must be noted that most human rights were already protected through common law prior to 2000, and in many ways, the HRA was just a formalisation of those rights. However, some of those rights were not given as much judicial protection as they have post-2000,[69] meaning the HRA should still be recognised as very significant.

The constitutional significance of the HRA can also be questioned in the sense that courts cannot strike down incompatible legislation, they can merely make a declaration of incompatibility,[70] which Parliament are entitled to ignore. So human rights are not as strictly protected as they could be, and the Act has not affected Parliamentary supremacy[71] as much as it could have. Parliament is still free from judicial control. In support of this, Wadham writes of how the HRA 'protects the principle of Parliamentary sovereignty because it does not permit the Convention to be used so as to override primary legislation.'[72] Wadham can be agreed with here. It appears the HRA was deliberately constructed so its constitutional impact was not too invasive of Parliamentary sovereignty. The HRA is less directly and explicitly binding on future Parliaments, than the European Communities Act, for example.

However, in reality, despite the lack of strike-down power, the s4 declaration of incompatibility is still a powerful 'weapon'[73] and it is very unlikely that Parliament would ignore such a declaration. Section 4 still provided the judiciary with a significant new influence over Parliament, which must be acknowledged.

Roger Smith sums up the capabilities of s4 perfectly: 'ministers retain the legal power to legislate irrespective of the

---

69 Smith (n 65) 276.

70 s 4.

71 Lyon (n 6) 453.

72 John Wadham, Helen Mountfield, Elizabeth Prochaska and Christopher Brown, *Blackstone's Guide to the Human Rights Act 1998* (6th edn OUP Oxford 2011) 9.

73 Bogdanor (n 26) 64.

HRA but, in fact, their political powers are somewhat contained.'[74] This suggests that s4 provides an implied understanding that Parliament will respect the views of the judiciary and take action following a s4 declaration, meaning Parliamentary sovereignty is implicitly undermined. Whilst it is technically possible for a government to defy a s4 declaration, it would ordinarily be 'politically inexpedient,'[75] meaning it would only be ignored in exceptional circumstances.[76] So the immense political and public pressure means that the government is almost compelled to address the incompatible statute. This reiterates the constitutional significance of the HRA.

Sales and Ekins believe this pressure means that the HRA 'has created a system which is closer to a constitution in which courts have the power to strike down legislation than is often supposed.'[77] Therefore, it is partly the *indirect* repercussions of the HRA which make it so significant. I agree. Section 4 is much more powerful than it *prima facie* appears, as political pressure plays a highly influential role.

Another sign of the HRA's significance has been the vast amount of important human rights cases since 2000. A fine example is *A and Others v Secretary of State for the Home Department*,[78] where a s4 declaration was made against s23 (detention without trial of foreign nationals) Anti-terrorism, Crime and Security Act 2001 due to its discriminatory nature. This use of the HRA by the courts led to the eventual replacement of this provision with non-discriminatory Control Orders[79] in the Prevention of Terrorism Act 2005. This demonstrates the important

74 Smith (n 65) 279.
75 Bellamy (n 66) 101.
76 Sales and Ekins (n 68) 230.
77 ibid.
78 [2005] 2 AC 68.
79 s 1.

impact the HRA has had on the courts, making the judiciary directly involved in legislative law-making.

One way the HRA could have been more significant would be if it was entrenched into the constitution, similarly to the US Bill of Rights, which would have been a greater contravention of Parliamentary sovereignty.[80] The HRA can be repealed at any time, as David Cameron intends to, replacing it with a British Bill of Rights, having set up a Commission to introduce this.[81] Therefore, the constitutional significance of the HRA is by no means long-term or permanent. But if it is repealed, it will undoubtedly be replaced with something similar, so Labour's 1998 Act will still have a lasting effect regardless of what the future brings, having instigated this greater recognition of human rights. I believe that, on the whole, it is very difficult to doubt the constitutional impact of the Human Rights Act.

### *G. Reforming the Membership of the House of Lords*

Another major reform that Labour was responsible for was the House of Lords Act 1999, which involved a drastic overhaul of the Lords' membership.[82] This meant removing most hereditary peers, followed by the introduction of mainly life peers, which produced a Labour majority for the first time.[83] A key aim was to defeat the overwhelming Conservative majority that had existed in the Lords for centuries. Bogdanor writes of how the 1999 Act 'transformed the upper house,'[84] suggesting the alteration to be quite revolutionary. Producing such a 'markedly different

80 Alison Young, *Parliamentary Sovereignty and the Human Rights Act* (Hart Publishing 2009) 1.

81 Ministry of Justice, 'Commission on a Bill of Rights' (29 March 2012) <http://www.justice.gov.uk/about/cbr> accessed 20 April 2012.

82 For more detail, see Dorey (n 30) 123-140.

83 Walters (n 25) 233.

84 Bogdanor (n 26) 157.

composition'[85] to the membership of the Lords was an extremely courageous step taken by Labour.

However, some might argue that passing the 1999 Act was perhaps easier than it could have been, because despite there being opposition, the Lords reluctantly agreed to pass it. So because of the ease in which the Act was passed, it can be deemed slightly less of an achievement by Labour. The Act was passed with no delay due to the compromise made between Blair and the Lords, allowing 92 hereditary peers to remain. But this compromise also makes this reform seem less significant, only partially completing what it set out to do. I believe such compromise shows weakness on Labour's behalf, but the removal of hundreds of hereditary peers was still a very dramatic reform nonetheless.

In concordance with the 1999 Act, Labour vowed to carry out a second stage of Lords reform;[86] transforming it into a primarily elected chamber. This second stage was not attempted by Labour, suggesting they only completed half the reform that they set out to achieve. So when considering this larger picture, the 1999 Act seems less significant, as it was only one step in an incomplete master-plan. Walters suggests that all the parties have shown laziness towards Lords reform, Labour included. He writes of how the 1999 Act was 'an easier option to comprehensive change,'[87] implying that the most important reform of the Lords, to make it democratically elected, was avoided. Dorey also believes Labour had 'kicked House of Lords reform into the constitutional long grass.'[88] This indicates that despite their efforts, Labour were unwilling to carry out what would have been an even more spectacular reform. A complete reform of the Lords has been longed for since 1911, and Labour, like the Conservatives, have not committed to going the full

85 Forman (n 11) 211.

86 The Labour Party, 'New Labour: New life for Britain' (1997) 29.

87 Walters (n 25) 232.

88 Peter Dorey, '1949, 1969, 1999: The Labour Party and the House of Lords Reform' [2006] Parl Aff 599.

nine yards; merely taking a partial step to appease those demanding full reform. Nonetheless, what Labour did achieve should not be discredited; the 1999 Act was still one of the most significant constitutional reforms of the 20th Century.

Despite the Conservatives' Life Peerages Act 1958 providing necessary tools for Labour to carry out the 1999 Act (as mentioned previously), I feel Labour's reform of the Lords' membership was much more significant, drastically altering the composition of the Lords in a mass exodus, rather than just providing a means for slight improvement. However, if the Coalition government successfully pass the House of Lords Reform Draft Bill,[89] then such a reform, including a partly elected House (through STV proportional representation), would perhaps overshadow previous reforms.

### *H. The Constitutional Reform Act 2005*

A further Labour reform to be discussed was the Constitutional Reform Act 2005, which aimed to achieve a more definitive separation of powers between the judiciary and legislators.[90] A key provision was to ensure the independence of the Lord Chancellor[91] from the judiciary and House of Lords, taking the new role of Secretary of State for Justice. This provision was only a minor constitutional reform, simply shifting certain responsibilities to different positions.

However, Part 3, which created a UK Supreme Court was much more historic. After centuries of the House of Lords being the highest court in the land, it is now the Supreme Court, independent of the Lords, which some feel was a significant step. But I must argue that this reform was rather cosmetic, merely changing the title and location of the

89 HM Government, *House of Lords Reform Draft Bill* (Cmd 8077, 2011).

90 For more information on the 2005 Act see Lord Mance, 'Constitutional Reforms, the Supreme Court and the Law Lords' [2006] CJQ 155.

91 Part 2.

highest court. Malleson points out that the new court does not have 'greater authority or a higher status,'[92] and that it is 'a change in form rather than substance,'[93] being the same as the Appellate Committee in the House of Lords which it replaced. Malleson can be whole-heartedly agreed with. This reform had no deep impact on the constitution, as the court operates exactly the same as prior to 2009, and possesses no greater constitutional powers.[94]

The Supreme Court title is also rather misleading,[95] as the UK court does not have the same strike-down powers as other Supreme Courts, such as in the US and Canada. Malleson believes the UK Supreme Court 'does not comply with the generally recognised prerequisites of a constitutional court.'[96] Lord Woolf also feels the UK court 'would be a poor relation among the Supreme Courts of the world'[97] with no strike-down powers, suggesting the introduction of this inferior Supreme Court to be of little importance. This again reiterates that the 2005 Act was merely a superficial alteration, and should not be considered a significant part of Labour's constitutional reform accomplishments.

## *I. Gordon Brown's Reforms*

Finally, some slightly less important reforms introduced by Brown's government were the Parliamentary Standards Act 2009 and Constitutional Reform and Governance Act 2010. The former was used by Labour to introduce the Independent Parliamentary Standards Authority to regulate MP expenses,[98] and the latter granting

92 Kate Malleson, 'The Evolving Role of the Supreme Court' [2011] PL 754, 765.

93 ibid 754.

94With the exception of acquiring the power to resolve disputes over devolution legislation from the Judicial Committee of the Privy Council.

95 Ryan (n 32) 158.

96 Malleson (n 92) 757.

97 Clare Dyer and Patrick Wintour, 'Woolf Leads Judges' Attack on Ministers' *The Guardian* (London, 4 March 2004).

98 s 3.

the civil service statutory recognition for the first time,[99] and requiring any new treaty signed by Britain to be ratified by Parliament.[100] Both Acts introduced relatively important changes, but neither can be considered one of the most significant constitutional developments since 1900.

## IV. Liberal and Liberal Democrat Party

Traditional Liberal Party philosophy imposes a 'distinctive'[101] commitment towards constitutional reform, having historically provided an ambitious alternative to the Conservatives' passive attitude. It has been customary for the Liberals to focus on constitutional matters, rather than the socio-economic issues like the Conservatives; purposefully planning long-term constitutional reform, instead of reforming the constitution out of forced necessity, due to the circumstances of the time, as has arguably been the case with the Conservatives.

The Liberals are extremely ambitious, consistently seeking a radically new constitutional order,[102] and have longed for a codified constitution, for example. However, I feel they are perhaps only so ambitious because they have been the third party since the 1920s, and until 2010, have had no realistic opportunity to implement such radical ideas. The Liberals perhaps try to achieve electoral support by focusing on a political area that the main parties have less time to focus on, using it almost as a "wild-card." Bogdanor writes of how in constitutional matters, Labour and Conservative policies are 'marked by a cautious and sceptical pragmatism, while the Liberal Party has adopted a holistic and utopian approach entirely at variance with the politics of

99 Part 1.

100 Part 2.

101 Vernon Bogdanor, 'Liberal Party and Constitutional Reform' in Vernon Bogdanor (ed), *Liberal Party Politics* (OUP 1983) 173.

102 Rodney Brazier, *Constitutional Reform: Reshaping the British Political System* (3rd edn OUP 2008) 1.

gradualism.'[103] I believe that such fearless ambition towards reform is largely influenced by the Liberal Party's usual[104] inability to legislate their proposals.

Having been the most committed party towards constitutional reform, the Liberals have been extremely influential, but have not received the credit they deserve, rarely having the means to put their initiatives into effect.

### *A. The Parliament Act 1911*

One of the most significant constitutional reforms achieved by the Liberals was the Parliament Act 1911, marking a 'fundamental change' [105] in British politics. Following frustrations in failing to pass a finance Bill, the 1911 Act was passed by Asquith's government as a means of preventing the Lords from vetoing Bills, awarding them the ability to *delay* Bills only. As undoubtedly one of the most prominent reforms of the 20th Century, the 1911 Act radically altered the balance of power between the Houses, and how Parliament legislates.

Walters believes the effect of the 1911 Act was 'profound,'[106] creating 'an assertion of the primacy of the Commons,' meaning that 'a chamber of veto was forced to reinvent itself as a chamber of scrutiny.' This signifies the Act to be of remarkably importance, which can be agreed with, as I believe it symbolised the first official reduction of the Lords' power. However, I do feel Ridley's claim that 'the Unionist [Conservative] view of a bicameral legislative was finally defeated,' [107] is overly exuberant, as the 1911 Act merely curbed the Lords' powers, rather than completely eradicating the Upper House.

---

103 Bogdanor (n 101) 187.

104 When not in a coalition.

105 Bogdanor (n 26) 149.

106 Walters (n 25) 192.

107 Jane Ridley, 'The Unionist Opposition and the House of Lords, 1906-1910' [1991] Parliamentary History 253.

The Act also improved democracy, providing for greater respect of the electorate's wishes by giving more power to the elected Commons. Weill writes of how the Lords could no longer 'coerce an election,'[108] with the 1911 Act creating 'popular sovereignty by which the people's voice in constitutional matters was retained.'[109] This conveys a sense that the Liberals helped instil greater legitimacy into the constitution, making their reform highly commendable.

However, the significance of the Act can be questioned slightly. Firstly, the 1911 Act has only been used on seven occasions across an entire century, so it is not a reform that affects our constitution regularly. Ekins also argues that the 1911 Act 'does not seek to redefine Parliament,'[110] which is true. It merely curbs certain powers and amends the way legislation is passed. Bogdanor also believes the Lords were still left with 'considerable powers.'[111] I strongly agree; the ability to delay a Bill by two years should not be underestimated.

Nevertheless, the 1911 Act was still highly significant in what it *did* do, and the role it played in instigating further future reforms. However, a century later, another Liberal, Nick Clegg, is still looking to achieve an elected 'popular' Second Chamber with the House of Lords Reform Bill. Whether the Coalition can finally achieve the aims set out by Asquith, only time will tell.

## *B. Voting Reform*

A significant reform under Lloyd George was the Representation of the People Act 1918. However, despite the government at the time having a Liberal leader, it was a Coalition heavily populated by Conservatives, meaning both

---

108 Rivka Weill, 'Centennial to the Parliament Act 1911: the Manner and Form Fallacy' [2012] PL 105, 116.

109 ibid 118.

110 Richard Ekins, 'Acts of Parliament and the Parliament Acts' [2007] LQR 91, 106.

111 Bogdanor (n 26) 151.

parties deserve credit. The Act allowed all men to vote in elections (regardless of property status), and also granted women over 30 the right to vote, subject to certain property requirements. This Act was introduced after the Great War because it was felt that certain men who had fought a war to protect British democracy now deserved the vote, regardless of property status. With a persuasive input from Suffragettes and Suffragists, the government also felt it necessary to award women the right to vote, following their valiant contribution towards the war effort.

The 1918 Act was an unforgettable legislative achievement, [112] and significant modernisation of our constitution, making it more democratic and representative, allowing for a greater number of citizens to cast their opinion on who should run the country. Blackburn writes of how the Act 'laid the foundations for the country's present-day voting and electoral system'[113] and 'was a symbolic measure of immense significance to the constitution,'[114] illustrating it as a milestone in constitutional history. Blackburn can be strongly agreed with; granting women the right to vote was one of the most memorable advancements in democracy in modern times.

However, the government do not deserve too much credit as the 1918 Act was rather consequential of the times, with the First World War and the Suffragette/Suffragist movements being the most important contributions, rather than unprompted initiatives of the Liberals and Conservatives. Hobbs believes the Act was the outcome of 'the greatest political caucus of modern times,'[115] implying that the primary influence was the women's rights movement, not a long-term Liberal/Conservative objective. However, as

---

112 Arthur Hobbs, *A Guide to the Representation of the People Act, 1918* (Butterworth 1918) 1.

113 Robert Blackburn, 'Laying the Foundations of the Modern Voting System: The Representation of the People Act 1918' [2011] Parliamentary History 33.

114 ibid 47.

115 Hobbs (n 112) 1.

is often the case, I believe it was a combination of the two; the government still played an important role. Blackburn also writes of how Lloyd George deserves a 'great deal of credit'[116] due to 'strong leadership' when passing the Act, indicating that the government *did* deserve some recognition. He also describes the Act as 'the only true liberal achievement of Lloyd George's premiership,' implying that the 1918 Act was particularly influenced by Liberal members of the Coalition, emphasising the need to grant Liberals *some* recognition, despite a Conservative dominance in government.

In 2011, Nick Clegg attempted to achieve another Liberal reform of the voting system with the AV referendum,[117] but failed. Had the public voted 'Yes,' I would have denoted his accomplishment as an extremely significant constitutional reform, but I am clearly unable to make such a statement.

### *C. Irish Independence*

Another significant reform introduced by the 1916-22 Coalition was granting the Republic of Ireland independence from the UK, and forming Northern Ireland. This was achieved through the Government of Ireland Act 1920 and the Irish Free State Constitution Act 1922. Such action was reactionary to the Irish War of Independence 1919-21, and the agreed ceasefire.[118]

The reformation of Britain's physical constitution and structure in such a drastic manner was extremely significant indeed; nothing short of a constitutional revolution. Bogdanor also describes the decision to keep

---

116 Blackburn (n 113) 47.

117 Parliamentary Voting System and Constituencies Act 2011.

118For more detailed information on the build up to the Governance of Ireland Act 1920 see HT Dickinson and Michael Lynch, *The Challenge to Westminster: Sovereignty, Devolution and Independence* (Tuckwell Press 2000) 102-111; and Francis Costello, *The Irish Revolution and its Aftermath 1916-1923: Years of Revolt* (Irish Academic Press 2003).

Northern Ireland excluded from the South as a 'crucial decision,'[119] implying that the action taken had important ramifications for the future, meaning acknowledgement of the Coalition's efforts is due, regardless of whether the impact was positive or negative.

However, it must once again be stressed that the emergency circumstances of the Irish situation played a crucial role, and it is doubtful that independence would have been granted had the uprising not occurred. Additionally, any governmental credit can once again be shared between the Liberals and Conservatives.

## V. The Current Coalition Government

As this article was intended to assess constitutional reform retrospectively up to present day, I will only consider the current Coalition government, and their future plans, very briefly. As mentioned previously, the introduction of a British Bill of Rights will be a noteworthy reform, but in most regards it will merely be a cosmetic modification of the HRA.

If the House of Lords Reform Bill is successfully passed, introducing a partially-elected Second Chamber, it will be of exceptional significance, achieving something that has been avoided for a century.

One reform the Coalition has already achieved is the Fixed-term Parliaments Act 2011, providing for fixed elections every five years. This guarantee will provide the Coalition the best amount of time possible to complete their intended legislative programme. Such a change will have a noticeable impact, as it means the Prime Minister cannot tactically select a general election date.

## VI. Conclusion

Despite New Labour introducing a vast catalogue of significant constitutional reforms in a short time period, I must conclude that the single-most recognisable reform since

119 Vernon Bogdanor, *Devolution in the United Kingdom* (OUP 1999) 65.

the beginning of the 20th Century was the UK entry into the EEC by the Conservatives. The impact of the European Communities Act on our constitution has been colossal, having a profound effect on Parliamentary sovereignty and the foundations of law-making, completely reshaping the basic democratic structure of the UK.

Labour must nevertheless receive credit for the sheer number of reforms they engineered under Blair. So too must the Liberals for their influential attitude towards constitutional matters, making important contributions, for example, in the Labour-Liberal Committee prior to the 1997 election. In my view, the Conservatives seemingly introduce constitutional reforms through necessity of the times (with the recent exceptions of the Bill of Rights and Lords reform plans), whereas the Liberals and New Labour have a genuine ambition to improve our constitution for purely constitutional reasons. It is of great irony that the party most supportive of constitutional reform, the Liberal (Democrat) Party, has been the least influential and the party that has played the most important role, the Conservative Party, is traditionally averse to large-scale change. But that is the nature of politics.

## BIBLIOGRAPHY

### Journal Articles

Allan T, 'Parliamentary Sovereignty: Law, Politics and Revolution' [1997] LQR 443

Batey A, 'Scotland's other Parliament: Westminster Legislation about Devolved matters in Scotland since Devolution' [2002] PL 501

Bellamy R, 'Political Constitutionalism and the Human Rights Act' [2011] IJCL 86

Blackburn R, 'Laying the Foundations of the Modern Voting System: The Representation of the People Act 1918' [2011] Parliamentary History 33

Dorey P, '1949, 1969, 1999: The Labour Party and the House of Lords Reform' [2006] Parl Aff 599

Dyer C & Wintour P, 'Woolf Leads Judges' Attack on Ministers' *The Guardian* (London, 4 March 2004)

Ekins R, 'Acts of Parliament and the Parliament Acts' [2007] LQR 91

Feldman D, 'None, One or Several? Perspectives on the UK's Constitution(s)' [2005] CLJ 329

Forsyth C, 'The Definition of Parliament after Jackson: can the life of Parliament be Extended under the Parliament Acts 1911 and 1949?' [2011] IJCL 132

Jamieson I, 'Playing Politics with the Law? Scottish Parliament's Power to Legislate for a Referendum on Independence' [2012] SLT 61

Leyland P, 'Devolution, the British Constitution and the Distribution of Power' [2002] NILQ 408

Leyland P, 'The Multifaceted Constitutional Dynamics of UK Devolution' [2011] IJCL 251

Malleson K, 'The Evolving Role of the Supreme Court' [2011] PL 754

Lord Mance, 'Constitutional Reforms, the Supreme Court and the Law Lords' [2006] CJQ 155

Ridley J, 'The Unionist Opposition and the House of Lords, 1906-1910' [1991] Parliamentary History 253

Ryan M, 'The House of Lords and the Shaping of the Supreme Court' [2005] NILQ 135
Sales P & Ekins R, 'Consistent Interpretation of the Human Rights Act 1998' [2011] LQR 217
Smith R, 'Human Rights and the UK Constitution: can Parliament Legislate "Irrespective of the Human Rights Act"?' [2006] LIM 274
Starmer K, Klug F, 'Incorporation through the "Front Door": the First Year of the Human Rights Act' [2001] PL 654
Wade W, 'Sovereignty – Revolution or Evolution?' [1996] LQR 568
Weill R, 'Centennial to the Parliament Act 1911: the Manner and Form Fallacy' [2012] PL 105

## Books

Allison J, *The English Historical Constitution: Continuity, Change and European Effects* (Cambridge University Press 2007)
Betten L, *The Human Rights Act 1998: What it Means* (Kluwer Law International 1999)
Blackburn R & Plant R, *Constitutional Reform: The Labour Government's Constitutional Reform Agenda* (Addison Wesley Longman Ltd 1999)
Bogdanor V (ed), *Liberal Party Politics* (Oxford University Press 1983)
—— *Devolution in the United Kingdom* (Oxford University Press 1999)
—— (ed), *The British Constitution in the Twentieth Century* (Oxford University Press 2003)
—— *The New British Constitution* (Hart Publishing 2009)
Brazier R, *Constitutional Reform: Reshaping the British Political System* (3rd edn Oxford University Press 2008)
Charmley J, *A History of Conservative Politics, 1900-1996* (Macmillan Press 1996)
Costello F, *The Irish Revolution and its Aftermath 1916-1923: Years of Revolt* (Irish Academic Press 2003)

Deacon R, *Devolution in Britain Today* (2nd edn Manchester University Press 2006)
Dickinson HT & Lynch M, *The Challenge to Westminster: Sovereignty, Devolution and Independence* (Tuckwell Press 2000)
Dorey P, *The Labour Party and Constitutional Reform: A History of Constitutional Conservatism* (Palgrave Macmillan 2008)
Douglas R, *Liberals: A History of the Liberal and Liberal Democrat Parties* (Hambledon & London 2005)
Evans M, *Constitution-making and the Labour Party* (Palgrave Macmillan 2003)
Forman FN, *Constitutional Change in the United Kingdom* (Routledge 2002)
Hobbs A, *A Guide to the Representation of the People Act, 1918* (Butterworth 1918)
O Hood Phillips, P Jackson & P Leopold, *Constitutional and Administrative Law* (8th edn Sweet & Maxwell 2001)
Jowell J & Oliver D, *The Changing Constitution* (Oxford University Press 2011)
Knox C, *Devolution and the Governance of Northern Ireland* (Manchester University Press 2010)
Law J & Martin E, *Oxford Dictionary of Law* (7th edn Oxford University Press 2009)
Layton-Henry Z (ed), *Conservative Party Politics* (Macmillan Press 1980)
Lyon A, *Constitutional History of the UK* (Cavendish Publishing Ltd 2003)
Oliver D, *Constitutional Reform in the UK* (Oxford University Press 2003)
O'Neill M, *Devolution and British Politics* (Pearson Education Ltd 2004)
Purvis J & Houlton S, *Votes for Women* (Routledge 2000)
Smith MJ & Spear J, (eds) *The Changing Labour Party* (Routledge 1992)
Wadham J, *Blackstone's Guide to the Human Rights Act 1998* (6th edn Oxford University Press 2011)

Young A, *Parliamentary Sovereignty and the Human Rights Act* (Hart Publishing 2009)

## Cases

*A and Others v Secretary of State for the Home Department* [2005] 2 AC 68
*Jackson v Attorney General* [2005] 3 WLR 733
*Macarthys Ltd v Smith* [1981] QB 180
*R v Secretary of State for Transport, Ex p Factortame Ltd (No 2)* [1991] 1 AC 603

## Command Papers

The Constitution Unit, *Scotland's Parliament: Fundamentals for a New Scotland Act* (Cmd 3658, 1997)
HM Government, *House of Lords Reform Draft Bill* (Cmd 8077, 2011)
The Home Department, *Rights Brought Home: The Human Rights Bill* (Cmd 3782, 1997)

## Statutes

Abdication Act 1936
Anti-terrorism, Crime and Security Act 2001
Charter of the United Nations 1945
Constitutional Reform Act 2005
Constitutional Reform and Governance Act 2010
Criminal Justice and Court Services Act 2000
European Communities Act 1972
European Convention on Human Rights
Fixed-term Parliaments Act 2011
Freedom of Information Act 2000
Government of Ireland Act 1920
Government of Wales Act 1998
Government of Wales Act 2006
Greater London Authority Act 1999
House of Lords Act 1999
Human Rights Act 1998
Hunting Act 2004
Ireland Act 1949

Irish Free State and Constitution Act 1922
Life Peerages Act 1958
Maastricht Treaty 1992
Merchant Shipping Act 1988
Northern Ireland Act 1998
Northern Ireland Constitution Act 1973
Parliament Act 1911
Parliament Act 1949
Parliamentary Standards Act 2009
Parliamentary Voting System and Constituencies Act 2011
Peerage Act 1963
Prevention of Terrorism Act 2005
Regency Act 1937
Representation of the People Act 1918
Scotland Act 1998
Single European Act 1986

## Reports and other Official Documents

Commission on a Bill of Rights, 'Do we need a UK Bill of Rights?' (2011)
The Labour Party, 'New Labour: New life for Britain' (1997)
Liberal-Labour Constitutional Committee, 'Report of the Joint Consultative Committee on Constitutional Reform' (1997)

## Websites

Ministry of Justice, 'Commission on a Bill of Rights' (29 March 2012) <http://www.justice.gov.uk/about/cbr> accessed 20 April 2012

# A Radical Interpretation of Individual Self-Defence in War

Tanzil Chowdhury

## Abstract

*The general prohibition on the use of force as expounded by the UN Charter identifies individual self-defence as one of the exceptions to this proscribed rule. Yet self-defence has developed ideologically in the past 60 years in a manner that has arguably undermined the spirit of the UN Charter. Notions of pre-emptive and anticipatory self-defence have been formulated to justify a reading of article 51 which allows for the use of force before any armed attack has occurred. The objective of this article is to observe the legal, moral and strategic underpinnings of the varying schools of thought and address their shortcomings. In doing so, I shall look in detail at the applied ethics of killing in self-defence. This analysis will involve the so-called 'domestic-analogy' which compares self-defence between states to self-defence between persons. Having looked at the various theories, I shall explain my own interpretation on the law of self-defence by using a radically different methodology. This will be premised on the notion of how our perceptions of violence alter when we receive new information about the incident in question. I shall also apply these theoretical interpretations to some case studies. It shall then become clear that the current law is symptomatic of iniquities with respect to international relations which observe hegemonic states regularly abusing the doctrine of self-defence. The conclusion shall then illustrate how the new interpretation aims to question the distortion of these bedrock principles upon which more powerful states rely, and to re-evaluate the legality and morality (or lack thereof) of their actions in the international arena.*

## I. Introduction

Manifestations of self-defence invoke the antiquated 'chicken and egg' conundrum - which came first? Quarrelling states either apportion blame to one another or rationalise their conduct. Whether it is contextualising the attack as pre-emptive or condemning this initial attack as aggression, it is easy to become immersed in a quagmire.

Antediluvian conceptions, derived from the canonical texts and the classical literature often imbue the purist *lex taliones* inclination, more commonly known as the eye-for-an-eye paradigm. Question, does not the aggressor no matter how brutal his actions are, have an 'inalienable right to life'[1] or is that forfeited once he has struck the first blow? Indeed one may ask, need he strike a blow? Surely a 'threat' of a blow is sufficient. So why is it necessary we worry ourselves with questions of moral reasoning on issues pertaining to law?[2]

The law of self-defence, specifically individual self-defence in the laws of war shall be the focus of discussion. The primary document, the United Nations Charter[3], codifies the law (albeit in frustrating generality as we shall later discover) to restrict the use of force between nations. It came after the horrors of two World Wars and the atrocities of the Nazi holocaust; perhaps one of the most horrific acts ever to blemish the tapestry of time. Political decolonization swept the world with varying degrees of success and a new Pan-European Belle Epoque promoting diplomacy, liberal democracy and human rights set the agenda for international relations. Whereas war had previously been commonplace, cited in various literary works on the *bellum justum,* the UN Charter appeared to adopt a more 'restrictionist approach', making war the exception rather than the rule. Indeed self defence, whether individual or collective, were the only caveats within the general prohibition on the threat or use of force against the territorial integrity and political independence of any state[4]. However, this premature optimism is quelled when one historically considers the number of wars and hostilities – since the inception of the

1 David Rodin, *War and Self-Defence* (OUP 2002) 50.

2 Indeed the nexus between law and morality (or lack thereof) forms the basis for much jurisprudence literature- see also Ian McLeod, *Legal Theory* (5th edn. Palgrave 2010) and Brian Bix, *Jurisprudence: Theory and Context* (6th edn. Carolina Academic Press 2012).

3 The Charter of the United Nations adopted 26 June 1945 articles 2(4) & 51.

4 ibid.

Charter. Weltman puts it well when he cynically says war has been present three times more frequently than it has been absent.[5] Schachter echoes similar sentiments when he states that reality seems to mock them [UN Charter].[6] Yet a contrasting discourse, a caveat, emerges and it seems to present war with the peaceful use of military[7] (force) (referring to the inherent right as enshrined in customary international law) as a conception that is far more dynamic.[8] Something seems askew here; the literature often treats invocations of self-defence with contempt yet as much scholarship cites its instrumental use as a means to apparently thwart belligerent states. Diversity of views is one thing, but polarity is quiet another. Are we in danger of retreating into the Pre-Grotian era in which war was a state's prerogative power? Or are we already there?

Why is there such a diversity of interpretation of the infamous self-defence enshrined in article 51? Surely law is law? Indeed, such is the uncomforting realisation of many students new to public international law that quashes their undeveloped 'Austinian' conception of law as rigid legal rules issued by a sovereign and backed by a threat. These questions create a *lacuna* in the law; a demand for the elusive virtue of clarity in this area. The impetus therefore is driven by the inadequacy of the status quo and a need to prevent abuse of this inadequacy.

The paper will begin by critically assessing the various approaches and applications of article 51. This includes a brief look at the history, impetus and ideology behind the notion of individual self-defence and explaining

---

5 John Weltman, *World Politics and the Evolution of War* (The John Higgins University Press 1995) 1.

6 Oscar Schachter, 'The Right of States to use Armed Force' (1984) 82 Michigan L Rev 1620.

7 Robert J Art and Kenneth N Waltz , *The Use of Force: Military Power and International Politics* (6th edn Rowman and Littlefield Publishing Group 2004) 3.

8 Tarcisio Gazzini, *The Changing rules on the use of force in International Law* (Manchester University Press 2005) 123.

why the popular notions of anticipatory and pre-emptive self-defence (terms which I will use interchangeably) garnered much support as time progressed. Two heterodox schools of thought exist, each with varying opinions within them[9] that discuss the meaning of the provision in very different terms. Analysis and criticism will therefore be the concern. The focus shall be on the imminence of threats as this is often the most contentious issue; therefore there will be little discussion on the intrinsic requirements of necessity, proportionality or the deliberation over distinguishing between reprisals and armed attacks.

Following on from this descriptive and diachronic analysis, I shall turn my focus abruptly from a critical mode to a creative one. I will attempt to cultivate a radically new interpretation of self-defence using history as a basis for this concept. In much the same way anticipatory and pre-emptive self-defence determine the imminence of a threat by previous acts of the belligerent, I will use historical events as an excusatory and justificatory basis: firstly, for understanding the use of violence; and secondly for apportioning charges of aggression elsewhere. This focuses on a very different methodology to that employed by many of the contemporary scholarships based on how our perceptions of violence radically alter, when we are exposed to new information concerning certain events. This requires a thorough critique of different approaches to killing in self-defence before moving-on to elaborate my own position.

Having formulated this interpretation (in addition to acknowledging some of its potential misgivings) I will apply it to a few sample cases. Observations as to whether or not strong and weak states can successfully use this defence will determine its success and I shall account for any potential extra-interferences with my results.

The conclusion discusses how this new doctrine would affect current knowledge concerning international

9 Christine Gray, *International Law and the Use of Force* (3rd edn Oxford University Press 2008) 117.

relations. This considers questions of how certain states fighting for self-determination or against external aggression maybe able to use this form of defence, not exclusively to correct past injustice, but rather a way to help us understand their conduct in the wider framework of international relations. It will also provide some insight about state hegemony in the context of international law.

## II. Counter-restrictionist vs. Restrictionist

Let us recall article 2 (4) and article 51 of the United Nations Charter respectively:

> All Members shall refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state, or in any other manner inconsistent with the Purposes of the United Nations.
>
> Nothing in the present Charter shall impair the inherent right of individual or collective self-defence if an armed attack occurs against a Member of the United Nations, until the Security Council has taken measures necessary to maintain international peace and security.[10]

Clemmons and Brown believe self-defence is a powerful and necessary concept.[11] Indeed, the interest stimulated scholars from Plato to Cicero and continues to form the topics of heated debate in contemporary international relations. Its importance however, has been temporally relative. During the development of the just war doctrine, the resort to war was still unlimited and remained a state prerogative. Self-defence was of little consideration as states were the final arbiters in determining their right to engage in war.[12] These ideas were symptomatic of a realist

10 UN Charter (n 4).

11 Commander Byard Q Clemmons and Major Gary D Brown, 'Rethinking International Self-Defence: The UN emerging role' (1988) 45 Naval L Rev 217.

12 Gazzini (n 8) 123.

conception of international anarchy[13] in which order was established through demonstration of power. The first noteworthy discussion of self-defence was the Caroline Incident in 1837. During the Mackenzie Rebellion against British rule in Canada, a steamboat from the sympathetic US, provided men and materials to a rebel-held island. The colonial rulers therefore set fire to the boat killing several American seamen.[14] Britain pleaded self-defence but US Secretary of State Daniel Webster proposed that such an invocation would only be realistic if necessity is instant, overwhelming, leaving no choice of means, and no moment for deliberation. Webster's now canonised analysis earmarked a new appreciation for a concept once considered as peripheral. It also laid the basis for the ensuing law of self-defence in customary international law which, as we shall see, would be the envy of the textual literalists. Some endorsed Webster's view whilst others question its academic merit and practical application.[15]

Various charters such as the Chapultepec Treaty and the Kellogg-Briand Pact had been created to renounce war as a measure of national policy and a means to solve inter-state conflicts.[16] However, neither had the far-reaching applicability, nor the controversy such as that in the UN Charter. The Charter appeared to adopt a near-ban on the use of force[17] with an effort to substitute law for force[18] placing great emphasis on the notion of sovereignty and sovereign equality. However, several issues of indeterminacy emerged from the rules; the standards of necessity and

13 Leo Van Den Hole, 'Anticipatory Self-defence under International Law' (2003) 19 American University International L Rev 70.

14 Malcolm Shaw, *International Law* (6th edn Cambridge University Press 2008) 1131.

15 Yoram Dinstein, *War, Aggression and Self-defence* (4th edn Cambridge University Press 2005) 249.

16 Van Den Hole (n 13) 71.

17 UN Charter (n 3).

18 China Miéville, *Between Equal Rights: A Marxist Theory of International Law* (Pluto Press 2006) 286.

proportionality, interpretations of key words in the text; not to mention differing perceptions of events.[19] It is here that our two diverging schools of thought emanate.

To use Arend and Beck's terminology, we have the restrictionists and the counter-restrictionists.[20] Unfortunately, even within these differing schools of thought, we have neither a holistic nor a homogenous opinion. The former assumes the position that article 51 is absolute, from which there is no derogation. Indeed, it has been likened to a jus cogens principle[21] and signatory states may only invoke the defence in response to an actual armed attack. The latter consists of a hybrid argument which looks at the article in the backdrop of the political and military realities that have developed since 1945 and, most importantly, the influence of customary international law. I shall firstly focus on the restrictionists.

A focus on this school of thought requires reference to the counter-school of thought; an exercise of endorsement through critique in that the restrictionist school is partially given legitimacy through abrogating the counter-restrictionists - arguably a lesser of two evils. Brownlie condemns the classical law and its anachronistic custom, that being the pillar of strength in the counter-restrictionists.[22] He suggested that treaty law aimed to clarify and tidy-up developments in international law between 1920 and 1930[23] with the later UN Charter being the apotheosis of such law.

In addition to this, adherents would say that any case of anticipatory self-defence would require a *lex scripta* more

---

19 Oscar Schachter, *International Law in Theory and Practice* (Martinus Nijhoff Publishers 1991) 141.

20 Anthony Clark Arend and Robert J Beck, *International Law and the Use of Force* (Routledge 1993) 73.

21 Gazzini (n 8) 122.

22 Ian Brownlie, 'The Use of Force in Self-Defence' (1961) 37 British Year Book of International Law 184.

23 ibid 197.

vividly worded that just armed attack.[24] One of the fallacies, and perhaps why Brownlie is so faithful to this perspective, is that applications of this defence were entirely subjective and as a result, he contested it could lead to absurd results. If we are to take Brownlie's conception of article 51, then identifying an armed attack would be easily and objectively determinable.

Dinstein, although affirming these views, placed himself as a conduit between the two opposing schools. His rejection of the naturalists approach (more akin to the counter-restrictionists) and adoption of a positivist mode dispels the right of self-defence as inherent. This he attributes as an anachronistic residue when international law was dominated by ecclesiastical doctrines. [25] But his view does not dismiss the customary right altogether. In contrast to Brownlie, he addresses both items of customary international law and the UN Charter distinctly. In this way he adopts a very strict approach to the text, but he acknowledges a very complex qualification of the customary international law right. Dinstein accepts that self-defence maybe invoked if an aggressor state embarks upon a course of irreversible action.[26]

This restrictionist method was acquiesced by the majority judgments in The Republic of Nicaragua v. United States.[27] Although the case was very critical of US conduct in providing logistical support for the Contras, the judges were reluctant to affirm any particular position [28] stating that 'attributions and assertions of self-defence is a political question which no court, including the ICJ should judge.'[29] However, dissenting Judge Schwebel did provide vociferous

24 Van Den Hole (n 13) 84.
25 Dinstein (n 15) 180.
26 ibid 191.
27 1984 ICJ Reports.
28 Gray (n 9) 130.
29 Dinstein (n 15) 212.

judicial opposition. He sympathised with the counter-restrictionists examining the potential for incongruous results should a state have to wait for an actual armed attack to respond. He denied that treaty law abrogated the 'inherent right' under customary international law.

The judges perhaps did provide a ray of scholarly light on an otherwise dark and unclear rule (on the scope of armed attacks and their gravity). The judgement focussed on the scale and effects of an attack[30] by distinguishing between the gravest forms of the use of force and other less grave forms.[31] The Iranian Oil Platforms[32] case echoed similar views when expanding the interpretation of pre-emptive strikes. In considering whether the aggressor had to be a state or whether blame could be apportioned to non-state actors, the court held the latter as amendable provided that the state's involvement was clear.[33] This was subject to some criticism by scholars suggesting that the threshold had been set too low, but this was affirmed by the ICJ and their reliance upon article 3 (g) General Assembly Resolution on the definition of aggression.[34]

The Nicaragua case [35] suggests an exhaustive approach to article 51. However, this is far from convincing given the lack of a clear and pronounced judgement. The silence rather tacitly ratifies the literal approach. One of the interesting points is that the judges did consider article 51 as part of customary international law. The problem is that this blurs the lines between the naturalist law conception of self-defence and the positivist approach. On the one hand their opinion about whether an armed attack needs be actual is unclear and, yet they seem to endorse its heritage in

30 Rodin (n 1) 114.

31 Shaw (n 14) 1133.

32 *Islamic Republic of Iran vs United States of America* ICJ Reports 2003.

33 Gray (n 9) 130.

34 Definition of Aggression annexed to General Assembly Resolution 3314 <http://jurist.law.pitt.edu/3314.htm> accessed 15 May 2010.

35 Nicaragua (n 27).

customary international law. These contrasting positions create a 'legal oxymoron' by entrenching the ideas portrayed firmly in opposing fields of thought. Perhaps the case law is not the best source of clarity.

Amongst many others, one of the main reasons adherents like Brownlie thought it was important to have a strict approach rather than a lithe interpretation, was to prevent mistake and fraudulent claims. How could states determine if an attack was imminent? The presupposition was that states were purely altruistic. The worry therefore also came from a fear that states could use self-defence as a 'carte blanche' for aggression.[36] There was also a very realistic consequence of more powerful states subjugating weaker ones. Indeed, 'in a world hard pressed to stop aggressive war, it makes little sense to open a loophole large enough to accommodate a tank division.'[37]

The arguments are valid and the intentions are noble. One would certainly endorse this perspective but history appears to also demonstrate some evidence to the contrary. The figurative thorn in the side of the rose is the inherent right established under customary international law. This needs to be discussed for us to understand advocates of the liberal school; one which sees the UN Charter as supplementary rather than superior to the classical law.

Franck, an advocate of the 'preventive self-defence doctrine' (which found its most devout supporters in the Bush administration), stated that 'common sense rather than textual literalism is often the best guide to interpretation'.[38] The arguments committed to the counter-restrictionist theory appear far more compelling and seem to incorporate an element which is more sensitive to developments in the manner in which wars are fought. The position seems to garner substantial strength from its reliance on the customary

---

36 Van Den Hole (n 13) 87.

37 Byard and Brown (n 11) 229.

38 Thomas Franck, *Recourse to Force: State Actions Against Threats and Armed Attacks* (Cambridge University Press 2002) 98.

international law right which maintains the inherent right of self-defence. Indeed some scholars even challenge the idea that self-preservation should even be subject to law.[39]

The *traveaux preparatois* seems to suggest that article 51 was not meant to limit the broader notions of self-defence as imbued by state practice[40] including the Chapultepec Treaty.[41] In addition, case law judgements have re-enforced the criteria of necessity and proportionality as principles of self-defence. These criteria are nowhere referred to in the UN Charter, but rather traditional legal norms from customary international law detailed through the annals of history. The following arguments are two–fold; one is lexical and the other is based on 'strategic concerns'.

The lexical position looks at the wording and approaches the text in a common sense manner. The focus on the word 'inherent' is an explicit reference to a pre-existing right which the treaty was not meant to circumvent.[42] Additionally, armed attack is deliberately vague in order to encourage reference to the customary law; a trigger-word if you will. Van Den Hole uses a logic which renders a literal reading *reductio ad absurdum*.

> If A then B is not equivalent to if A, *and o*nly if A, then B.

In other words, the logic that 'if one is subjected to an armed attack, one can invoke self-defence' is not necessarily the same as the logic that 'if one is subjected to armed attack, and only to an armed attack, then one can invoke self-defence'.[43] However, this gap-filling exercise is perhaps a secondary rather than a primary claim to this

39 Oscar Schachter , 'Self-Defence and the Rule of Law' (1989) 83 The American J Int L 259.

40 Van Den Hole (n 13) 75.

41 Schachter (n 6) 1633.

42 Arend and Beck (n 20) 73.

43 Van Den Hole (n 13) 85.

avenue of thought. One of Van Den Hole's most compelling arguments however, is the reference to article 2 (4) UN Charter. This refers to a prohibition of a threat of force. Read in conjunction with article 51 and the 'inherent right,' the meaning seems to ring truer with the notion of pre-emption.

Another alternative argument is that the UN Charter provision only refers to self-defence in response to an armed attack and that its vagueness is deliberate. The Charter assumes that pre-emption of armed attacks are dealt with by the customary law.[44]

The 'strategic' argument takes into consideration other determining factors such as the development of military technology and tactics. It recognises that if such a strict reading were upheld, it could impede a state's ability to avert the attack in question. McDougall's sitting duck analogy provides the rationale for the customary international law right.[45] It illustrates that such a rule would be senseless. So to wait for an attack you know is coming would be an intolerable doctrine.[46] It is only when you have suffered an attack that you may respond: particularly absurd if there is an impending attack and you have the resources to avert it. Indeed this argument is compelling even for Dinstein who typically holds a fairly conservative approach. He acknowledges that when a state has committed itself to the deployment of an attack from which it cannot backtrack, then the state may use 'interceptive self-defence'.[47] Indeed, to wait for an attack may allow more time for 'the aggressor state' to formulate a greater attack or not pre-emptively attacking may destroy possibilities to attenuate the harm.

One of the key contributions of this school of thought is the development of anticipatory self-defence.

---

44 Shaw (n 14) 1132.

45 Myres S MacDougal, 'The Soviet-Cuban Quarantine and Self-Defence' (1963) 57 American J Int L 597.

46 Rodin (n 1) 113.

47 Dinstein (n 15) 191.

Doctrines of pre-emptive and preventive self-defence have also been created and are sometimes used as interchangeable terms. I will refer to the anticipatory self-defence notion and may, from time to time, refer to it as pre-emption.

Anticipatory self-defence was developed as a theory by Micheal Walzer and it does offer very compelling arguments for the doctrine. He asserts that the moral theory underlying UN Charter, or 'the legalist paradigm'[48], is that the international society is made of independent states and the society has laws which determine territorial integrity and political sovereignty as embodied by article 2 (4) of UN Charter. In addition, any force or threat of force equals a crime of aggression in which case the circumstance may in turn justify self-defence and be justly punishable. Contrasted with Webster's test, Walzer announced that to invoke self-defence, 'there must be a manifest intent to injure, actual preparation for an attack and if one was to wait, it would greatly magnify the risk.'[49] In other words, a state may use force in the face of a threat of attack, in situations when not doing so would impinge the notions set out in article 2 (4). A textbook example of this was the 1967 Arab-Israeli war.[50] According to a still-disputed testimony, surrounding Arab armies lined their troops on their borders. Egypt expelled UN peacekeeping forces and their charismatic President Nasser made threats to interfere with shipping in the Straits of Tiran, at the entrance to the Gulf of Aqaba which was part of an overall plan of aggression against Israel.[51] Finally, and perhaps more crucially, the Egyptian forces purposefully delayed their attack, knowing that this would inevitably place Israel in a state of readiness. Such a state of uncertainty would ultimately sap Israel's ability to fight, and would

48 Micheal Walzer, *Just and Unjust* (4th edn Basic Books 2006) 61.

49 ibid.

50 Gray (n 9) 161.

51 Stanimir A Alexandrov, *Self-Defence Against the Use of Force in International Law* (1996) Kluwer Law International 153.

paralyse the state in fear. Subsequently Israel fired the first attack (this is assuming we accept these facts).

Walzer makes strong consequentialist calculations about the rigour of his theory and its rationale in averting the untold sorrow of war. In consequentialist moral reasoning, you place the good before the right. Generally speaking, you weigh out all the evils that you would avoid if you take a certain action against the evils – were that action not taken. What I find even more interesting is Luban's criticism of these ideas. Apart from the general criticism ascribed to all utilitarians in that their theories require calculations which are messy at best[52] and ignore the opaqueness of war,[53] Luban looks at the morality of Walzer's proposal. It is apparent that such interpretations and their wide acceptance and disapproval often swing on questions of moral and strategic implications. This is why it is important to look at the ethical underpinnings of such a theory and acknowledge that international legal norms do not exist in isolation.

There are certain aspects of Walzer's doctrine and Luban's criticism which are of interest to me. Even though Walzer's theory is forward looking, he makes considerations of past facts to determine the imminence of a threat. Such was the case in the 1967 war. This distinction between an imminent threat rather than a general one is deduced with reference to previous facts; it looks backward. Luban also makes the distinction between determining the 'rapaciouness'[54] (to use Vattel's terminology) of a state in probabilistic rather than temporal terms.[55] He cleverly uses

52 David Luban, 'Preventive War' (2003) 32 Philosophy and Public Affairs 225.

53 David Rodin, 'The Ethics of Asymmetric War' in Richard Sorabji and David Rodin (eds) *The Ethics of War: Shared Problems in Different Traditions* (Ashgate Publishing Limited 2006) 172.

54 Emer de Vattel, 'Droit des gens; ou, Principes de la loi naturelle appliqués à la conduite et aux affaires des nations et des souverains' (translated: Law of Nations or the Principles of Natural Law Applied to the Conduct and to the Affairs of Nations and of Sovereigns) (J Robinson London 1797) 445.

55 Luban (n 52) 230; I would suggest that probability exists within a temporal framework.

this probability aspect to separate the moral basis between pre-emptive wars and preventive wars (the latter responding to general threats). More importantly, he identifies certain characteristics which would increase the probability of an attack. This he does by defining what is a rogue state; one which favours a violent and militarist ideology. In order to ascertain whether such a state is rogue, he refers to their 'track record'[56]. This is of particular interest because I feel the literature lacks texture and sophisticated consideration of past events in determining one's claim to self-defence. Some scholars do to a certain extent but I contend that this is insufficient and I shall elaborate, on why, later.

The controversy over interpretation of this article can be summed up in this sporting mantra: 'the best defence is a good offence.' This supports the view that self-defence can be launched prior to an actual armed attack and with good reason; in essence, the basis of anticipatory and pre-emptive self-defence. A lack of affirmative declarations throws the question of relationships between the treaty law and the pre-existing right into disrepute. Had there been an express repudiation of the 'inherent right' in the treaty, then a strict interpretation of it would have been legally and morally sound (but I hasten to affirm the simplicity of such a task). However, the retorts are scathing; that the pre-existing right does still exist and that military technology is such that it would be dangerous to assert such reasoning. What needs to therefore be established is an alternative interpretation; one which respects the pre-existing right but also the well-willed intentions of the UN Charter in restricting force and preventing abuse. It also needs to follow the sprit of Van Den Hole that should a state invoke self-defence, they are subject to rigorous procedures to determine the claim of their right to invoke self-defence. I acknowledge that claims to anticipatory self-defence are often gratuitous and abused, not surprisingly by the regional and world hegemonic-powers for

56 ibid 232.

non-altruistic ends. The counter-restrictionist position tells us something quite clear about the nature of international relations. On the one hand they aim to justify their actions within the framework of the law; yet by the same token 'they are quick to interpret every legal restrain upon building power potential as an inhibition of their self-protection.'[57] But given what we have, we need to develop a theory which is encompassing of these traditional legal norms but which redresses an unfair imbalance in the abuse of this doctrine.

## III. Re-thinking the ethics of killing in self-defence

We need a theory which respects the spirit of the UN Charter as embraced by article 2 (4) and a theory which recognises the customary international law right. But perhaps most importantly, we need a theory which is not open to abuse and considers very strongly the concerns that Brownlie makes with reference to the 'carte blanche for aggression'. It needs to be a rule which paradoxically is both proscribing and prescribing and which is not susceptible to mistake or fraud. To begin this experiment, I need to take the law of self-defence and strip it down to its fundamental unit. Walzer refers to the 'domestic analogy' which compares international self-defence to personal self-defence: what we would understand as self-defence in criminal law. Walzer makes the point that the two are isomorphic.[58]

All things considered (when looking at other theories and piecing together my own), I will talk about the law of self-defence between individual units/persons and the assumption is that these can refer to individual states. Identifying problems with the 'domestic analogy' shall be dealt with later.

Self-defence poses questions of 'morality in extremis.'[59] It quite literally requires us to make life and death assessments of situations and offers one of the few

57 Werner Levi, *Law and Politics in the International Society* (Sage Publication 1976) 55.

58 Rodin (n 1) 108.

59 ibid 1.

instances in which killing is morally and legally permissible. My strong feeling is therefore, that when analysing the ethical basis for self-defence, the paramount mode of moral reasoning has to be intuitionism. If we have a situation which seems to have favourable consequentialist outcomes but feels counter-intuitive, it will be difficult to assume this as morally acceptable (a process of ad hoc reflective equilibrium).

We are dealing with a moral asymmetry in which a situation is created wherein one of the actors has a right to kill another. There are several approaches which I shall go through and critique. However, the ultimate aim is not necessarily to develop and improve these ideas. To the contrary, it will be to identify what I perceive as a weakness in their methodology. Note, all the following examples work on the presumption that the only way for a victim to save himself is by killing his aggressor.

Thomson's approach is widely accepted as having ignited the debate of killing in self-defence. Her argument is incrementally built up using several examples trying to force a universal moral conclusion on each scenario. She begins with the 'villainous aggressor'[60] which illustrates the textbook case of morally justified self-defence. You are approached by x who wants to kill you and the only way to prevent x from doing so is killing him. We then move on to the 'innocent aggressor'[61] who, like the villainous aggressor wants to kill you, and to stop him from doing so would require you to kill him. Yet his aggression has come from an ephemeral lapse of sanity. The final scenario identifies the 'innocent threat'[62] wherein a fat man is perniciously pushed off a cliff in your direction and the only way to save yourself is through deflecting him (and as a result killing him). Controversially, Thomson sees no moral difference between the three and

60 Judith Jarvis Thomson, 'Self Defence' (1991) 20 Philosophy and Public Affairs 283.
61 ibid 284.
62 ibid 287.

thinks that it would not only be morally excusable but justifiable to kill.

This rights-based account[63] makes no demarcation between fault or moral agency; so even though the innocent threat has no autonomy of his act, Thompson asserts that it is still morally acceptable to kill in self-defence – a position that David Rodin, as we shall soon see, disagrees with.[64] Her formulation rests on the premise that we have rights against one another not to be killed. Upon aggression by x, x forfeits his right not to be killed and therefore the victim may kill him without violating x's rights.[65] This means that x's right to life includes a claim against others that they not kill him. When he aggresses, he forfeits this right and looses his claim against the victim. McMahan rightly acknowledges that something, particularly with the innocent threat, is not quite right here. It seems counter-intuitive that a falling fat man, with no fault or moral agency could therefore lose his right to not be killed. This forfeiture framework doesn't seem adequate to explain certain problem cases; indeed, it leads us into a philosophical quagmire.'[66] Rodin correctly identifies that forfeiture of rights can turn on facts 'about status, condition, actions and intentions of both parties.'[67]

If we are to elevate this latest scenario to the situation between states in the context of *jus in bello*; soldiers in conflict maybe entirely innocent (even going as far as opposing their presence in a particular country) despite the illegality of the *jus ad bellum*. However, because their default position is that of being under orders to kill, in spite of the fact they have no fault, it would expose them to being justifiably killed (arguably they have agency). Quong raises equally valid concerns regarding the lack of moral agency of

63 Jeff McMahan, 'The Basis of Moral Liability to Defensive Killing' (2005) 15 Philosophical Issues 387.

64 Rodin (n 1) 79.

65 Thomson (n 60) 302.

66 Rodin (n 1) 70.

67 ibid 76.

the 'innocent threat', who therefore should not be subject to moral duties.[68] The main principle here is that it seems counter-intuitive to subject someone who has no intention to kill to these moral burdens. Whilst we can excuse an individual from killing the falling fat man, it seems against our intuition that such an act is even morally permissible.

One of the aspects that I find appealing about this theory is the notion of loosing a right not to be killed. Many of the problem cases, when elevated to the level of states, would unlikely occur. A state will never commit aggression in a lapse of sanity in the conventional sense. But it seems plausible that a state, through some type of aggression it commits in the present, or in the past (in respect of a series of attacks) may forfeit its right to defend itself.

Another account refers to culpable liability.[69] This works on the premise that when an individual puts himself in a situation, such as pointing a gun towards your head, he makes himself liable to be killed. Depending on the victim's epistemic limitations it will therefore be determined whether the act is morally permissible or morally excusable. For example, assume the gun was not actually loaded but for some reason, neither individual knew this. If we couple culpable liability with an objective account of facts it would only make it morally excusable. However if we couple it with a subjective account then it becomes morally permissible.[70] The latter seems far more attractive but it seems peculiar to make an act morally different based on a lack of knowledge.

Culpable liability is perhaps the basis of article 51 of the UN Charter. It says that should a nation state commit an aggression, it is liable to attack based on the principle of self-defence (under the parameters of necessity and proportionality). This is all good and well but we can't celebrate too prematurely for we are forcibly pulled back into

68 Jonathan Quong, 'Killing in Self-Defence' (2009) 119 Ethics 515.
69 McMahan (n 63) 397.
70 ibid 398.

our original *harmatia*: determining the difference between aggression and pre-emptive attack.

McMahan also puts forward his most favourable account which is the justice-based account[71]. Here, distribution of harm is attributed to those most responsible for the harm, other things being equal. It does not necessarily require harm but does require agency. Therefore what it advocates is very similar to the tort law of negligence in terms of causal proximity between breach and harm suffered. A person's liability increases when one's action is more fault-inclined. For example, when driving a car you account for all the consequences of potential accidents regardless of how remote they maybe to your actual driving the car.

This at first seems attractive but is open to many objections. For example, how can we determine who is the initial moral agent responsible? McMahan suggests the extreme possibility that potentially the mother of the villain could be liable.[72] One of the things I feel a lot of these theories lack is a focus on the rights of the victim; instead they look at the rights, fault or agency of the aggressor. Quong however, switches the focus and asserts that 'each person is understood to have a powerful agent-relative permission to avoid sacrificing or significantly risking their own life for the sake of others (in the absence of any obligations voluntarily incurred)'.[73] I find this particularly convincing, as it seems to appeal to our intuition. The following example will illustrate this idea.

Suppose Tanveer is slowly being engulfed by quicksand. I can rescue him but I know that in all likelihood I will loose my watch. Many would conclude that I am required to rescue Tanveer. Suppose however, a lion is embedded in the quicksand and will most likely devour my

71 ibid 403.

72 ibid 405.

73 Quong (n 68) 51.

legs if I try to rescue Tanveer. If we focus on the rights of Tanveer, they far outweigh my legs being eaten. But it seems difficult to morally compel me. One way is to firstly look at the agent-relative value which changes the moral outcomes. If I translate this to the level of nation states, one could interpret it as states thinking that their life is important to them. In situations where this is being threatened, they can take certain measures (albeit necessary and proportional) to prevent such outcomes. This fits in neatly with the aforementioned point in terms of a realist's conception of international anarchy.

The theories I have briefly discussed are by no means comprehensive and they all offer some interesting commentary on the philosophy of self-defence. I have highlighted the view that when a state aggresses against another, then it violates and infringes a right to life and integrity of the other state, in such a way that makes them morally susceptible to attack. This seems accordant with what we understand as an inherent right of self-defence. I also see that in a world of sovereign nation-states, it is understandable that they would be self-interested before looking out for others. This idea is in line with Quong's 'agent-relative value'. After all, it is the nature of wars in self-defence in that they are different to general wars; they are special-interest wars.[74]

One final point; Rodin makes very good criticisms of why the 'domestic analogy' is philosophically misleading. To understand why, he makes reference to what is known as the 'Hohfeldian correlate'.[75] He says that the domestic analogy refers to a normative relationship in which the following elements are present; a subject (defender), object (aggressor), act (homicide) and the end (to protect yourself). The aggressor, as he is morally at fault, loses his right upon this aggression and can be killed (circumventing the so called

74 Luban (n 52) 221.
75 Rodin (n 1) 75.

inalienable right of life paradox).[76] My right not to kill you is the logical correlate of your duty not to kill me. Therefore my right to kill you in self-defence is the logical correlate of your failure to possess the right that I not kill you.[77] He elaborates further saying that rights not to be killed are interpersonal and require reciprocity, as the Hohfeldian liberty illustrates. When we consider this in the realm of national self-defence, we must consider its relationships in times of war and acknowledge that they expose very different elements in war. He cites that there are two levels of war;[78] between peoples and states, and it is a moot point whether national self-defence is conceived as a right against people or states. Zohar refers to this as moral vertigo.[79]

My response to this is subtle. This analogy does make a lot of assumptions about the content of relationships between peoples and states and their apparent similarities. Because these theories work on a rights-based approach, they are exposed to this criticism. However my approach doesn't consider rights per se. Rodin's analysis only falters your domestic analogy if one assumes the normative relationship in the Hohfeldian sense described. I acknowledge that they can be surmised in this way but I would refute that this has a monopoly on the framework of explanation.

## IV. History and the Notion of Pre-Emption

Many may have always held the normative conception of international law and enforcement as having the potential to ensure real justice. Such perceptions work on the notion that we right every wrong regardless of its time or place in history.

---

76 Cheyney C Ryan, 'Self-defence, Pacifism and the Possibility of Killing' (1982) 38 Ethics 510.

77 Rodin (n 1) 64.

78 ibid 65.

79 Noam J Zohar, 'Collective War and Individualistic Ethics: Against the Conscription of "Self-Defense"' (1993) 21 Political Theory 615.

History itself is a very important concept in law. We, as lawyers, must determine facts which happened in history from an objective viewpoint. The problem with adjudication is that we always assume that cases exist in isolation; in a vacuum in which time and space cannot enter. This legal black hole as it were, hinders the real effectiveness of law in restorative justice. One thing that all these theories tend to lack is a consideration of historical elements. Whilst I appreciate that history is itself elusive,[80] a veneration of what has happened rather than what will happen I think is infinitely more useful in our determination and application of self-defence. I explain how my interpretation works and how it alters the differing restrictionist and counter-restrictionist schools of thought.

In every second of our lives, our minds subconsciously make decisions about the behaviours of others, marking them with a moral tick or cross; not desirable perhaps but perfectly understandable. I observe a woman helping an old man with his shopping and assess this as a good thing. When we observe certain actions in isolation, we arrive at certain conclusions. Take the following for example:

Ashley is walking down the street minding her own business when her attention is accosted by an incident across the road. She sees Willard, a neighbour, being hit, with some rigour, in the stomach by the school's head boy, Carleton. Like most of us, she makes a fairly uncontroversial moral assessment of that particular act, in isolation, as being wrong. The two are quickly reprimanded by the local officer, PC Phil.

Ashley, much to her chagrin, visits the police station to go through some formalities as a witness. Later she learns some interesting information about Willard, in that he was the school bully and had a long history of picking on Carleton. This included stealing his money, yelling

80 See also Friedrich Nietzche, *On the Advantage and Disadvantage of History for Life* (Hackett Publishing1980).

profanities at Carleton and sometimes even hitting him. Upon the revelation of this new information, her earlier moral assessment of Carleton's action changes and although she does not necessarily think it was morally justifiable, she begins to think his action maybe excusable. Something happened when she received new information. It would seem history changed her evaluation of the facts-past temporality has normative value.[81]

With this interpretation, rather than looking at either agent individually, the approach is radically a holistic approach and so it observes them all – in their entirety. Entirety here entails not just the present facts, but the past facts. It determines, regardless of how far back in time, who the initial aggressor was. This appears a very difficult task and indeed I will address the problems with this analysis. But if it becomes possible, and that we may answer this question with unanimity, then we extinguish all of the criticisms afforded to both schools of thought.

I shall explore, a little further, the importance of history and context before I explain how such a law should be worded. My interpretation emanates from disillusionment with liberal theories of justice which isolate people's choices from their cultural and temporal context. This type of thinking is essential in determining behaviours of other people. For example the morality (and therefore temporality) of a terrorist is different from the morality of a pacifist. It would be seemingly absurd to afford the moral standards of the former to the latter. Much like philosophers that say culture is the context of choice, [82] I consider circumstances. I also make the proposition that states can behave in very similar ways where their choices are subject to a veritable wealth of external influences; including history. Indeed, to accept this choice (in this case of self-defence) 'at face value, one has to take into account that their current

81 This forms one of the basis of my doctoral research.

82 Jan Van Der Stoep, 'Towards a Sociological Turn in Contextualist Moral Philosophy' (2004) 7 Ethical Theory and Moral Practice 134.

convictions and behaviour are shaped by circumstances of domination.'[83] History acknowledges the complexity of situations and is crucial in the 'intellectual task of generating or discovering principles which require choices to be made.'[84]

What I have briefly argued is that when we are revealed past facts about certain situations, we change our moral perspectives. I think this is fairly uncontroversial as we tend to look favourably on things put into context and unfavourably on things put out of context. I have also looked at how this is important in determining the choices people make relative to their circumstances.

The key premise we have is history. This does two important things; it helps to identify the initial aggressor but also helps to evaluate the severity of a threat. Let us assume that the UN Charter is absolute (like an act of legislation in English law); a peremptory norm that codified rules of customary international law. This means we embrace the spirit of article 2 (4) UN Charter which respects sovereignty and tries to eliminate the use of force. But what it also means is that we take on the Caroline Incidents idea of pre-emption. However, with this we add a few adjustments; a provision clarifying circumstances when pre-emptive self-defence maybe used. This will introduce the element of history and what Luban brilliantly considers as rogue states. To take Lubans' militarism definition that is 'ideology favouring violence [with] a track record of violence and a build up in capacity to pose a genuine threat.'[85] He says that if we are to justify a preventive war, it can only be based on determining the character of whom our attack is aimed at. One way to establish such a character is by looking at history.

By looking at history, we can piece together whether or not a state poses a genuine threat. But herein lies a

---

83 ibid 137.

84 Will Kymlicka, *Contemporary Political Philosophy: An Introduction* (2nd edn OUP 2002) 35.

85 Luban (n 51) 231.

hurdle; the very job of a historian is to establish one universally accepted narrative of history. How can we therefore determine which history is attributable to the 'rogue-ness' of a state and which is in fact a response or consequence of repression by another rogue state. To illustrate this problem clearly, let us make some assumptions that it maybe suggested that guerrilla groups are inherently militaristic. However, often guerrilla groups are responses or reactions of a people, under repression, vying for political and economic emancipation. Their rogue-quality is not instinctive but manufactured: they have been coerced into violence. To put it simply they are a product of their circumstances. However, violence of other state/non-state actors show a propensity for violence that is innate, rather than a product of circumstances and indeed, such propensities can be evidenced with historical records. Admittedly this is over simplified, but the role of this thought experiment is to highlight the merit and importance of looking backwards.

One pressing question therefore, is how far back do we have to go to determine the 'rogue-state'? This brings us back to the very problem with all self-defence theories; identifying the initial aggressor. Assuming that the requirement for a rogue character had been satisfied, could for example the Americans try to justify self-defence against Britain citing British imperialism and colonial aggression as the initial aggression? If one can objectively verify that Britain was a belligerent state and had been since, then it is difficult not to arrive at a conclusion that self-defence would be justified. To clarify, I shall illustrate using the following example:

Let us assume the worst; that Nazism was an enduring ideology and succeeded to this day. As part of its policy, it continues to commit pogroms against Jewish, black and disabled people. There is a state neighbouring Germany which is home to these groups. Germany, in its unremitting commitment to its fascist ideology, occupied or militarily intervened in this country. Over time, this fictional state

grows more incensed and irate as death and casualties accrue. They begin to mobilise a guerrilla resistance and fire missiles into Nazi Germany. Now we can *prima facie* observe that Nazi Germany is a rogue state. What is also apparent is that they were the initial aggressors. This is verifiable through historical accounts of these pogroms. If the guerrilla army were therefore to conduct operations in self-defence, their violence would be excused because it is as a response to the rogue state. History therefore, has determined the initial aggressor and the severity of the threat. This argument perhaps reduces complex state relationships to a rather simplistic formulation but complexity is not an argument against producing authoritative rulings.

We can take some inspiration from McMahan's analysis of preventive war. It is fairly uncontroversial that preventive war is illegal under article 51 of UN Charter (as it responds to general threats which could be far back in time) but McMahan puts forward moral arguments for it albeit inside a lattice of very strict moral constraints. Consider the battered woman case; her husband has a history of violence against her and she has a reasonable belief that her husband will attack (although the threat is not imminent). The problem emanates from insufficient evidence to establish the probability of the attack.[86] Luban picks up on this by saying that 'we re-characterise imminence in probabilistic rather than temporal terms.'[87] Whilst I would suggest probability and temporality share some common ground, it naturally follows that determination of this can be based on evidence in history. McMahan says that intuitively we would accept a war of prevention if compelling evidence that the state would unjustly attack us, that waiting would lessen effective response and that peaceful means have been exhausted.[88] I think this

86 Jeff McMahan 'Preventive War and the Killing of the Innocent' in Richard Sorabji and David Rodin (eds) *The Ethics of War: Shared Problems in Different Traditions* (Ashgate Publishing 2006) 172.

87 Luban (n 52) 231.

88 McMahan (n 86) 172.

is true. However, I would not necessarily endorse the theory because it would be prone to abuse (but discussion of these ideas means that history has some relevance).

We shall establish 'imminence' by introducing the 'probabilistic/temporal' element. This will be evidenced by the history of the aggressor to whom the self-defence is being invoked against. If the threat is ongoing, it will make the probability of the attack more likely. This is often referred to as the accumulation theory [89] and rightly identifies the difficulty distinguishing between self-defence, reprisals and so-called pinprick attacks.

This is an all-encompassing interpretation of the self-defence law as embedded in the treaty provisions, customary international law and (taking into consideration) moral and ethical considerations.

> Let us call it historical self-defence.
>
> 'A state may invoke its inherent right before an armed attack if it is imminent and the state in which it is invoked is considered rogue. The state in question's historical record of aggression against the state wishing to invoke self-defence will determine whether a state is rogue and if the attack is imminent.'

To clarify the last sentence; this follows a circular reasoning. The historical record of events informs us of whether a state is rogue and whether a state is rogue informs us whether an attack is imminent. Imminence is measured in probabilistic terms; which is in turn, determined by the historical record.[90] All areas of the hypothetical provision are intimately linked.

---

89 Alexandrov (n 51) 166.

90 Note, it would also be subject to high evidential standards and would require an impartial and objective mechanism for determining 90 the facts and historical record. My proposal shall offer two levels of exculpation; when the attack is imminent (one of the requirements of anticipatory self-defence), then it will be justifiable; if the threat is more remote, it is only excusable.

The best way to demonstrate this new interpretation and indeed to identify its weaknesses is by looking at a few short examples.

## V. Application: Teasing Out the Problems

To really put this theory to the test, it would be most useful to use a case study whose history is most disputed. The question concerning Palestine indulges us into two separate narratives of history; one which saw 1948 as the triumphant declaration of Israeli statehood and the other which the Palestinians mourn as their 'Al-Nakba' or 'catastrophe'. The events I will use are going to be as all encompassing of both accounts of history as possible. The aim is not necessarily to condemn or condone the other; rather it is to demonstrate the theory.

Let us consider the war in the Gaza Strip in 2008. The charge often levelled at the government in Gaza's military wing is the constant rocket fire into the southern Israeli cities of Ashkelon and S'derot.[91] The Israeli Ministry of Foreign Affairs states that 1750 Qassam rockets and 1528 mortar bombs were deployed in 2008. This has become the propaganda discourse for Israel's war with Gaza. When Israeli forces executed its Operation Cast Lead on December 27th 2008, self-defence was cited as the basis for its use of force.[92] Prior to the war there had been a four-month ceasefire in which the number of rocket attacks dropped to virtually zero yet the Israeli Defence Forces committed targeted assassinations on Gazan government leaders.[93] Let us assume the facts to be true (in all likelihood they are; the

---

91 Intelligence and Terrorism Information Centre at the Israel Intelligence Heritage & Commemoration Centre 'Anti-Israel Terrorism in 2007 and its trends in 2008' (2008) 26 available at <http://www.terrorism-info.org.il/malam_multimedia/English/eng_n/pdf/terror_07e.pdf>.

92 Rose Mishaan, 'Introduction- Recent Events in Gaza' (2009) 32 Hastings Int L Rev 639.

93 Victor Kattan 'Gaza: Not a War of Self-Defence' January 15 2009 Jurist: Legal News and Research <http://jurist.law.pitt.edu/forumy/2009/01/gaza-not-war-of-self-defense.php> accessed 15 May 2010.

dispute is the selection of relevant facts rather than convenient ones[94]).

The Qassam rocket attacks, like the example we used above regarding Ashley, Willard and Carleton, is an intrinsically wrong act. However, if we reverse back into history, maybe our perceptions of this violence will change. To quote Kattan, 'one cannot ignore the conduct of Israel's armed forces in the occupied territories and examine the rocket attacks in isolation.'[95]

History will determine the severity and imminence of a threat. So if the government in Gaza wants to justify the rocket attacks (rather than excuse them), it has to prove that an attack is imminent, much like Walzer's anticipatory self-defence. And it has to prove that the attack to which the state is directed, is rogue. Rogue, as mentioned before, is ascertained by a forensic analysis of historical documentation. This is where we potentially hit our first snag in the theory. What facts are relevant? Which are merely convenient? What is objective and what is subjective? I have acknowledged this problem before; but I do think it is important and achievable. Recall my conceptualisation of justice as righting every wrong regardless of time. History is required to do such a thing. If we can create an exercise which is able to determine objective history (or at least a historical account of the facts which has popular consensus) then this formula will surely work.

The most recent UN Security Council resolution 1860 [96] re-affirms the infamous Resolution 242 [97] (and

94 Zohar (n 79) 612 Zohar refers to HLA Hart's terminology citing casual attributions as 'ascriptive' rather than descriptive in that they ascribe responsibility to the agent rather than reporting the objective sequence of events.

95 ibid.

96 UN General Assembly Resolution supporting the immediate ceasefire according to Security Council Resolution 1860 23 January 2009 A/RES/ES-10/18 available at: <http://www.unhcr.org/refworld/docid/49917ee92.html>.

97 UN Security Council Resolution 242 (9 November 1967 S/RES/242 (1967) available at: <http://daccess-dds-

subsequent Resolutions 338[98],1397[99], 1515[100] and 1850[101]) that Palestine continues to be occupied by Israel. Also international law, even after the disengagement of 2004, recognises that Israel continues to occupy Gaza. [102] Furthermore, we have Resolution 799[103] condemning the deportation of hundreds of Palestinian civilians in contravention to Israel's obligations as an occupying power under the Fourth Geneva Convention; Resolution 904[104] expressing shock at the appalling massacre committed against Palestinian worshippers in Hebron; Resolution 673 [105] adopted unanimously with reference to Israel refusing to receive the mission of the then UN Secretary-General[106]; Resolution 106 admonishing Israel's pre-arranged and planned attacks inside the Gaza Strip.[107] All in all, Israel has accumulated 223 UN Security Council Resolutions

---

ny.un.org/doc/RESOLUTION/GEN/NR0/240/94/IMG/NR024094.pdf?OpenElement>.

98 <UN Security Council, Resolution 338 October 1973 S/RES/338 (1973) available at: http://daccess-dds-ny.un.org/doc/RESOLUTION/GEN/NR0/288/65/IMG/NR028865.pdf?OpenElement>.

99 UN Security Council Resolution 1397 S/RES/1397 (2002) available at: <http://www.un.org/News/Press/docs/2002/sc7326.doc.htm>.

100 UN Security Council, Resolution 1515 S/RES/1515 (2003) available at: <http://unispal.un.org/unispal.nsf/0/71B2C135FCA9D78A85256DE400530107>.

101 UN Security Council, Resolution 1850 S/RES/1850 (2008) available at: <http://www.refworld.org/cgi-bin/texis/vtx/rwmain?page=publisher&publisher=UNSC&type=&coi=PSE&docid=4950d67c2&skip=0>.

102 UN Human Rights Council, UN Human Rights Council : Report of the Special Rapporteur on the Situation of Human Rights in the Palestinian Territories Occupied since 1967 available at <http://www.unhcr.org/refworld/docid/461e52b12.html>.

103 UN Security Council Resolution 799 S/RES/799 (1992) available at: <http://www.unhcr.org/refworld/docid/3b00f15f4f.html>.

104 UN Security Council Resolution 904 S/RES/904 (1994) available at: <http://www.unhcr.org/refworld/docid/3b00f15f4f.html>.

105 ibid.

106 UN Security Council Resolution 673 S/RES/673 (1990) available at: <http://www.unhcr.org/refworld/docid/3b00f13844.html>.

107 UN Security Council Resolution 106 S/RES/106 (1955) available at: <http://www.unhcr.org/refworld/docid/3b00f13934.html>.

condemning its use of aggression against Palestine and other Arab states; more so than any other state in the world. It continues to defy humanitarian law and has not resolved any of the resolutions which are annually re-affirmed. This seems enough to determine the severity of the threat. It is apparent that the state has a propensity for aggression, but determining the initial aggressor is not something which has been ratified by law. This is where we run into a potential *cul-de-sac*. Now we must rely purely upon history to determine the origins of the conflict. I think history can, and indeed does, healthily inform law. But it all depends on whose arguments we find more compelling. But is this not the job of a litigant? If jurists could objectively determine the origins of the conflict, coupled with the affirmation of the history as it represents violence by Israel, it gives more credence to the imminence of a threat. So let's put this back into our case study and relate it to our very first hypothesis.

If we recall the example we used earlier of Willard and Carelton and contextualised violence; if Qassam rocket attacks were fired in response to what they gauged, and can be objectively verifiable by an independent and neutral body, as an imminent threat, based on Israel's history, these would be justified. However, if the threat were far more remote the rocket attacks would only be excusable.

If my theory were to stand true, how could a weak state ever commit an act of aggression and use unlawful force? Surely they would always justify every use of force using this theory, in essence exercising the very type of arbitrary force that powerful states use. This accentuates the need for an enforcement body which can create a system that subjects all states to the rigour of due process, unlike the United Nations Security Council.

The questions for deliberation in the light of the Falklands war are very interesting. The claim from Argentina is that they were exercising their right of self-defence since they had territorial claims pursuant to article 2 (4) of the UN Charter. They said that Britain had usurped the island 149 years ago and used continuous force. Alexandrov worried

that if the Argentine claim was tenable, it would allow the very thing which my interpretation permits: 'claims for restoration of the status quo ante.'[108] What of the Osirak case where Israel destroyed a nuclear reactor in the Tuwaitha Nuclear Facility? This is a fascinating case as the situation is not clear-cut. Although condemned by the international community under Resolution 487,[109] this was an example of a far remoter threat. Had for example, Iraq had a history of violence against Israel, Israel's action would have only been excusable. If the Iraqis had developed their nuclear facility to a level capable of manufacturing nuclear weapons, it maybe argued that Israel's actions were then justified, but this is conjecture and subject to empirical evidence to suggest the counter. This also raises the broader question of nuclear weapons; would a state be able to use this in historical self-defence? I shall not go into this here but it is safe to say that I would have affirmed Judge Schwebel's remarks that nuclear weapon is an exception and under no circumstances should it be used by anyone in self-defence.[110]

## VI. Conclusion: Implications and the Real Need For a Radical Interpretation

> 'We must...recognise that by this temporary submission of the Vanquished[...]a new political order is initiated, which, although without moral basis, may in time acquire such a basis, from a change in the sentiments of the inhabitants of the territory transferred, since it is always possible that through the effects of time and habit and mild government...the majority of the transferred population may cease to desire union - when this

108 Alexandrov (n 51) 132.

109 UN Security Council Resolution 487 S/RES/487 (1981) available at <http://www.undemocracy.com/S-RES-486(1981).pdf>.

110 Legality of the Threat or Use of Nuclear Weapons - Advisory Opinion of 8 July 1996 - General List No 95 (1995-1998) <http://www.icj-cij.org/docket/files/95/7515.pdf?PHPSESSID=dc8e14d6e87c61e15b4d964715510a0a>.

> change has taken place, the moral effect of the unjust transfer must be regarded as obliterated; so that any attempt to recover the transferred territory becomes itself an aggression.'[111]

The implication by Sidgwick's quote is that it is acceptable to let past injustices go unanswered; that time is the great healer. The assumption is that past transgression, etched in the consciousness of a people can be all too easily forgotten. History demonstrates to the contrary. This mode of thought is precarious and it once again sucks us back into the vacuum of our legal black hole.

What international law in general tells us is descriptive of the distribution of power in international relations. One very simple example of this revolves around whether the definition of armed attack includes economic sanctions; favoured by the hegemonic states like the US as exemplified in Iraq, and disfavoured by the smaller ones.[112] The fact that it politicises the law should be no surprise. International society is shaped by the very interests of states.[113] The creation of law, whether crystallised in treaties or developed through customary international law, is determined by those states for they are the de facto lawmakers. Indeed there exists the notion of sovereign equality, but without an independent arbiter with wide reaching jurisdiction (indeed even the history of the UN has revealed 'a very high degree of complicity with the politics of power and imperialism'[114]), it is difficult not to see this as anymore than a legal fiction. The legal institution in international relations could be a force to reckon with but 'its influence is diluted, however, and sometimes outweighed, by other forces in a developing international society.'[115]

---

111Henry Sidgwick, *The Element of Politics* (Cambridge Library Collection 2012) 268.

112 ibid 111.

113 Levi (n 57) 152.

114 Miéville (n 18) 290.

115 Rein Müllerson, *International Law, Rights and Politics* (Routledge1994) 6.

By illustrating the implications and their effects, I think it will demonstrate the need for such an interpretation as posited, or at least one which is historically reflective. The beauty of international law is that it is wonderfully abstract and prosaic; and herein lies its enigma, in that it is wonderfully abstract and prosaic: it allows a whole spectrum of different interpretations. My position perhaps emanates from a favourable view of secessionism or redress for weaker states that are subject to annexation, or occupation in contravention of international law; or who are regularly subjected to human rights violations. The claims of self-determination for example, are all too easily quashed because they aim to address and alleviate the grievances of the weaker party. For example, in January 1978, Australian Minister for Foreign Affairs, Mr Peacock, deplored the use of force by the Suharto government against East Timor but accepted its integration into Indonesia. Inaction and accepting something *de facto*, or out of reality is effective complicity in these types of crimes.[116] It is interesting to note, had the East Timorese mobilised a national liberation army, under my interpretation, they would have been justified in using historical self-defence. Under the restrictionist approach, they would have to wait for an armed attack from a much more powerful state which could have bombed it to oblivion (and most likely diminished its ability to respond). On the other hand the counter-restricitionist school, whilst they may have been able to pre-emptily attack, the moral justification would have been far less. But it is hardly surprising that the literature has developed in this way; it is politically and economically in the interests of states to be able to use violence with few constraints and thus legitimise such action.[117] Flagrant use of violence is not a result of super-power politics, it is constitutive of it. Given that the nature of law is that which is created by states, it is unlikely

116 Tanzil Chowdhury, Nothing is Something Dangerous <http://www.e-ir.info/2012/09/06/nothing-is-something-dangerous/> accessed on 15 March 2013.
117 Miéville (n 18) 287.

that they will create or affirm laws which are a hindrance to their exponential power growth.[118]

This theory bores out of an inadequacy of the respective restrictionist and counter-restrictionist camps. Breaking away from orthodoxy, which suggest that only the latter benefits the more powerful states, I suggest that both do. The former, although a small hurdle, is counteracted by the stronger state's ability to quickly and efficiently respond to an actual armed attack. The latter, as we have discussed in detail, is prone to far reaching exploitation. An anomaly one could identify with my interpretation, is its bifurcation - the effective creation of two different laws: one for strong states, and one for weaker ones. This is not evident in the wording of the text but perhaps in the way the interpretation manifests itself. The very idea of having different laws or rather different standards was formulated and developed by Rodin when he refers to the 'ethics of asymmetric war.'[119] In it he talks about the trials and tribulations of *jus in bello* where we often have strong versus weak with a common aim to make war as less bloody and short as possible. The weak cannot fight using the conventional methods that the strong does. They do not have the smart-weapons or laser-guided missiles which target only combatants. Should they be subject to the same humanitarian laws as the strong? Rodin suggests no when detailing his argument and I think this has some resonance in the *jus ad bellum*. The law as it exists is inherently unfair towards weaker states. It makes assumptions, through its distorted lens of so-called sovereign equality, that states have equal military capabilities.

This interpretation aims to contextualise all conflicts. International law, more than anything, presents agglutination between the disciplines of law and politics. Both have an intimate relationship which informs one another. In addition to just being a legal rule, the most important thing it is meant

---

118 Levi (n 57) 53.
119 Rodin (n 1) 161.

to do is stimulate a serious reflection within the international arena. It is meant to provide small groups of people, weaker or repressed states the means for greater legal recourse. This is based on an acknowledgment that the origins of their suffering are often etched in history and embedded within the positive (and natural) law. If these types of entities are able to use this new interpretation to justify seemingly violent acts (what some may even refer to as terrorism) it may make us all begin to think critically about the parameters of such conduct. Why are such acts of violence being committed and yet they are accepted as morally and legally permissible? Inevitably, many of the questions will be as a result of their past transgressions.

This more general approach to the question illustrates the main objective; not necessarily to convince people that this is good law and that it should be law. But rather to encourage a critical reflexive attitude when it comes to claims of self-defence. There is a lot at stake here, not just national pride, but national integrity and, most importantly, lives. Thus the motivation is not one compelled by historical revisionism but historical affirmation. It is a process which informs the law and that ensures historically legal justice. But before international law can be serious about self-defence, states, particularly the powerful ones, need to be serious about international law.

## BIBLIOGRAPHY

### Books

Alexandrov S, *Self-Defence Against the Use of Force in International Law* (Kluwer Law International The Hague 1996)

Arend A and Beck R, *International Law and the Use of Force* (Routledge London 1993)

Art R and Waltz K (eds) *The Use of Force: Military Power and International Politics* (6th edition Rowman and Littlefield Publishing Group Oxford 2004)

Bix B, *Jurisprudence: Theory and Context* (6th edition Carolina Academic Press 2012)

Cassese A, *International Law in a Divided World* (OUP Oxford 1986)

Dinstein Y, *War, Aggression and Self-defence* (4th edition CUP Cambridge 2005)

Franck T, *Recourse to Force: State Actions Against Threats and Armed Attacks* (CUP Cambridge 2002)

Gazzini T, *The Changing Rules on the Use of Force in International Law* (Manchester University Press Manchester 2005)

Gray C, *International Law and the Use of Force* (3rd edition OUP Oxford 2008)

Kymlicka W, *Contemporary Political Philosophy: An Introduction* (2nd edition OUP Oxford 2002)

Levi W, *Law and Politics in the International Society* (Sage Publication London 1976)

McLeod I, *Legal Theory* (5th edition Palgrave 2010)

Miéville C, *Between Equal Rights: A Marxist Theory of International Law* (Pluto Press London 2006)

Müllerson R, *International Law, Rights and Politics* (Routledge London 1994)

Nietzche F, *On the Advantage and Disadvantage of History for Life* (Hackett Publishing1980)

Rodin D, *War and Self-Defence* (OUP Oxford 2002)

Schachter O, *International Law in Theory and Practice* (Martinus Nijhoff Publishers Dordrecht 1991)

Shaw M, *International Law* (6th edition CUP Cambridge 2008)

Sidgwick H, *The Element of Politics* (Cambridge Library Collection 2012)

Sorabji R and Rodin D (eds) *The Ethics of War: Shared Problems in Different Traditions* (Ashgate Publishing Limited Aldershot 2006)

Vattel E, 'Droit des gens; ou, Principes de la loi naturelle appliqués à la conduite et aux affaires des nations et des souverains' (translated : *Law of Nations or the Principles of Natural Law Applied to the Conduct and to the Affairs of Nations and of Sovereigns*) (J Robinson London 1797)

Walzer M, *Just and Unjust* (4th edition New York Basic Books 2006)

Weltman J, *World Politics and the Evolution of War* (The John Higgins University Press London 1995)

## Articles

Brownlie I, 'The Use of Force in Self-Defence' (1961) 37 British Year Book of International Law 183 – 268

Clemmons B and Brown G, 'Rethinking International Self-Defence: The UN Emerging Role' (1988) 45 Naval Law Review 217 – 46

De Los Rios C, 'Understanding Political Violence' (2004) 4 (1) LLC Review 29 – 43

Luban D, 'Preventive War' (2003) 32 (3) Philosophy and Public Affairs 207 – 48

MacDougal M, 'The Soviet-Cuban Quarantine and Self-Defence' (1963) 57 American Journal of International Law 597 – 604

McMahan J, 'The Basis of Moral Liability to Defensive Killing' (2005) 15 Philosophical Issues 386 – 405

Mishaan R, 'Introduction – Recent Events in Gaza' (2009) 32 Hastings International Law Review 639 – 43

Ochoa Ruiz N and Salamanca-Aguado E, 'Exploring the Limits of International Law Relating to Self-Defence' (2005) 16 (3) The European Journal of International Law 499 – 524

Quong J, 'Killing in Self-Defence' (2009) Ethics 119 507 – 37
Rodin D, 'War and Self Defence' (2004) 18 Ethics and International Affairs 63 – 8
Ryan C, 'Self-Defence, Pacifism and the Possibility of Killing' (1982) 93 (3) Ethics 508 – 24
Schachter O, 'The Right of States to Use Armed Force' (1984) 82 (5) Michigan Law Review 1620 – 46
Schachter O, 'Self-Defence and the Rule of Law' (1989) 83 (2)The American Journal of International Law 259 – 77
Thomson J, 'Self Defence' (1991) 20 Philosophy and Public Affairs 283 – 310
Van Den Hole L, 'Anticipatory Self-defence under International Law' (2003) 19 American University of International Law Review 69 – 106
Van Der Stoep J, 'Towards a Sociological Turn in Contextualist Moral Philosophy' (2004) 7 (2) Ethical Theory and Moral Practice 133-146
Zohar N, 'Collective War and Individualistic Ethics: Against the Conscription of 'Self-Defence" 21 Political Theory 612 – 22

## Cases

Military and paramilitary activities in and against Nicaragua (*Nicaragua v United States of America*) Jurisdiction and Admissibility 1984 ICJ Reports 392
Oil Platforms (*Islamic Republic of Iran v United States of America*) Judgment 2003 ICJ Reports

## Advisory Opinions

'Legality of the Threat or Use of Nuclear Weapons' Advisory Opinion of 8 July 1996 - General List No 95 (1995 – 1998) (http://www.icj-cij.org/docket/files/95/7515.pdf)

## Reports

Intelligence and Terrorism Information Centre at the Israel Intelligence Heritage & Commemoration Centre 'Anti-Israel Terrorism in 2007 and its trends in 2008' (2008) pp

26 available at:
http://www.terrorism-info.org.il/malam_multimedia/English/eng_n/pdf/terror_07e.pdf

## UN Security Council Resolutions

UN Security Council, *Resolution 799* S/RES/799 (1992) available at:
http://www.unhcr.org/refworld/docid/3b00f15f4f.html
UN Security Council, *Resolution 904* S/RES/904 (1994) available at:
http://www.unhcr.org/refworld/docid/3b00f15e14.html
UN Security Council, *Resolution 673* S/RES/673 (1990) available at:
http://www.unhcr.org/refworld/docid/3b00f13844.html
UN Security Council, *Resolution 242* S/RES/242 (1967) available at: http://daccess-dds-ny.un.org/doc/RESOLUTION/GEN/NR0/240/94/IMG/NR024094.pdf?OpenElement
UN Security Council, *Resolution 338* S/RES/338 (1973) available at: http://daccess-dds-ny.un.org/doc/RESOLUTION/GEN/NR0/288/65/IMG/NR028865.pdf?OpenElement
UN Security Council, *Resolution 1397* S/RES/1397 (2002) available at:
http://www.un.org/News/Press/docs/2002/sc7326.doc.htm
UN Security Council, *Resolution 1515* S/RES/1515 (2003) available at:
http://unispal.un.org/unispal.nsf/0/71B2C135FCA9D78A85256DE400530107
UN Security Council, *Resolution 1850* S/RES/1850 (2008) available at: http://www.refworld.org/cgi-bin/texis/vtx/rwmain?page=publisher&publisher=UNSC&type=&coi=PSE&docid=4950d67c2&skip=0
UN Security Council, *Resolution 106* S/RES/106 (1955) available at:
http://www.unhcr.org/refworld/docid/3b00f13934.html
UN Security Council, *Resolution 487* S/RES/487 (1981)
http://www.undemocracy.com/S-RES-486(1981).pdf

UN Security Council, *Resolution 904* S/RES/904 (1994) available at: http://www.unhcr.org/refworld/docid/3b00f15f4f.html
UN Security Council, *Resolution 673* S/RES/673 (1990) available at: http://www.unhcr.org/refworld/docid/3b00f13844.html

## UN General Assembly Agreements

General Assembly Resolution 3314 http://jurist.law.pitt.edu/3314.htm
UN General Assembly, General Assembly resolution supporting the immediate ceasefire according to Security Council resolution 1860 A/RES/ES-10/18 available at: http://www.unhcr.org/refworld/docid/49917ee92.html

## UN Human Rights Council Report

UN Human Rights Council, UN Human Rights Council : Report of the Special Rapporteur on the Situation of Human Rights in the Palestinian Territories Occupied since 1967 available at http://www.unhcr.org/refworld/docid/461e52b12.htm

## Treaties

Charter of the United Nations (entered into force 24 October 1945)

## Websites

Kattan V 'Gaza: Not a War of Self-Defence' 15 January 2009 Jurist: Legal News and Research available at http://jurist.law.pitt.edu/forumy/2009/01/gaza-not-war-of-self-defense.php
Tanzil Chowdhury, 'Nothing is Something Dangerous' September 6 2012 E-International Relations available at http://www.e-ir.info/2012/09/06/nothing-is-something-dangerous/

# Merchandising and Brand Extension in the Music Industry

Magdalena Borucka

## Abstract

*Use of brand extension in the music industry has become a focal point in current marketing trends. Musicians need to accept that in order to be successful they need to actively engage themselves in this process. This article presents the functional appeal of brand extension and strategies used to extend musicians' brands into different categories of products and services, identifying the most advantageous of them. While brand stretching is omnipresent in all genres of music, this work focuses on hip hop as the most prominent source of successful brand extensions into the areas of entrepreneurship and leadership, with names such as Diddy and Jay-Z being known even by people not familiar with their music. Although artists are still torn between love for music in its pure form and financial success, acceptance for their right to take advantage of their work spreads. It is concluded that multitasking may therefore become a standard not only in the hip hop industry but across all genres of music.*

## I. Introduction

In our everyday life, we tend to rely on brands when we make our purchase decisions without even realising how they affect our final choices. Branding is not a new device used by producers to attract consumers, neither is brand extension. Use of brand extension in the music industry, however, seems to be more and more important. It looks like every artist needs to develop his or her own brand in order to really exist on the market. It seems that musicians cannot just be performers anymore, and that there is a strong pressure for them to become entrepreneurs in their own right.

This research paper will attempt to present the functional appeal of brand extension and the strategies used in the music industry in order to successfully extend musicians' brands into different categories of products and

services. This article will attempt to find the reason for the growing popularity of using merchandising and brand extension in the music industry and what are the most successful techniques used by artists to extend their brands. Other questions that this work will answer are what are the reasons for using brand extension instead of creating new brands, why celebrity-endorsed and celebrity-owned brands are more appealing to consumers than regular brands and why the music industry, especially in the hip-hop sector, has become such a popular playground for brand extension. In order to answer all these questions, this paper will first discuss the main function of branding and the rationale behind brand extension. Next, it will consider the growing popularity of celebrity endorsement, along with brands owned by celebrities, referring to brand personality and lifestyle brands and how they are used in the brand extension process. Following this discussion, this work will try to show the rationale behind using brand extension in the music industry and what possible obstacles musicians might need to overcome. Lastly, this research paper will assess the strategies for a brand extension in the hip-hop industry along with the most prominent examples of the most successful brand extensions.

## II. Brands – functional significance

The word "brand" is derived from the Old Norse word meaning "to burn", as branding was the principal means by which animals were marked by the owners of livestock. Nowadays, branding is still the means by which a business can differentiate its goods. The benefits of this process have become more important as producers started becoming more and more distant from the buyers. As a result, a brand now serves as a means of assuring product authenticity and, most importantly, its quality. They act as an

assurance that the characteristics, functions and features of the branded product will remain the same for every item.[1]

Brands identify the goods and services of one producer and differentiate them from those of competitors. Particular brand equity is exhibited when consumers respond more favourably to the marketing actions of one producer than they do to his competitor. The image of the producer in the consumers' minds is the basis for this brand equity.[2]

What is especially important is that a brand is much more than only the name or the object that it identifies. Consumers buy products for many more reasons than just their quality. The main difference between products and brands, therefore, is that products are made in factories, whereas brands are made and exist only in consumers' minds. Hence, the creation of a strong image and identity is a significant part of brand management. This is why it is not the brand on its own that is the real asset, but the brand loyalty created as an intellectual concept.[3] Brand loyalty is a type of affective commitment – if this commitment is high, it could motivate consumers to continue the relationship between the brand and themselves.[4]

It needs to be emphasized, however, that the word "brand" does not refer solely to consumer products anymore. It includes places, companies, industrial products and services but also people, such as movie stars, politicians and musicians.[5]

---

1 Richard Cree, 'Papa's Got A Brand New Brand: An Investigation of Brand Strategy in the UK Music Industry' 4 (2) The International Journal of Urban Labour and Leisure, 4 <http://www.ijull.org/vol4/2/cree.pdf> accessed 23 August 2012.

2 Kusum L Ailawadi, Kevin Lane Keller, 'Understanding Retail Branding: Conceptual Insights and Research Priorities' (2004) 80 Journal of Retailing 331, 332.

3 Cree (n 1) 4-5, 8.

4 Tsan-Ming Choi et al, 'Fast Fashion Brand Extensions: An Empirical Study of Consumer Preferences' (2010) 17 Brand Management 472, 474.

5 Cree (n 1) 1.

## III. Brand extensions

Establishing a new brand name in international markets requires a big investment, sometimes well over $100 million, which makes it beyond the capability of most companies. This initial cost is not the only issue, companies need to face the struggle to broaden the market base of their products, avoid the price competition and differentiate products.[6] Launching new products is then a business activity that is connected with high risk and high costs. Since success rates are usually below 50%, many companies tend to resort to brand extension strategies in order to make their new offers more attractive for consumers.[7]

Brand extension is usually defined as the use of an established brand name in order to enter new product classes or categories. It can be classified into two general forms – horizontal and vertical extensions. Horizontal extensions involve the application of an existing brand name to a new product. It can be either a product in a similar class or in a category new to a producer. Vertical extension refers to introducing a similar brand in the same product category, but with a different quality and price.[8]

Between 1977 and 1984, 40% of the brands introduced into the United States supermarkets were brand extensions. The main advantage of this strategy is the reduction in product introduction risk.[9] It needs to be noted, however, that while brand extension is now one of the most frequently employed branding strategies, it is not risk-free. The failure rates of brand extensions in many high-tech

6 Ashley Lye, P Venkateswarlu, Jo Barrett, 'Brand Extensions: Prestige Brand Effects' (2001) 9 (2) Australasian Marketing Journal 53, 53.

7 Eva Martinez, Teresa Montaner, Jose M Pina, 'Brand Extension Feedback: The Role of Advertising' (2009) 62 Journal of Business Research 305, 305.

8 Choi (n 4) 473-474.

9 Mary W Sullivan, 'Brand Extensions: When To Use Them' (1992) 38 (6) Management Science 793, 793.

and fast moving industries are close to 80%.[10] While chances of success are higher than in the regular new brand introduction strategy, it is never possible to precisely predict the market reaction and consumers' needs.

Another type of brand expansion is brand stretching. While the difference between brand extension and brand stretch is often difficult to define, brand stretch generally involves stretching brand names way beyond the original core product area and as such, it generates the greater risk.[11]

The main advantage of all these types of brand expansion is its speed and lower costs. Building an entirely new brand with an unknown name takes both time and investment. Obviously, simply extending the same name also places some limitations on diversification and may in turn lead to a lack of creativity and innovation. The decision as to when to use the same brand name depends not only on the core brand image but also on internal company capabilities. It is very common for companies to be overconfident and stretch the brand too far outside of its original industry base. The best attitude seems to be a slow systematic progression to different but related industries rather than rushing from one end to another.[12]

While the risk of introducing brand extension always exists, consumer acceptance of a proposed extension is higher if the perceived quality of the brand is high. Another factor is a perceived match between product categories, especially if the skills seem to be transferable and the products are complementary.[13] The success of brand

---

10 Ali Besharat, 'How Co-Branding Versus Brand Extensions Drive Consumers' Evaluations of New Products: A Brand Equity Approach' (2010) 39 Industrial Marketing Management 1240, 1241.

11 Mike Bastin, 'How Far Can You Stretch Your Brand?' *China Daily* (18 November 2011) <http://usa.chinadaily.com.cn/opinion/2011-11/18/content_14118755.htm> accessed 23 August 2012.

12 ibid.

13 Lorraine Sunde, Roderick J Brodie, 'Consumer Evaluations of Brand Extensions: Further Empirical Results' (1993) 10 International Journal of Research in Marketing 47, 47.

extension depends on how consumers perceive the extended product and whether it can satisfy their needs. Other factors that may affect the consumer behaviours towards brand extensions include self-image, brand loyalty, and brand concept, consistency and involvement.[14]

Brand extension is a form of permanent and free advertising. If the product that was initially advertised has a unique and coined name, every addition to the basic product line to which the name is attached, makes the name grow stronger.[15] It is also a balancing act and producers need to be very careful not to extend their brands too far as this may in turn become harmful to the brand. Careless licensing and attaching a logo or trademark to all kinds of unrelated products may destroy the integrity of the brand.[16]

## IV. Celebrity endorsement

Celebrities are a common feature in the contemporary marketplace – they often become faces of consumer products, brands and organisations. Brands make use of well-known and liked celebrities by leveraging their equity.

> By pairing a brand with a celebrity, a brand is able to leverage unique and positive secondary brand associations from a celebrity and gain consumer awareness, transfer positive associations tied to the celebrity onto the brand, build brand image and ultimately enhance the endorsed brand's equity.[17]

Celebrity endorsement is then a very useful tool used to improve communication with potential consumers by

---

14 Choi (n 4) 472, 474.

15 Victoria Slind-Flor, 'Money and Mayhem' [2007] Intellectual Asset Management 15, 15.

16 ibid 17.

17 Jasmina Ilicic, Cynthia M Webster, 'Effects of Multiple Endorsements and Consumer-Celebrity Attachment on Attitude and Purchase Intention' (2011) 19 Australasian Marketing Journal 230, 230.

creating connections between the advertised brand and consumers. The role of celebrity endorsement cannot be underestimated as it facilitates breaking down cultural barriers, helps to reposition brand images and, as a result, improves sales of the endorsed product. Even announcing an endorsement contract affects stock returns. An example of that could be an announcement of Tiger Woods' endorsement deal which increased Nike's stock value.[18] There is no doubt then that celebrities are a worthwhile investment.[19]

The power of celebrities is not limited to selling products and brands; they influence popular culture and public life, which in turn has an impact on consumer perceptions and attitudes. This is why companies try to attract celebrities for various campaigns within different product categories. Such an overexposure, however, does not help a brand since consumers perceive celebrities endorsing multiple product categories as less credible.[20] Various researches show that the image associated with a celebrity is transferred onto the brands he or she endorses, and then from the celebrity to consumers through their brand selection, which communicates their self-concept that, in turn, forms a self-brand connection.[21]

The most important qualities affecting the effectiveness of celebrity endorsement include their personal attractiveness, likeability, familiarity, believability, expertise and credibility. Consistency of the celebrity endorser's image with the image of the brand or product is also an important factor. If a celebrity matches the product and brand, he or

---

18 Christopher R Knittel, Victor Stango, 'Celebrity Endorsements, Firm Value and Reputation Risk: Evidence from the Tiger Woods Scandal' *Massachusetts Institute of Technology* (25 August 2010) 4 <http://www.econ.ucdavis.edu/faculty/knittel/papers/Tiger_latest.pdf> accessed 24 April 2013.

19 Ilicic (n 17) 230, 230.

20 ibid.

21 ibid 230-231.

she is perceived as more credible and persuasive so it is important that the spokesperson's characteristics are relevant for the attributes of the brand.[22] It is, therefore, important for both the brand and the celebrity, to pick up the match wisely since consumers perceive celebrity endorsers as believing in the brands they support and look for reasons for the endorsement. This is why they react more favourably if a celebrity endorses only one product. The persuasive power grows if an endorser is seen as an expert in the given category, making the decision to buy the brand easier.[23]

The reason why celebrity endorsement works so well for selling products is that the consumer attachment to a celebrity affects consumer attitude toward the brand and influences purchase intentions. According to the attachment theory, the basic human need to make strong emotional attachments with significant others results in the relationship between a consumer and a brand. In this scenario then a brand acts as a link to a significant other, i.e. to a celebrity to which a consumer feels emotionally attached. Consumer attachment to a celebrity leads to a higher attitude towards the brand endorsed by a celebrity because consumers see celebrities as their role models or at least people they would like to be associated with. If a celebrity endorses a brand, consumers see that as their rational decision and perceive them as believing in the endorsed brand or product. Buying such products is a way of getting closer to the celebrity. They may not have the same lifestyle but at least they wear the same clothes, use the same perfumes or drive the same cars.[24]

### *A. Celebrity-owned brands*

It seems that not only producers noticed how big the power of celebrities is but also celebrities themselves are

22 ibid 231.

23 ibid.

24 ibid 232, 235.

ready to take advantage of their persuasive power. In recent years, more and more celebrities decided to use their name as a brand on its own right and started developing ranges of products and services. It is no longer about brands extending into different categories and using celebrity endorsement as an effective marketing and promotional tool. Rather about celebrities using their names as brands and extending themselves into completely new categories. It usually starts with the support from some other brand, that already exists in the market, but being a silent partner in the business partnership sealed with a celebrity. The use of the professional expertise combined with the known name, brings astounding results.

In the past when a celebrity decided to lend his or her image to further a product it was seen as a sell-out. Nowadays, however, getting behind a product and using the fame to support it is seen as a good example of entrepreneurship.

> The kind of active investment we are seeing from celebrities like 50 Cent, Ashton Kutcher, Sean Combs and Leonardo DiCaprio is marked by market research, personal engagement in the product and an ownership stake. Sure, a lot of these glitzy moguls-in-the-making have business managers and research teams, but compare their entrepreneurial endeavours to what they could be doing – renting out their likenesses to underwear ads – and you have got to admit they are a little bit more engaged.[25]

It seems that it is not enough to be an artist anymore; being a celebrity investor is now an essential ingredient if a celebrity wants to be seen as successful. 'It is hip to be an entrepreneur, maybe more so than ever, and being hip while making money could be reason enough to become an

---

25 Karsten Strauss, 'Celebrity Entrepreneurs on the Rise?' *Forbes* (16 May 2012) <http://www.forbes.com/sites/karstenstrauss/2012/05/16/celebrity-entrepreneurs-on-the-rise/> accessed 23 August 2012.

investor'.[26] The list of celebrities becoming entrepreneurs is endless. Lady Gaga, after engaging in many endorsement deals, such as the one with MAC where she endorsed her own line of cosmetics, has now become a major shareholder in Backplane – a platform connecting music and sports stars with their fans through social networks. Ashton Kutcher founded a venture firm, A-Grade Investments that aims at scanning for tech start-ups to which it could lend money. Justin Bieber has stakes in the messaging platform Tinychat, the social-app Stamped, as well as Spotify and Sojo Studios.[27] These are just a few examples which are only a tip of the celebrity investments' iceberg and with the current trends, it seems like this iceberg can only grow bigger in the future.

## *B. Brand personality*

What is particularly interesting is how it is possible that celebrities with no professional expertise will manage to become successful in endorsing other brands or promoting their own. Celebrities can convince consumers to buy particular products even though they did not take part in the creative process and all they did was lend their face to support it. In other cases, they might even be active in the creation of a product but it is still quite a mystery why consumers trust a good singer to also be a good engineer.

The answer to these questions may be in the theory of brand personality. Brand personality is defined as the set of human characteristics associated with a given brand, including gender, age, human personality and socioeconomic class. [28] By associating brands with certain human characteristics, consumers are anthropomorphizing them. These personalities differentiate brands in consumers' minds even if consumers are not able to articulate these differences. 'The colourless, odourless and tasteless vodka product

26 ibid.

27 ibid.

28 Choi (n 4) 476.

category is a case in point. One vodka may be seen as "cool" and "hip", whereas another may be described as "intellectual" and "conservative"'.[29]

This process of association is indeed a very strong marketing device. Consumers, who identify themselves with a particular personality, will have a greater preference for the brands matching this dimension.[30] Brand personality, therefore, plays a pivotal role in attitudes of consumers and their purchase intentions. Consumers are familiar with the rugged persona associated with Harley-Davidson and Marlboro, the youthful excitement of Pepsi or sophistication of Mercedes Benz and they react to them accordingly. This is why it is crucial for producers to understand their target audience and try to build the personality of their brand that would match the one of the potential consumers.[31]

Brand personality has a few positive effects as it influences consumer preferences, elicits their emotions, encourages self-expression and stimulates active information processing. What is the most important, though, is that brand personality not only increases levels of loyalty and trust, but also influences brand attitudes and associations while at the same time providing a basis for product differentiation. By creating favourable brand associations among consumers, which they would regard as satisfying their needs, a favourable brand personality is created. Brand personality is then a non-product-related attribute of a brand that has the power to detract or add to a consumer's impression of the brand.[32]

As it was explained, brand personality gives a brand human-like features. In order to attract consumers,

---

29 Bernd Schmitt, 'The Consumer Psychology of Brands' (2012) 22 Journal of Consumer Psychology 7, 11.

30 Choi (n 4) 476.

31 Traci H Freling, Jody L Crosno, David H Henard, 'Brand Personality Appeal: Conceptualization and Empirical Validation' (2011) 39 Journal of the Academy of Marketing Science 392, 392.

32 ibid 393, 395.

producers need to manage their brands so that they appeal to the target audience. Creating a brand as a person may be a lengthy and complicated process. Using a celebrity is then an alternative route to this goal saving time, money, effort and resources. The features associated with a particular celebrity are easily transferred to the product so the brand adapts the personality of the chosen celebrity. The same logic works in the case of celebrity-owned brands – if a celebrity decides to launch a product, this product will be seen as having the same features as the celebrity in question. Using celebrities to promote products is then not only about attracting the attention by using a famous face, which would only have a short-term effect. It is about matching the personalities, which could turn out to be profitable in the long-term.

The important part is then the right choice of celebrity so he or she matches the product or brand. Using a celebrity who is seen as tacky, controversial and is associated with a hard-partying lifestyle, such as Paris Hilton, for promoting a high-end product could only backfire and tarnish the luxury brand image. She is, however, very successful in promoting her clothing line, perfumes, handbags, watches, stationery, bedding and footwear.[33] The reason for this is probably that she is known for her love of luxury and splendour so consumers looking for products having these features, are inevitably choosing the products she promotes. Others may actually like the product but try to hide the label in order not to be associated with such a persona.

On the other side of this spectrum, there are celebrities who are associated with high quality products, prestige and respect. A good example could be Bruce Willis who is an award-wining Hollywood actor, and is also engaged in his own business activities, such as the Planet Hollywood

---

33 Jen Ortiz, 'Surprise? Paris Hilton Earns Over $10M a Year From 17 Different Product Lines' *Business Insider* (1 June 2011) <http://articles.businessinsider.com/2011-06-01/entertainment/30076238_1_piers-morgan-paris-hilton-product-lines> accessed 23 August 2012.

restaurant chain. It was not a surprise when he agreed to endorse Belvedere's Sobieski vodka, which is known for its quality and is portrayed as a top shelf product in its category. The famous actor did not agree only to become the ambassador of the product, which is now advertised as 'designed by Bruce Willis', but he also obtained a 3.3% stake in the company in exchange for signing a four-year contract to promote the Sobieski vodka brand world-wide.[34] Such an endorsement contract benefits both sides – it strengthens the prestigious image of the Sobieski vodka by associating it with a respected celebrity, but it also reinforces Bruce Willis' image as a good businessman. If the sides of this deal were not equally respected, i.e. if the actor agreed to support a low-end product, instead of improving the appeal of the promoted brand, he would tarnish his own. A good example of such a disaster could be Donald Trump advertising "Trump Steaks", the decision more than questionable. Sometimes such a mismatch may only raise some eyebrows like in the case of "Sex and the City" star – Kim Cattrall promoting Super Mario for Nintendo DS. Nevertheless, obtaining the right balance between the product/brand and the celebrity chosen to endorse it is crucial since a mistake may negatively affect both.

### *C. Lifestyle brands*

Another consideration when trying to understand the power of brand extension is a creation of "lifestyle brand". Harley Davidson is a lifestyle brand since the consumers associating themselves with this brand use the product as a lifestyle, i.e. they associate themselves with the images linked to the brand. In other words, it is not only about a purchase decision, being a Harley Davidson consumer means living a certain lifestyle. Products with such a status can extend the

34 Amelie Baubeau, David Kesmodel, 'Bruce Willis Sees Spirits in Equity Deal With Belvedere' *The Wall Street Journal* (23 December 2009) <http://online.wsj.com/article/SB10001424052748703478704574611690552812758.html> accessed 23 August 2012.

brand to all kinds of different areas and categories that consumers might buy. Sometimes the extension might go into fields completely unrelated to the original product. For Harley Davidson it resulted in extending the brand not only to t-shirts, leather jackets, helmets and beer but also perfume, cribbage boards, wedding-cake toppers, condoms or Barbie dolls.[35]

Creating a lifestyle brand is, therefore, a very worthwhile undertaking. Once a brand is established, the possibilities for the extension are endless. However, the process of giving the brand the lifestyle dimension might be lengthy and it might not always succeed since not all brands have the potential to have such an appeal. Some industries, however, are particularly prone to welcome lifestyle brands. The music industry is one of such examples. Music is not only about the sound, it is about certain flair, the aura that surrounds the artist and with which the audience wants to be associated. It is easy to manipulate consumers' attitudes and make them believe that whatever their favourite artist has is a must-have. This is why establishing a brand instead of just performing is the focus of the music marketing nowadays. It is an unstoppable machine with unlimited power, which is exploited by the biggest players in the show business game.

## V. Brand extension in the music industry

There is no doubt that a musician's name is one of his or her most important assets. It encompasses the reputation an artist built around it and what consumers use as a point of reference to identify the artists they enjoy. An artist is able to protect this asset by obtaining registered trademark rights to a word, series of words, stylised words or logo. Artists need to consider in advance what they want to do with their brand when applying for a trademark. If they decide to extend the brand, they need to ensure it is protected with regard to all relevant goods or services. A

35 Slind-Flor (n 15) 15.

brand name can also be protected even without a registration if an artist has managed to establish a reputation associated with a given name over a significant period.[36]

Trademark protection gives an artist a legitimate right to use his or her brand with regard to certain goods or services. It also guarantees a right to prevent other parties producing either music or merchandise from using the artist's mark, as it could lead to a loss of revenue for the given artist or damage to his or her reputation. Another merchandiser selling, for example, t-shirts of poor quality with the artist's name or logo, not only takes over the revenue that this artist could gain. He also potentially tarnishes the positive association with the artist's brand name if consumers are disappointed with the quality of the product they purchased not knowing that it was not in fact a product authorised by their idol. Another significant benefit of a trademark registration is that once registered, the brand can be used to exploit other valuable revenue channels, including merchandising, licensing or sponsorship.[37]

Although brand stretching is a very popular marketing strategy, there used to be a certain reluctance to take advantage of it in the music industry. The money-art dichotomy and the tension between artistic integrity and popular success were the main forces stopping artists from exploiting their fame in order to earn more money. This attitude, however, has shifted as it became obvious that artists were simultaneously the product, producer and brand.[38] What used to be seen as a sell-out is now seen as a good decision. Artists do not have to hold back and refuse good business deals out of fear of being seen as "not artistic" enough. In fact, they pursue their careers and get involved with various industries like the best businessmen, with their

36 Georgina Harris, 'The Sugababes: the trademark rights associated with band names' (2010) 21(5) Entertainment Law Review 165, 165-167.

37 ibid 167.

38 Krzysztof Kubacki, Robin Croft, 'Markets, Music and All That Jazz' (2011) 45 (5) European Journal of Marketing 805, 805.

fans patting them on the back and happily spending their money on products endorsed by their idols.

It is estimated that in 2004, Ozzy Osbourne earned US $35 million from concert sales and another US $15 million from merchandising. It shows how big a business merchandising is in the music industry and how important revenue stream it constitutes. Stretching a brand name beyond the usual classes such as video/sound recordings and entertainment services into fields including obvious merchandise goods such as clothing, stickers or mugs is usually the first step of the brand extension strategy. Next step is to extend the brand into more unusual items such as dolls, as registered by the Spice Girls, or suntan lotion and perfume, as registered by the Pussycat Dolls.[39]

The music industry is very peculiar though, with its self-obsession and desire to stand apart from all the other markets. While there is still some reluctance to fully embrace the principles of branding, it cannot ignore modern corporate pressures. Four of the big five major record companies – Sony, BMG, Warners and Universal - are owned by multinational companies with no music interests but with important brand assets. Even though these corporations tend to leave their music subsidiaries to run themselves independently, there are still pressures to integrate the successful management model across all its divisions.[40]

The music industry is especially also interesting because 'musicians' identities are multiple and fluid, adapted to social conditions encountered in everyday life, and inseparable from the art work'[41] and as such are transferred to their music. They are not, however, free to be whoever they want to be. The society and the values it believes in determine what kind of people can become musicians, what

39 Harris (n 35) 167.

40 Cree (n 1) 4.

41 Kubacki (n 35) 806.

types of musicians can be recognized and what social position is given to them. Musicians need to be very careful since they are not only judged by their performance, but they are also presented with various opportunities to express their identities in their everyday working lives and everything they decide to do, may affect their overall image.[42]

The universal routes that every artist needs to follow in order to become a brand on his/her own right are now an inevitable part of an effective marketing strategy. An artist starts out as just another new and unknown act and depends heavily on solid product he or she is trying to sell, as well as good marketing in order to make an impact. Over time, an artist may start being recognized as a source of consistently high quality and each act is expected to be as good as the one before. Gradually an artist manages to develop into the personality of a star and then eventually, those with the strongest brand reach the iconic status. Artists, therefore, just like regular branded products, need to go through the entire process of brand creation. They start as an unbranded product, then brand working as a reference point, brand as personality, and they end up as icons. At this stage, identities of many of the biggest and most successful stars begin to become tied up with the identity of their label. The reason for that is often that they outsell other artists so that they become the major name bringing the highest revenues or that they simply start their own labels and start promoting new and upcoming artists.[43]

Because of technological innovation and digital revolution, all the relationships between stakeholders in the music industry are being redefined. Due to this change, all music producers, but especially the major record companies, have to look for new ways of creating and sustaining competitive advantage for their releases. One of the potential solutions is the development of strong brands to

42 ibid.
43 Cree (n 1) 7-8.

which consumers would feel strongly attached. Such a brand not only guarantees a bigger chance of success of new releases but it also allows sustaining a price differential over competitors, based on the perceived extra value of the brand. A strong brand name gives an unquestionable advantage. When we think about artists such as Madonna, Michael Jackson or Lady Gaga, we do not only think of recording artists, these are images connecting a number of different areas of culture. [44] These associations attract a bigger audience which is exactly what record companies are looking for.

The need to develop a strong brand which can later be stretched originates in the said digital revolution as, according to the Recording Industry Association of America, illegal music downloads were responsible for a 23% decrease in sales of music CDs worldwide only between 2000 and 2006. It was reported that music sales fell from 449.2 million in 2007 to 360.6 million in 2008. [45] Looking at these numbers, it should not come as a surprise that artists need to start looking for alternative sources of income. Illegal downloading created a major dent in profits, forcing the artists to look for new and innovative ways to ensure they can remain entertainers without having to search for a new profession in order to pay their bills. [46]

Using brands to boost the income is certainly not a new device and different strategies were used in order to take advantage of brands' power. Licensing tracks for use in movies, television and advertising helped Moby's 1999 album "Play" sell over 10 million copies worldwide even though it underperformed commercially upon release.

---

44 ibid 13.

45 Julian Gratton, 'How the Music Industry Is Using Brands and Advertising to Plug the Gap Left By Illegal Downloads' (*Red C,* 20 October 2009) <http://www.redcmarketing.net/blog/marketing/how-the-music-industry-is-using-brands-and-advertising-to-plug-the-gap-left-by-illegal-downloads/> accessed 23 August 2012.

46 ibid.

Almost all set to be a total flop; it is now the number one selling electronica album of all time.[47] Similarly, 'Too Close', a song by Alex Clare released in the United Kingdom in 2011, without much initial outcome, has become international success and has now reached over 43 million views on YouTube after being featured as the soundtrack to Microsoft's advertisement for Internet Explorer 9 in 2012.[48] Another route the artists may follow in order to increase the revenue are sponsorship deals. Every concert is sponsored by some often completely musically unrelated brand, with the major ones for the sponsorship deals being Pepsi, Coca-Cola or Heineken.[49]

Brands and music are working together to generate new revenue streams and they succeed in these endeavours. It is estimated that to date, brands devote 5% of their advertising budgets to music, and brand managers strongly believe that music is an effective way of building brand awareness. At the world's biggest music industry trade fair – MIDEM 2009, key players of both music industry, such as Sony BMG and EMI, and brands agencies, including Coca Cola, met behind closed doors to discuss and explore practices for the most effective music-brand collaborations.[50] In March 2013 Universal Music Group announced a new global partnership with Bang & Olufsen, the Danish provider of high-end audio and video products. The official goal of this collaboration is to allow music lovers to 'experience recorded music in the highest possible quality'.[51] The strategic rationale seems to be achieving a perfect blend of the creative image of UMG and the high quality and excellence associated with Bang & Olufsen.

---

47 ibid.

48 http://www.youtube.com/watch?v=zYXjLbMZFmo&ob=av2n

49 ibid.

50 ibid.

51 http://www.universalmusic.com/corporate/detail/2460

While some of the brand extensions in the music industry seem to be fairly rational, like Britney Spears promoting her own line of perfumes or Nicki Minaj, known for her eccentric looks, promoting her own line of makeup for MAC, others are astonishing, to say the least, with JLS condoms being just one of such examples.

### *A. Hip-hop industry*

While the brand extension is a phenomenon occurring in every sector of the music industry, it seems like the hip-hop music industry is the breeding ground for the biggest amounts of strong brands extending into many different categories of products. The origins of hip-hop can be found in the neighbourhoods of poor black and Latino families in New York City. From the days it was born in the 1970s when it was mostly underground and spread in the streets to the present day, it has evolved into a multi-billion dollar global phenomenon and its contribution to the economy of the United States is estimated to be in the billions of dollars. Its growth was rapid and 'by the end of the 1980s it became the single most potent global youth force in a generation'.[52]

Hip-hop is not only the music, it is a culture and the core four elements that emerged in the 1970s include MCing, DJing, breaking and graffiti art. This culture is reflected not only in the music, but also in the clothing, art, film, literature, social advocacy, entrepreneurialism and politics. It provides a certain lifestyle that incorporates various groups of products, which can be associated with it. 'Hip-hop is the thread that holds together the fabric of today's urban-youth culture and it touches a multitude of

52 Valerie L Patterson, 'Engaging Hip-Hop Leadership: Diversity, Counter-Hegemony and Glorified Misogyny – (Free-Style Version)' 1 <http://www.ipa.udel.edu/3tad/papers/workshop2/patterson-newman.pdf> accessed 23 August 2012.

industries – from entertainment to apparel to marketing to technology. To put it bluntly, hip-hop is big business'.[53]

Hip-hop culture proved to be very demanding even for the artists and in the mainstream; it is not enough anymore to be a rapper. Nowadays it is expected that artists will also engage in entrepreneurship. The role model consists of a rapper turned chairman, with Shawn "Jay-Z" Carter, Sean "Diddy" Combs, Dana "Queen Latifah" Owens, Curtis "50 Cent" Jackson, to name only a few, being the examples of successful transition from hip-hop to corporate culture.[54]

Since rappers are often too controversial to be sponsored by a mainstream brand, they often create their own. Eminem, for instance, created Shade 45 Radio Channel, Shady Games, Shady Ltd. Clothing and Eight Mile Style LLC, stretching his brand name into different sectors while still maintaining the integrity of the core brand.[55] Creating their own brands is also a very productive move in terms of free advertising. It is very easy for musicians to incorporate brand names into lyrics and since they are the owners of the brand, they do not have to pay extraordinary amounts of money for such advertising. 'After all, if millions of people are downloading the track and singing along to it... what better way is there to get your brand on the lips of a nation?'.[56]

## VI. Brand extension strategies

When discussing brand extension strategies in the music industry, there are two names that can never be overseen – Jay-Z and Diddy. In the May 2007 issue of Ebony Magazine, it listed the most influential 'Blacks in America' and placed both of them in the business category

53 ibid 1, 3.
54 ibid 4-5.
55 Gratton (n 44).
56 ibid.

presenting them as president/CEO Def Jam and founder of Sean Jean and Bad Boy, respectively.[57] In a recent Forbes article, they were both featured as potential future billionaires.[58]

### *A. Shawn "Jay-Z" Carter*

Shawn Carter, known as Jay-Z, has been regarded as a very successful entrepreneur for a while now and his individual fortune is estimated to be around US $450 million. Most recently, Jay-Z and his wife Beyoncé topped Forbes' 2012 World's Highest-Paid Celebrity Couples list.[59] He has sold more than 50 million albums, won 10 Grammys and had more number one albums on America's Billboard chart than any other solo artist. Nevertheless, he is also an extraordinarily successful businessman.[60] Jay-Z holds stakes in a very wide range of businesses, including Brooklyn Nets, ad firm Translation, cosmetics company Carol's Daughter and the 40/40 Club chain. In 2008, he signed 10-year deal with Live Nation worth $150 million.[61] He has partnerships with Coca-Cola, Budweiser, Reebok, Microsoft and Hewlett Packard.[62]

---

57 Patterson (n 49) 5.

58 Zack O'Malley, 'Who Will Be Hip-Hop's First Billionaire?' *Forbes* (22 September 2011) <http://www.forbes.com/sites/zackomalleygreenburg/2011/09/22/who-will-be-hip-hops-first-billionaire-jay-z-diddy-dr-dre-birdman-50-cent/> accessed 23 August 2012.

59 Nida Dar, 'Jay-Z, Beyonce to Become Forbes Highest-Paid Couple' (*Business Recorder*, 9 August 2012) <http://www.brecorder.com/arts-a-leisure/261-life-a-style/72829-jay-z-beyonce-to-become-forbes-highest-paid-couples.html> accessed 23 August 2012.

60 Simon Hattenstone, 'Jay-Z: The Boy from the Hood Who Turned Out Good' *The Guardian* (20 November 2010) <http://www.guardian.co.uk/music/2010/nov/20/jay-z-interview-simon-hattenstone> accessed 23 August 2012.

61 O'Malley (n 55).

62 John Jurgensen, 'The State of Jay-Z's Empire' *Wall Street Journal* (22 October 2010) <http://online.wsj.com/article/SB10001424052702304741404575564092478617462.html> accessed 23 August 2012.

He started his music career in 1996 but despite his efforts, no major labels wanted to sign him. Together with two partners, he decided to form and independent label Roc-A-Fella Records. Following his commercial success, the label entered a joint venture with another major company Def Jam. By 2005, Jay-Z became the president and CEO of Def Jam.[63] Currently, the label evolved into The Island Def Jam Music Group and is comprised of Island Records, Def Jam Recordings, and Mercury Records. Representing artists such as Justin Bieber, Kanye West, Mariah Carey, Rihanna, Bon Jovi and Duffy, it is now recognized as one of the most successful labels in the industry.[64]

In late 1990s, Jay-Z launched his clothing line Rocawear. On the brand's official website, we can read that

> [it] represents a borderless, global lifestyle. With Roc-A-Fella Records serving as the initial launch pad, The ROC realized its prowess in creating culture far beyond the realm of music. Hence, the birth of the apparel company, Rocawear. Like Roc-A Fella Records, Rocawear quickly staked its claim in hip-hop history becoming the destination brand for street savvy consumers.[65]

The brand was, and still is, very successful, bringing US $700 million in sales annually. This allowed Jay-Z to sell his clothing label in 2007 for US $204 million, while still maintaining creative and operational control.[66]

Shawn Carter ventured into sports bar chains and lounges when he opened his 40/40 Club and a Greenwich Village bistro, the Spotted Pig. He created them as an extension of himself and delivered brand recognition by

---

63 ibid.

64 http://www.islanddefjam.com/

65 http://rocawear.com/

66 Rasha Proctor, 'Jay-z Brand Extension' (14 December 2011) <http://www.rashaproctor.com/wp-content/uploads/Jay-Z-Brand-Equity.pdf> accessed 23 August 2012.

mentioning the club in his songs.[67] He also retained creative control when he entered into a deal with Reebok and became the first non-athlete to have a sneaker line with this company. In 2006, he agreed to help with designing Budweiser Select campaign, shattering the myth of Champagne-drinking-rappers. He uses all these deals to shape his own public image and this is why he carefully considers each of them before rushing into any decisions.[68]

Jay-Z created his brand based on high quality product he provides, namely, his music. His brand is not heavily dependent on his own personality but he maintains its appeal by entering into various business ventures. Instead of creating a different stage persona for himself, he is now a successful and respected businessman, taken seriously by all his partners.

### *B. Sean "Diddy" Combs*

Sean Combs, who has been known variously as "Puffy", "Puff Daddy", and "P.Diddy", in August 2005 decided that henceforward he wanted to be known as "Diddy". This brought about proceedings against him in the UK, as there was already one Diddy in the music game.[69] It does not look, however, that this dispute could in any way harm his already established brand. His net worth is estimated to be US $500 million with music making up less than 20% of his revenue. He has stakes in his own record label Bad Boy, clothing line Sean John and Enyce, marketing firm Blue Flame and most recently, he entered into a deal with Diageo's Ciroc vodka which brings him double-digit millions each year as he receives an annual cut of profits as well as a percentage from every bottle sold.[70]

---

67 ibid.

68 Jurgensen (n 59).

69 Michele Boote, 'Diddy Do It?' (2007) 197 Trademark World 18, 18.

70 O'Malley (n 55).

His success is not coincidental, it was all a carefully planned strategy he followed in order to establish his own brand. He started from establishing his own record label Bad Boy Records that then was turned into a multifaceted entertainment powerhouse – Bad Boy Worldwide Entertainment Group, of which he is the CEO and founder. Apart from its core function as a music label, it now encompasses a broad range of businesses including recording facility, music publishing, television and film production, artist management, apparel and restaurants. His success in music has translated into a collection of businesses, not limited only to entertainment but also fashion and fragrance. The fashion brand Sean John, which debuted in 1999, turned out to be a tremendous success with annual retail sales in the United States exceeding US $525 million. Combs made sure that his brand is not only popular but also luxurious and of the best quality and in 2004 he was honoured by the Council of Fashion Designers of America as Men's Wear Designer of the Year. Following the success of the clothing line, he decided to form a partnership with Estée Lauder Companies, and has launched three fragrances, of which two won the FiFi awards in the best men's fragrance prestige category.[71]

In 2007, Sean Combs signed a deal with Diageo to oversee and manage all marketing and branding initiatives for Ciroc Vodka. On its website we can read that the reason for that partnership was that 'the company understood that [he is] not just a celebrity endorser, [he is] a brand builder. [He is] a luxury brand builder'.[72] Even though many can see him as overly confident egocentric, his successful businesses only prove that he really is the person he claims to be.

His trust in his own capabilities and the power of the brand he now rightfully represents is reflected in everything he does. He launched a series of headphones, advertised by

71 http://www.seanjohn.com/
72 http://www.ciroc.com/

a slogan 'Diddybeats embody the celebration of the finest in music, luxury and pop music. In other words, these incredible sounding in-ear headphones represent all things Diddy'.[73] Egocentric or not, he certainly knows how to run the business.

### *C. Musical collectives*

What is especially popular, and what does not seem to be as prominent within any other genre, is the creation of musical collectives. Under one brand name, they unite not only rappers but also DJ's, producers, songwriters and often other non-musical members, such as photographers, filmmakers, designers or bloggers. They all cooperate in order to promote each other as individuals, as well as the umbrella brand they create together.

One of the most popular hip-hop collectives of all times, the Wu-Tang Clan is also the most revolutionary rap group of the mid-'90s but their music is only part of their influence. More importantly, the way they decided to operate completely changed the standard concept of a hip-hop crew. The Wu-Tang Clan emerged in 1993 as a loose congregation of nine MCs. They decided that instead of releasing one album after another, they would establish their brand with their debut album and then engage into as many side projects as possible. In this process, thanks to the initial attention brought by the strong first release, each individual member became a star on his own right as well as received individual royalty cheques.[74]

The Wu-Tang brand became very powerful and in short time they were introducing new associates as a kind of a brand-name franchise. While many argue that the Wu-Tang brand was suffering from inconsistency and overexposure, the commercial success was indisputable. They released a video game, a comic book and a clothing line; they also

73 http://www.diddybeats.com/

74 http://www.wutang-corp.com.

operate a record label, fashion house and film production company, all under the Wu-Tang brand.[75]

In March 2011, yet another collective emerged and attracted a lot of public attention. The Los Angeles-based hip-hop collective Odd Future (also known as Odd Future Wolf Gang Kill Them All and OFWGKTA) seems to follow the same strategy as their predecessors. The group consists of skaters, artists, photographers and friends, all living in Los Angeles. Through internet releases, they managed to create a very strong fan base, which in turn brought about the attention of the mainstream. Under the Odd Future umbrella brand, each artist releases his own tracks. Thank to this strategy, with every critically acclaimed release of an individual artist, the Odd Future brand becomes stronger and that appeal is then transpired onto each member of the collective. As in the case of Wu-Tang Clan, the strategy proved to be reliable.[76]

In only one year, Odd Future came from unknown underground artists to worldwide fame. They have created their own independent music label – Odd Future Record and gained their own slot on Adult Swim for their show Loiter Squad. They also opened their own pop-up shop selling all types of Odd Future merchandise, such as socks, skate decks, shirts and air fresheners. While it was intended to be only a temporary store open for a couple months, it still stands to this day. They also sell their merchandise in their online store, which enjoys unabated popularity since their fans want to follow their distinctive clothing style, inspired by the old school rappers.[77]

Manchester is also a home of successful musical collectives. The Murkage Cartel is an arts organisation made

---

75 Cree (n 1) 3.

76 Jarvis, 'OFT x CCS Present: The History of Odd Future | A Beginners Guide to OFWGKTA' (*Odd Future Talk*, 24 August 2012) <http://oddfuturetalk.com/2012/08/oft-x-ccs-present-the-history-of-odd-future-a-beginners-guide-to-ofwgkta/> accessed 29 August 2012.

77 ibid.

up of the core band Murkage, and a group of closely cooperating with it DJs, artwork designers, producers, promoters, bloggers, stylists, radio presenters, film makers and photographers. They all support each other in their endeavours, promoting both the Murkage Cartel brand, as well as their own names. It would be interesting to see where they are going to be in a few years' time, especially since they have already started being involved in merchandising and sell their originally designed tees, which proved to be the first step in almost every brand extension strategy.[78]

While each of the artists presented above is different and they adopt different marketing strategies, there seems to be a pattern that they all have in common. If we wanted to create a model for brand extension in the hip-hop industry, it would a follow a few basic steps. First one would be the same as it is for every emerging brand – deliver a good quality product, gain recognition and establish brand as a reference point of consistently high quality products. Once an artist's name earns this positive association, an artist may start slowly extending his or her brand. It looks like the most common step is to create an independent label, which will promote not only the core artist but also new talents he or she discovered. These new artists are assumed to represent the same quality as the founding artist and as such are his or her brand extension as well. If they succeed, it only reinforces the label's appeal, which in turn also transfers to the core artist.

Another essential element in the early steps of the brand extension strategy is the creation of clothing lines. It usually starts with simple T-shirts with artists' logos but following the success of the artists, they can become rightful fashion labels. The number of artists expanding into fashion has been on the rise in the recent years, giving consumers

78 Daniel Nolan, 'Interview: Murkage Cartel' *City Life* (6 May 2011) <http://www.citylife.co.uk/news_and_reviews/news/10019404_interview__murkage_cartel> accessed 28 August 2012.

plenty of options to express their individuality and style.[79] This is why it is important for the artist to create a unique style, which he or she represents and which consumers may follow by buying the clothes he or she endorses.

Once these three pillars – strong brand name, record label and fashion label – are established, the possibilities are endless. Depending on their own personal style and attitude, they may want to create a luxury brand, like Diddy, or become entrepreneurs, like Jay-Z. Artists need to be very careful though as all their actions influence and affect their brand names and images, both, with their fans and their investors. This is why the hip-hop artists need to make sure that their brand extensions are all reflections of themselves.[80]

## VII. Conclusion

The intense competition in the dynamic marketplace, combined with high cost of investment needed to enter new markets, have pushed companies to adopt new and innovative brand strategies[81] Since brands belong to their owners who helped create and nourish them, they should have a right to exploit them and prevent others from taking advantage of their work. This research paper aimed at explaining why brand extension is so popular in the music industry. Since artists are producers, products and brands at the same time, they should have a right to take advantage of their work.

Hip-hop culture has produced very successful artists who have managed to extend their brands into the areas of entrepreneurship and leadership. This global phenomenon is a symbol of innovation and creativity offering useful lessons for various mainstream organizations in the public and private arena. There is still a struggle for the artists as to how to extend further into the mainstream without losing the

79 Proctor (n 63).

80 ibid.

81 Besharat (n 9) 1247.

street appeal, which is supposed to be inseparable from the hip-hop culture.[82] Looking at the successful artists becoming entrepreneurs, there is no doubt that the acceptance for the business deals creeping into the hip-hop culture is now a common feature. If this trend remains, the multitasking artists may become a standard in the industry, as hip-hop is definitely not the lonely island. The only question is how much consumers are willing to accept and, consequently, how far the brand can extend.

82 Patterson (n 49) 11.

## BIBLIOGRAPHY

### Journal Articles

Ailawadi KL, Keller KL, 'Understanding Retail Branding: Conceptual Insights and Research Priorities' (2004) 80 Journal of Retailing 331

Besharat A, 'How Co-Branding Versus Brand Extensions Drive Consumers' Evaluations of New Products: A Brand Equity Approach' (2010) 39 Industrial Marketing Management 1240

Boote M, 'Diddy Do It?' (2007) 197 Trademark World 18

Choi TM et al, 'Fast Fashion Brand Extensions: An Empirical Study of Consumer Preferences' (2010) 17 Brand Management 472

Cree R, 'Papa's Got A Brand New Brand: An Investigation of Brand Strategy in the UK Music Industry' 4 (2) The International Journal of Urban Labour and Leisure <http://www.ijull.org/vol4/2/cree.pdf> accessed 23 August 2012

Freling TH, Crosno JL, Henard DH, 'Brand Personality Appeal: Conceptualization and Empirical Validation' (2011) 39 Journal of the Academy of Marketing Science 392

Harris G, 'The Sugababes: the trademark rights associated with band names' (2010) 21(5) Entertainment Law Review 165

Ilicic J, Webster CM, 'Effects of Multiple Endorsements and Consumer-Celebrity Attachment on Attitude and Purchase Intention' (2011) 19 Australasian Marketing Journal 230

Kubacki K, Croft R, 'Markets, Music and All That Jazz' (2011) 45 (5) European Journal of Marketing 805

Lye A, Venkateswarlu P, Barrett J, 'Brand Extensions: Prestige Brand Effects' (2001) 9 (2) Australasian Marketing Journal 53

Martinez E, Montaner T, Pina JM, 'Brand Extension Feedback: The Role of Advertising' (2009) 62 Journal of Business Research 305

Schmitt B, 'The Consumer Psychology of Brands' (2012) 22 Journal of Consumer Psychology 7
Slind-Flor V, 'Money and Mayhem' [2007] Intellectual Asset Management 15
Sullivan MW, 'Brand Extensions: When To Use Them' (1992) 38 (6) Management Science 793
Sunde L, Brodie RJ, 'Consumer Evaluations of Brand Extensions: Further Empirical Results' (1993) 10 International Journal of Research in Marketing 47

## News Articles

Bastin M, 'How Far Can You Stretch Your Brand?' *China Daily* (18 November 2011) <http://usa.chinadaily.com.cn/opinion/2011-11/18/content_14118755.htm> accessed 23 August 2012
Baubeau A, Kesmodel D, 'Bruce Willis Sees Spirits in Equity Deal With Belvedere' *The Wall Street Journal* (23 December 2009) <http://online.wsj.com/article/SB10001424052748703478704574611690552812758.html> accessed 23 August 2012
Dar N, 'Jay-Z, Beyonce to Become Forbes Highest-Paid Couple' (*Business Recorder*, 9 August 2012) <http://www.brecorder.com/arts-a-leisure/261-life-a-style/72829-jay-z-beyonce-to-become-forbes-highest-paid-couples.html> accessed 23 August 2012
Gratton J, 'How the Music Industry Is Using Brands and Advertising to Plug the Gap Left By Illegal Downloads' (*Red C,* 20 October 2009) <http://www.redcmarketing.net/blog/marketing/how-the-music-industry-is-using-brands-and-advertising-to-plug-the-gap-left-by-illegal-downloads/> accessed 23 August 2012
Hattenstone S, 'Jay-Z: The Boy from the Hood Who Turned Out Good' *The Guardian* (20 November 2010) <http://www.guardian.co.uk/music/2010/nov/20/jay-z-interview-simon-hattenstone> accessed 23 August 2012
Jarvis, 'OFT x CCS Present: The History of Odd Future | A Beginners Guide to OFWGKTA' (*Odd Future Talk*, 24 August 2012)

<http://oddfuturetalk.com/2012/08/oft-x-ccs-present-the-history-of-odd-future-a-beginners-guide-to-ofwgkta/> accessed 29 August 2012

Jurgensen J, 'The State of Jay-Z's Empire' *Wall Street Journal* (22 October 2010) <http://online.wsj.com/article/SB10001424052702304741404575564092478617462.html> accessed 23 August 2012

Knittel Ch R, Stango V, 'Celebrity Endorsements, Firm Value and Reputation Risk: Evidence from the Tiger Woods Scandal' *Massachusetts Institute of Technology* (25 August 2010) 4 <http://www.econ.ucdavis.edu/faculty/knittel/papers/Tiger_latest.pdf> accessed 24 April 2013

Nolan D, 'Interview: Murkage Cartel' *City Life* (6 May 2011) <http://www.citylife.co.uk/news_and_reviews/news/10019404_interview__murkage_cartel> accessed 28 August 2012

O'Malley Z, 'Who Will Be Hip-Hop's First Billionaire?' *Forbes* (22 September 2011) <http://www.forbes.com/sites/zackomalleygreenburg/2011/09/22/who-will-be-hip-hops-first-billionaire-jay-z-diddy-dr-dre-birdman-50-cent/> accessed 23 August 2012

Ortiz J, 'Surprise? Paris Hilton Earns Over $10M a Year From 17 Different Product Lines' *Business Insider* (1 June 2011)
<http://articles.businessinsider.com/2011-06-01/entertainment/30076238_1_piers-morgan-paris-hilton-product-lines> accessed 23 August 2012

Patterson VL, 'Engaging Hip-Hop Leadership: Diversity, Counter-Hegemony and Glorified Misogyny – (Free-Style Version)'
<http://www.ipa.udel.edu/3tad/papers/workshop2/patterson-newman.pdf> accessed 23 August 2012

Proctor R, 'Jay-z Brand Extension' (14 December 2011) <http://www.rashaproctor.com/wp-content/uploads/Jay-Z-Brand-Equity.pdf> accessed 23 August 2012

Strauss K, 'Celebrity Entrepreneurs on the Rise?' *Forbes* (16 May 2012) <http://www.forbes.com/sites/karstenstrauss/2012/05/16/celebrity-entrepreneurs-on-the-rise/> accessed 23 August 2012

## Websites

http://www.ciroc.com/
http://www.diddybeats.com/
http://www.islanddefjam.com/
http://rocawear.com/
http://www.seanjohn.com/
http://www.universalmusic.com
http://www.wutang-corp.com/
http://www.youtube.com/

# Do No Harm: 'Best interests', patients' wishes and the Mental Capacity Act 2005

Sarah L Morgan

## Abstract

*The Mental Capacity Act 2005 provides a mechanism for decisions to be made on behalf of individuals who are deemed incapable of making decisions for themselves. Central to the Act is the application of the 'best interests' principle, whereby any decision made must primarily consider what is best for the individual in question. Whilst this principle could be seen as potentially paternalistic in nature, leading to ignorance of individual's wishes, the Act and its code of practice positively encourage the involvement of non-capacitor's in decision making, regardless of the extent of their incapacity. This discussion explores the nature of 'best interests' and the complex legal ramifications of making decisions on behalf of others. It explores the nature of capacity and whether an individual's wishes should always override what is thought to be in their 'best interests'. The discussion concludes that the focus on 'best interests' within the Mental Capacity Act 2005 does not undermine the wishes of individuals who do not have capacity to make decisions on their own behalf. Best interests must, however, be viewed holistically and prospectively, considering all elements (social, emotional and medical) which may have a bearing on the outcome of an individual case.*

## I. Introduction

An individual's capacity for thought and decision making for all aspects of his life and especially in relation to his medical care is something that many people take for granted. If, however, this capacity is diminished or lost (or may never have truly been present), how and by whom should such decisions regarding one's life be made? The Mental Capacity Act 2005 (MCA)[1] provides a mechanism for decision-making on behalf of individuals over the age of 16

1 Mental Capacity Act 2005.

who for any reason (whether temporarily or permanently) lack capacity. The Act focuses on the concept of 'best interests'. Section 1(5) states: 'An act done, or decision made, under this Act for or on behalf of a person who lacks capacity must be done, or made, in his best interests'[2]. The application of this principle in a medical setting (as for all settings) requires thought and input from a number of sources, including the individual for whom the decision is to be made. In this essay, I argue that the use of the best interests concept does not undermine the ability of non-capacitor patients' wishes to be respected. Determining a patient's best interests and where their wishes fit into a decision making process is complex, yet the Act encourages open discussion with all parties to ensure best outcomes. It is the determination of capacity and potential changes in capacity over time, along with an individual's change in their own wishes as time progresses, that has more impact on whether their wishes can always be respected.

## II. Best Interests Prior to the MCA 2005

The best interests concept is central to the MCA and the decision making for and on behalf of individuals who lack capacity. The application of this concept to the provision of medical treatment for adults who lack capacity (and to more general substitute decision making on their behalf) is derived from the case of *Re F (Mental Patient: Sterilisation)*[3], in which permission was sought for the sterilisation of a 35-year-old female who was a resident of a home for the mentally disabled. The individual in question had a mental age of approximately 5 to 6 years but had embarked on a sexual relationship with a male resident at the home. This sparked concerns regarding her potential to become pregnant, a situation that would have been disastrous for her due to her lack of understanding. The request for

2 ibid.

3 *Re F (Mental patient: Sterilisation)* [1990] 2 AC 1.

sterilisation followed a review of all potential contraceptive measures available and was made by the individual's mother and the authority responsible for the care home. At the time, there was no clear route in common law for determining how treatment of the mentally incapacitated should be decided upon – especially with regard to what was seen as such an extreme form of non-therapeutic treatment.

The House of Lords declared that that the operation could go ahead[4]. The judgment stated that medical treatment could be provided to an adult with mental incapacity (described as 'an inability to consent for one reason or another') if that treatment was shown to be in the patient's best interests[5]. Each of the Lords' judgments quoted the concept of best interests, yet no explanation was provided for what these best interests may be. It is interesting to note, however, considering the statement being discussed within this essay, that Lord Goff commented on how these interests should be judged. He suggested that it was not only the doctors who should make the decision and stated that he anticipated that 'an inter-disciplinary team will in practice participate in the decision' in determining whether treatment was appropriate[6]. This initial declaration, however, did focus on the use of the Bolam test[7] in determining whether a decision was 'best' based on the absence of negligence in the medical outcome chosen.

A number of further cases have expanded the common law application of the best interests concept in this context prior to it being codified within the MCA. A change in the approach to identifying a person's best interests can be seen, however, with a move away from applying the Bolam

4 ibid.

5 ibid, *cf* judgment by Brandon LJ.

6 *Re F* [1990] 2 AC 1.

7 The Bolam test was established in the judgment of McNair J in the case of *Bolam v Friern Hospital Management Committee* [1957] 1 WLR 582.

test in their assessment[8] to a wider (more holistic) welfare-based review of a person's life requirements (i.e. not simply focusing on the medical perspective) and to a more objective, reasoned basis for decision making[9].

## III. The Nature of Capacity

Capacity can be described as the ability of a person to make a reasoned decision, and under the Act, a person is defined as lacking capacity 'if at the material time he is unable to make a decision for himself in relation to the matter because of an impairment of, or a disturbance in the functioning of, the mind or brain'[10]. However, while a person may be deemed as lacking capacity at that particular point in time, this may not always be the case. There are several categories of capacity within which an individual could be placed – from a transient loss of capacity (as may be seen in cases due to episodic illness or drug induced) to cases where capacity has diminished over time (as seen in dementia sufferers whose ability to actively participate in decision making decreases as the illness progresses), to cases where capacity is lost completely and will never be regained (as seen in cases such as Persistent Vegetative State). At this end of the scale, however, there is another important category – those who may never have had capacity due to the presence of a mental disability. Capacity distinctions may be further complicated by the person's actual level of capacity – legally a person is described as being either competent (having capacity) or non-competent (incapacitated),[11] but in reality, this is a sliding scale. A person may not be deemed legally competent, but they may still have the capacity for thought and discussion and therefore be able to let their

8 Judgment by P Butler-Sloss in the case of *Re S (Adult patient: Sterilisation)* [2000] 3 WLR 1288, where she dismissed the use of the Bolam test in such cases.

9 The use of a balance sheet approach is described by Thorpe LJ in *Re A (Medical treatment: Male sterilisation)* [2000] 1 FCR 193, CA.

10 MCA s 2(1).

11 MCA 2005.

wishes be known. This should enable them to play a role in making reasoned decisions about issues that would affect their life.

Assessing capacity is the ultimate starting point when making best interest decisions – the Act is careful to point out that a competent person may make a decision that is seen by others as unwise and not in their best interests, but as they are competent, their decision must be accepted[12]. An unwise decision does not equal a lack of capacity[13]. When a person is deemed to be incapacitated, a functional test[14] to assess their competence would provide an indication as to their level of capacity and therefore their ability to be involved in decision-making. As will be discussed later, every effort must (and indeed should) be made to involve individuals in the decision process. Functional tests can therefore provide an indication of the level of involvement possible.

Issues in the categorisation of non-capacitors become apparent when providing direct guidance on how best interests are to be determined, and it also demonstrates that different approaches to each category may be required. For those with a transient loss of capacity (at least anticipated to be transient), there may be a reluctance to make certain decisions based on the potential for regaining capacity and the impact on future wishes. As part of the assessment process, the Act requires that a judgment be made on whether the person lacking capacity is likely to regain it at some point in the future[15]. Where individuals are losing or have lost capacity, then there may be a greater need to look back at previous thoughts, wishes, and decisions made by that individual when they had full capacity. For those who have

12 MCA s 1(4).

13 See *Re B* [2002] EWHC, in which a tetraplegic woman argued to have her ventilator switched off. She was found to be mentally competent although suffering from a severe disability. Her decision stood.

14 As described in the MCA s 3(1).

15 MCA s 4(3).

never had capacity, they may never have been given the potential to make decisions or to discuss their thoughts and wishes.

How, therefore, can their wishes be determined? Regardless of the category or level of capacity of the person in question, they will all have wishes and feelings that should be taken into consideration when decisions are made. Should the Act therefore distinguish between these categories? What may differ between these groups is the ability to determine what the individual's wishes and feelings may be and the impact of the decision to be made on the future well-being of that person. As the nature of capacity may change in an individual over time, so, potentially, will his or her best interests.

## IV. Defining Best Interests

The best interest standard or principle is applied in a number of legal situations when determining the course of action that can provide the best potential outcome for the individual in question. There is no clear definition of best interests in most cases; 'best' is difficult to describe as what is 'best' will differ not only between individuals, but also between different time points in an individual's lifetime[16]. Someone taking 'control' of a person's life through the making of decisions on their behalf can be viewed as paternalistic in nature. As Coggon states[17], it should, however, not be viewed simply as a concept; rather, it should instead be seen as 'a construct for good decision-making'. He additionally states that it is a 'goal to aim towards in all cases', acknowledging that a singular best interest may not be possible to determine in each case.

---

16 John Coggon, 'Best Interests, Public Interest and the Power of the Medical Profession' (2008) 16 Health Care Analysis 219; *cf* Loretta M Kopelman, 'The Best Interests Standard for Incompetent or Incapacitated Persons of All Ages' (2007) Journal of Law, Medicine, and Ethics 187.

17 Coggon (n 16) 219.

For a competent individual who is able to make their own decisions, a best interest choice is subjective, as it is based on their own thoughts and feelings at that time. When one is trying to determine the best interests of another, it should become an objective decision process based on a number of identified factors (the approach fostered by the MCA). If the individual is involved in the discussion, however, one confounding element in this process is that subjective elements (their thoughts, wishes, and feelings) must be included within the objective analysis. Does this potentially change the objective nature of the discussion? Can a truly objective discussion take place around issues that would normally be seen as subjective?

The Act itself does not try and define what someone's best interests might be. This could be seen as a weakness of the Act, considering the potential gravity of the decisions that may need to be made on another person's behalf. Would a more precise definition provide better guidance for how best outcomes are achieved from the application of the Act? The nature of the actual outcome is difficult to determine, as the Act can be applied across a range of decisions/acts that need to be carried out on behalf of an individual lacking capacity. This approach should instead be viewed as a strength as it allows a more flexible approach to decision making to be applied on an individual case basis.

The code of practice (CoP), written to provide guidance to users of the Act[18], clearly demonstrates this flexibility. The CoP provides a set of guiding principles for how an individual's best interests can be ascertained, including the identification of their views (past, present, and future); consulting others who may be able to provide a view on that individual; the restriction of rights (avoiding conflict with the European Convention on Human Rights); and, most importantly, encouraging participation of the individual

18 Mental Capacity Act Code of Practice 2007.

involved. What the Act encourages is a full and frank discussion of all the contributory factors (both good and bad) surrounding the decision to be made.

Trying to determine an individual's best interests is therefore a potentially complex process. It may not be readily apparent what the best interests of an individual may be, and a careful weighing up of all factors that could influence the potential outcome needs to take place. The preferred method for this process is the use of a 'balance sheet' approach, as first described by Thorpe LJ in his judgment on the case of *Re A (Mental patient: Sterilisation)*[19]. He proposed the use of a balance sheet (as applied in financial matters) to determine what a person's best interests might be. He described the drawing up of two columns, one containing a 'factor or factors of actual benefit' and then 'counterbalancing the dis-benefits to the applicant'[20]. If one column is in credit compared to the other, then that is where the balance of best interests is said to lie. This approach allows for an objective overview of the issues at hand.

The issues that need to be reviewed for each individual case will vary, but as illustrated in earlier common law examples prior to the MCA, they should encompass all aspects of the decision that will affect the individual. These are not confined to clinical issues alone, but should also include consideration of the social, emotional, and welfare issues that may affect the individual. As Thorpe LJ stated in *Re A*, the 'evaluation of best interests is akin to a welfare appraisal'[21]. Welfare appraisals, such as that used within the Children Act 1989, require a thorough review of all factors that impact the individual who is at the centre of the discussion, and the same approach should be applied in

19 *Re A* [2000] 1 FLR at 555.

20 ibid.

21 in *Re A* [2000] 1 FLR.

cases under the MCA. As McGuiness stated[22] in her discussion of best interests as a pragmatic approach, a 'best interests' standard should be seen as 'not telling us which interests to protect,' but that it can 'act as a general principle stating that we should reach the best decision overall'.

## V. Participation of the Incapacitated Individual

Mental incapacity may stop people from making a reasoned independent decision regarding a course of action that involves them. It does not preclude them, however, from being involved in the decision making process. The Act encourages active participation by the incapacitated person in the decision making process to ensure that their wishes are taken into account[23]. In terms of respecting the wishes of non-capacitor individuals, this can be seen as a strong point for the Act. By discussing the decisions that need to be made, including a review of all possible options and outcomes, then the wishes, thoughts, and feelings of the individual can be ascertained and should then be taken into account when the final decision is made.

There are potential problems, however, that need to be overcome, especially when considering varying levels of capacity. The major issue is that of communication, such as whether the person in question has the ability to communicate their thoughts and wishes. This could be due to a complete inability to communicate or a lack of clarity in their communication, whether verbal or nonverbal. There are clear cases where communication is impossible (e.g. patients in a persistent vegetative state); however, for many others, the forms of communication may be many and varied. Patients who have lost capacity may have diminished abilities in verbal communication or may require the use of nonverbal methods for communication, such as the use of

22 Sheelagh McGuiness, 'Best Interests and Pragmatism' (2008) 16 Health Care Analysis 208.

23 MCA s 4(4).

signboards or other methods that require translation by a carer. In cases where a person has never had capacity, their language abilities may be restricted and may require some form of interpretation, although the use of an interpreter or translator in such a process does potentially introduce an element of bias that is different from the true wishes of the patient. For example, carers may try to provide their version of the individual's thoughts, believing that to be of more benefit to the patient.

As discussed earlier, patients who have never had capacity may never have had the opportunity to make decisions on their own behalf, so it may be difficult for those individuals to articulate their wishes. Where an individual's lack of capacity is due to a mental disability, then they will likely not fully understand the issues facing them, and all that may be ascertained is their likes and dislikes around certain issues. These should not be ignored, however, as they may still have important bearing on the final decision-making.

Patients who have lost capacity at the point of decision-making, (and are unable to actively contribute to the process) do have another route by which their wishes can be taken into account – through the use of advance directives. The Act states that when these advance directives involve end-of-life treatment decisions, they must be adhered to[24] if they were written at a point in time when the person did have capacity. As part of the participative approach, other written orders should be reviewed; however, the nature of the orders and the level of capacity of the individual at that point must be taken into account.

## VI. The Decision Making Process – Who Decides?

In trying to determine all the factors that may be involved when defining an individual's best interests, a variety of people need to be involved in the decision making process

24 MCA section 24.

– this refers back to Lord Goff's comments in *Re F*[25] regarding the use of an 'inter-disciplinary team'. No person exists on his own – we all interact with a number of people, including family members, friends, carers, and medical personnel, during our lives. Each of these individuals may have an insight into what the incapacitated person's thoughts and wishes may be and should therefore be involved in the decision making process. The British Psychological Society suggests that a 'best interest group meeting' should be held and chaired by one individual in order to determine the best outcome for the person involved[26]. Some incapacitated individuals may have previously elected someone to be their voice in these matters, up to the level of Lasting Power of Attorney, but this will not be the case in all situations.

As discussed previously, the Act does appear to encourage discussion as an important part of any decision making process for determining best interests. By involving a group of people in the process, a fuller discussion of all the issues that play a role in this process can take place. The aim should be to gain group consensus in decisions, but with a variety of people, all with different emotional and other attachments to the individual in question, there are likely to be disagreements.

What must be remembered through all discussions is that the best interests of the individual in question should take primary place. When discussing the welfare – especially the social welfare – of an individual, it is difficult to discuss this in isolation from the persons who interact with the incapacitated individual on a regular basis, especially family members, carers, etc. Choudhry[27] compares the approach

---

25 *Re F* [1990] 2 AC 1.

26 British Psychological Society, 'Best Interests: Guidance on determining the best interests of adults who lack the capacity to make a decision (or decisions) for themselves' (2007) Accessible at: <http://www.mentalhealthcare.org.uk/media/downloads/Best_Interests_Guidance.pdf> Last accessed 29 April 2013.

27 Shazia Choudhry, 'Best Interests in the MCA 2005 – What Can Healthcare Law Learn From Family Law?' (2008) 16 Health Care Analysis 240.

under the MCA with that used under the Children Act 1989. While the Children Act openly mentions the impact of decisions on others involved, the MCA, as presented, is completely focused on the individual's interests. By focusing on the individual, their interests are given the primary place at all times – but should the interests of others be taken into account? This may move the discussion away from a patient's wishes, but due to their incapacity, they may not understand the implications of the decision they are involved in, which could ultimately lead to honouring their best interests at that point that ultimately has an adverse effect on their future wellbeing.

It is not appropriate for the Act to define exactly who should be involved in the decision making process, and the membership of the decision group should be formulated specifically for each individual case. There have been a number of discussions as to who should ultimately make a decision but to reach such an objective decision, a dispassionate view on the matter is required. Family members with strong emotional attachments may not be able to take this dispassionate view, while doctors may be too focused on clinical requirements. A multi-disciplinary team led by a neutral individual (potentially with some understanding of incapacity) may well provide the best outcome for the individual in question.

## VII. Alternative Approaches

Before making a final comment on whether the approach in the MCA undermines the best interests of an individual, there is a need to look at potential alternative approaches that could be applied. The closest alternative is that of substituted judgment[28] (as used in the United States), where an individual must try to place themselves in the position of the incapacitated patient when determining the decision that must be made. This is a much more restrictive

28 Kopelman (n 16) 187.

approach, however, in terms of determining the individual's thoughts and feelings and has higher potential for undermining their best interests.

It is not easy to define other alternative approaches aside from requiring that all individuals who have capacity to make advance directives or to at least letting their feelings and wishes be known at an early age. These would later need revising, as it is known that thoughts and feelings change over time, and when placed in the situation of losing capacity, how can one determine his or her own thoughts? We are, however, still left with the issue of those who have never had capacity.

## VIII. The Careful Balancing Act – Undermining or Empowering?

The balancing act that is required for determining best interests is, as has been discussed, not a simple one, nor should it be expected to be. When an individual has the ability to take part in this process, every effort should be made to determine what their wishes might be. These must not be ignored against the backdrop of clinical or other issues that are included in the decision making process—the individual's wishes should be given a voice, though not an overpowering one. They should not be allowed to override the other viewpoints, just as another person's viewpoint should not override that of the incapacitated person. At the same time, the process should not simply pay 'lip service' to the involvement of the incapacitated individual. Rather, the decision making team should fully consider their thoughts.

The Court of Protection[29] has an important role to play in this process. This court is only invoked when a decision cannot be made due to disagreements among the decision making team. When they are asked to make a decision, they must demonstrate that they have included the

29 MCA 2005.

wishes of the individual in their reasoning and demonstrate the balanced approach that is required in action.

## IX. Conclusion

Common law has led to the development of the best interests concept in relation to individuals who lack the capacity to make their own decisions. This has expanded over time to require a review of not just medical/clinical aspects of a person's health, but also their welfare, social, and emotional requirements. For a thorough review of these various factors, involvement from a number of parties with links to the individual in question is required to determine what that person's thoughts, views, and wishes might be. This is a complex process that is further complicated by the past, present, and future capacity of the individual in question. What must be remembered is that their welfare and their wishes must be considered first in all discussions, ensuring that they are respected when possible. This does not mean that their wishes must be adhered to, but that sound reasoning and balancing must be applied by the person designated to make the decision on their behalf to ensure that their viewpoint is respected at all times. The focus of the Act should remain on the discussion of best interests, not on defining what the outcome should be; therefore, decisions made by the Court of Protection must take care not to set a dangerous precedent in seeming to ignore a patient's wishes. As our understanding of capacity and the nature of decision making in those lacking it evolves, so too should their involvement in the process, although this does not negate the best interests approach for determining the decisions that must be made.

## BIBLIOGRAPHY


### Legislation

Mental Capacity Act 2005
Childrens Act 1989

### Cases

*Re F (Mental patient: Sterilisation)* [1990] 2 AC 1
*Re S (Adult patient: Sterilisation)* [2000] 3 WLR 1288
*Re A (Medical treatment: Male sterilisation)* [2000] 1 FCR 193, CA
*Re B* [2002] EWHC

### Journals

Buchanan A, 'Mental Capacity, Legal Competence and Consent to Treatment' (2004) 97 Journal of the Royal Society of Medicine 415

Choudhury S, 'Best Interests in the MCA 2005 – What Can Healthcare Law Learn From Family Law?' (2008) 16 Health Care Analysis 240

Coggon J, 'Best Interests, Public Interest and the Power of the Medical Profession' (2008) 16 Health Care Analysis 219

Donnelly M, 'Best Interests, Patient Participation and the Mental Capacity Act 2005' (2009) 17 Medical Law Review 1

Dunn, MC, ICH Clare and A Holland, 'To Empower or to Protect? Constructing the 'Vulnerable Adult' in English Law and Public Policy' (2008) 28(2) Legal Studies 234.

Heywood R, 'Parents and Medical Professionals: Conflict, Cooperation and Best Interests' (2012) 20 Medical Law Review 29

Kopelman LM, 'The Best Interests Standard for Incompetent or Incapacitated Persons of All Ages' (2007) *Journal of Law, Medicine and Ethics* 187

McGuiness S, 'Best Interests and Pragmatism' (2008) 16 Health Care Analysis 208

## Reports

British Psychological Society, 'Best Interests: Guidance on determining the best interests of adults who lack the capacity to make a decision (or decisions) for themselves' (2007) Accessible at: <http://www.mentalhealthcare.org.uk/media/downloads/Best_Interests_Guidance.pdf>

# Hard Cases

Dorota Galeza

## Abstract

*On the one hand, legal doctrine seems indeterminate, but it may be maintained that even in "hard cases", judges only "constantly talk about the answer they already knew in advance." Legal philosophers are divided in this respect. Dworkin provided a very convincing answer for the "one answer" model, whereas both inclusive and exclusive positivists and Critical Legal Studies and legal realists presented plausible responses to the "no one answer" model. This article provides a new insight into legal reasoning by linking Dworkin's theory with French existentialism. It tackles with most common criticisms of Dworkin's argument and states which facets of this criticism are most cogent.*

## I. Introduction

Is legal doctrine really indeterminate? In other words, do judges have discretional power to use legal doctrine as they wish? Or, even in "hard cases", do judges only 'constantly talk about the answer they already knew in advance'?[1] Indeed, the answer to this question can have a tremendous effect in relation to lawmaking. Is new law created in the courtroom each time a judge decides a case without a precedent or do judges only administer what is to be dispensed? Legal philosophers are divided in this respect. Dworkin provided a very convincing answer for the latter, whereas both inclusive and exclusive positivists and Critical Legal Studies (CLS) and legal realists presented plausible responses to the former. In this article, I will assess those theses and answer the difficult question whether in "hard cases" judges make law by enforcing their political and moral judgments or only state the underlying principles that are known already. I will spread my analysis to smaller themes, such as the political nature of adjudication and the language.

---

1 Albert Camus, *The Fall* (Penguin Books 1957) 107.

## II. The ambiguous concept of a "hard case"

First, I will examine the term "hard case". Different theories adopt different interpretations of this term. I will start with the positivistic approach. Twining and Miers[2] define a "hard case" as a case in which a judge (i) thinks the letter of the statute is clear (whether this is due to the fact 'that the text or the underlying intent), and (ii) has significant reservations about the application of the statute so interpreted.'[3] They distinguish a "hard case" from a "difficult case", where the latter case is such in which the judge thinks the letter of the statute (however regarded) is not clear.[4] A slightly different approach is taken by Dworkin, who, in reference to positivistivism, defines a "hard case", as follows: when a certain case cannot be resolved by the use of an unequivocal legal rule, set out by the appropriate body prior to the event, 'then the judge has, accordingly to that theory, a 'discretion' to decide the case either way.'[5] Dworkin, however, does not identify the characteristics of a "hard case" and he does not provide a judge with instructions on how to decide whether the contentious case is a "hard" one.[6] He merely provides very broad guidelines, such as '"hard cases" arise when "both in politics and law, ... reasonable lawyers ... disagree about rights";[7] "no established rule can be found";'[8] etc. In the light of the aforementioned, we can distinguish Dworkin's two types of "hard case": a) a case without a rule and, b) a case with a rule which offers 'incomplete,

2 William Twining and David Miers, *How to do Things with Rules* (3rd edn, Weidenfeld and Nicolson 1991).

3 John N Adams and Roger Brownsword, *Understanding law* (4th edn, Thomson Sweet and Maxwell 2006) 102.

4 ibid.

5 Roland Dworkin, *Taking Rights Seriously* (Duckworth 1978) 81.

6 Alan C Hutchinson and John N Wakefield, 'A Hard Look at "Hard Cases": The nightmare of a noble dreamer' (1982) 2 Oxford J Legal Studies 86, 88.

7 Dworkin (n 5) *xiv*.

8 ibid, 44; Hutchinson and Wakefield, (n 6) 86, 91.

ambiguous or confliction guidance'.[9] However, this typology may differ according to the American Realists, who casted a doubt upon the fact whether precedents could ever restrict the application of a legal rule. As they pointed out there were always factual differences that could be distinguished further.[10] The illustration of this mechanism is given by Schlag, who compares two interpretations of the term "vehicle". According to Hart, an automobile was clearly a vehicle.[11] However, this assumption neglected the fact that the word "vehicle" has a fundamental meaning, 'separate from and independent of the rest of the sentence – is just that – a legal move'.[12] Even if, as put forward by Hart, there is such a fundamental meaning, it is subject 'that this core meaning is or should be determinative of the meaning of the ordinance'.[13] This was supported by Fuller, who advocated that Hart's atomistic approach to interpretation of presumption that the term '"vehicle" has meaning in and of itself'[14] is pointless. It can result in illogical interpretations of the rule. This semantic approach utilised a legal matter. Fuller's purposive analysis of the legal rule was aimed not only on Hart's semantic grounds, but primarily on the premise advocated by Hart that atomistic word parsing would spoil 'a purposive "structural integrity"...of the law'. [15] Probably, those theoretical problems dissuade Hart from giving a classic definition of a "hard case" and to merely to give an example of it. The concept of "hard case" is too vague to be neatly put in words. For the convenience of this

9 Herbert LA Hart, 'Law in the Perspective of Philosophy: 1776-1976' (1976) 51 NYUL Rev 538, 547.

10 Michael DA Freeman, *Lloyd's Introduction to Jurisprudence* (7th edn, Thomson Sweet and Maxwell 2001), 1387.

11 Herbert LA Hart, 'Positivism and the Separation of Law and Morals' (1958) 71 Harv. L. Rev. 593, 607 and 615.

12 Pierre Schlag, 'No Vehicles in the Park' (1999) 23 Seattle Uni L Rev 381, 387.

13 ibid.

14 ibid.

15 ibid; Lon L Fuller, 'Positivism and Fidelity to Law -- A Reply to Professor Hart' (1958) 71 Harv L Rev 630, 663.

essay I will, nevertheless, adopt a classic definition, close to the one given by Twinig and Miers.

## III. Adopted approaches

The question whether in "hard cases" judges make new law by an exercise of moral and political judgments is inevitably interlinked with the version of sources of law adopted. According to Kennedy, in reference to sources, we can distinguish six different approaches: deduction and judicial legislation (Hart), judicial legislation (Unger), deduction, limiting rules and judicial legislation (Raz), deduction, coherence and judicial legislation (MacCormick), deduction, coherence and personal political theory (Dworkin), deduction and coherence (Civilians).[16] From the above, only Dworkin and Civilians do not accept that judges make new law. All the others are concurrent on the point that judges, while adjudicating cases, do make new law. The only difference between the rest of the theories is the *way* they make new law, meaning the scope of discretion they possess and the nature of judgments by which they are influenced (whether they are political or moral).

The concept of a hard case intertwines two completely different underlying notions - the ideal vision of law in an idyllic world, where every case is heard by Hercules, an ideal judge with all wisdom and knowledge,[17] and the dull, painful reality, where law is created and applied by humans, driven by their weaknesses.

## IV. The "no one answer" approach

I would like to start with the realistic vision of adjudication, where law is indeterminate, judges have a wide discretion, and they are ordinary mortals. The good insight into this world is given by CLS. The movement was

16 Duncan Kennedy, *A Critique of Adjudication* (Harvard University Press 1997) 37.

17 Michael BW Sinclair, 'Hercules, Omniscience, Omnipotence, And the Right Answer Thesis' (2003) 46 New York Law School Law Review 447.

internally inconsistent. Therefore, I will restrict my analysis to the American branch of the movement, primarily to Kennedy and Unger.

First, I will discuss the judge's actual state of mind. An insightful study of psychology of law was provided by Unger. He advocated that this novel approach to the nature heads to an antinomy in the comprehension of the relationship between the mind and the world.[18] He also believed that this antinomy has common trails with the pivotal problems of liberal psychology and political theory.[19] There are three non-exhaustive principles: the division between understanding and desire, the postulate that desires are arbitrary and the stipulation that knowledge is acquired by a mixture of 'elementary sensations and ideas'[20], which metonymically indicates that the acquisition of knowledge is basically "the sum of its parts".[21] If we agree that the law is imperfect, ambiguous, indeterminate and sometimes unjust, we ought to consider the state of mind of the adjudicator, i.e. this is what Hart called the 'internal point of view'[22] that depends on the state of the mind. We may think differently in the particular moment and this can affect our judgments. Each person has a different state of mind and this can vary from an individual to another one constantly or can change in response to certain events. Kennedy maintains we do not know 'what judge's actual state of consciousness of the issue of neutrality may be.'[23] In another words, the contention is that there are elements of legal debate that imply 'ideological

18 Roberto Unger, *Knowledge and Politics* (Free Press 1976) 30.

19 ibid.

20 ibid.

21 ibid.

22 Herbert LA Hart, *The Concept of Law* (2nd edn, OUP 1997) 242; Peter Winch, *The Idea of a Social Science and its Relations to Philosophy* (Routledge, 1958) ch 2; Ludwig Wittgenstein, *Philosophical Investigations* (2nd edn, Blackwell 1958) 197-241; Max Rhenstein, *Max Weber on Law in Economy and Society* (Harvard University Press 1954) 11-12; Neil MacCormick, 'On the "Internal Aspect" of Norms' in Neil MacCormick (eds) *Legal Reasoning and legal Theory* (OUP 1994).

23 Kennedy (n 16) 134.

influence even in the absence of any showing of ideological preferences or intentions, conscious or unconscious, in the person doing the argument.'[24]

The second area discussed by the movement is language. Kennedy provided an insightful study to the ideology of the language. The language itself is a source of political interpretation. As he notices, every language has a temporal (diachronic) and synchronic structure. Vocabulary and grammar change constantly over the years, as the concrete language is subjected to foreign influences, responses to "material" developments like technological and scientific innovations and is intentionally adjusted by users 'who see it as a locale for the playing out of conflicting social projects (Negro, black or African American? Stewardess or flight attendant?).'[25] These linguistic findings apply also to legal disputes[26], particularly hard cases. The choice between literal and purposive approach is a political decision.

Those concepts, advocated by CLS are very insightful and they certainly push the theoretical debate forward. Lucy, in reference to *The Critique of Adjudication*[27], said that '[t]he book is rich in ideas and engagingly written.'[28] But, conversely, CLS's concepts are too "descriptive" and they do not offer any robust vision. Perry contrasted this realistic conception of adjudication with its institutional counterpart. He favoured the latter, because he considered the process of adjudication as 'the essence of law'[29], which distinguishes it from 'the phenomenon of positive law'[30]. Some connection between natural law and

---

24 ibid; Hilaire McCoubrey and Nigel D White, *Textbook on Jurisprudence* (3rd edn, Blackstone Press 1999) 232-233.

25Kennedy (n 16) 134.

26 ibid.

27 ibid.

28 William Lucy, 'What is Wrong with Ideology?' (2000) 20 Oxford Journal of Legal Studies 283, 300.

29 Stephen R Perry, 'Judicial Obligation, Precedent and the Common Law' (1987) 7(2) OJLS 215, 216.

30 ibid.

positive law is necessary,[31] and, therefore, 'both fiat and reason ... [are] necessary elements of law'.[32] Furthermore, certain concepts advocated by CLS could be easily encompassed within the mainstream. For instance, MacCormick persuaded us that judges in "hard cases" need to apply a moderately political discretion[33]. Therefore, Lucy believed that the "reductionist"[34] account offered by CLS lacks 'a single, unified "enlightenment project"'[35]. He advocated that in order to accept the novelties proposed by CLS, some reference to 'the problematic nature of representation, truth, and the human sciences is required.'[36]

This is very close to the account of the movement offered by MacCormick, who thought that 'normative order'[37] is not an outcome of a natural course of things, but 'a hard won production of organizing intelligence.'[38] He believed in the usefulness of the fact that 'the materials are themselves produced through rational activity, at least partly informed by previous dogmatic reconstruction'[39]. This is the reason why Lucy called the movement 'heretical'[40]. Nevertheless, at the same time, MacCormick agreed to a certain extent with CLS that 'hard-case adjudication ultimately rests upon subjective, incommensurable, consequentiality value-choices'.[41] My reading of these two very close accounts of the movement is that judges do not

---

31 John Finnis, *Natural Law and Natural Rights* (Clarendon Press 1980).

32 Perry (n 29) 217, Lon L Fuller, 'Reason and Fiat in Case Law' (1946) 59 Harv L Rev 376; Lon L Fuller, 'Forms and Limits of Adjudication' (1990) 92 Harv L Rev 353, George P Fletcher, 'Two Modes of Legal Thought' (1981) 90 Yale LJ 970, 979.

33 MacCormick (n 22) chs 5-8; Lucy (n 28) 283, 299.

34 ibid 298.

35 ibid.

36 ibid.

37 Neil MacCormick, 'Reconstruction after Deconstruction: A Response to CLS' (1990) 10(5) Oxford Journal of Legal Studies 539, 558.

38 ibid.

39 ibid.

40 Lucy (n 28) 283.

41 ibid 299.

always make new law in "hard cases." Certainly, some judges may fall prey to ideology and personal sense of morality, but there are still judges, who can stand above those difficulties and who can conform with the letter of law, or if the letter of law is lacking, they can conform to uniform standards expressed by legislature and approved by the society. I would also like to criticise the movement on the ground of the method of research adopted. Although CLS's observations of the structure of legal system, based on logic and ideology of the movement, are justifiable and plausible; their empirical account of the ideology of the judge's mind cannot be accepted and justified on scientific grounds. It is not based on Process-Product research and as noticed by MacCormick it lacks theoretical underpinning. It misunderstands the nature of adjudication and cannot in scientific context be considered as fact.[42]

> What may result from everyday experience is either (a) common sense understanding of trial and error generalisations, which work more or less, or (b) question which puzzle us enough to stimulate some scientific endeavor, i.e. questions that may eventually *lead to* some scientific research. Before science comes into existence, there has to be, as already mentioned, a "rape of the senses", or a "breaking with everyday experience".[43]

For those reasons, I do not agree with the CLS' account in psychology of adjudication.

## V. The "one answer" model

Despite the merits presented by CLS, we have always to bear in mind the purpose of the law. The underlying aim of adjudication is more than to just provide *an* answer to a controversial issue. Adjudication is part of a larger system.

---

42 John H Chambers, *Empiricist Research on Teaching: a Philosophical and Practical Critique of its Scientific Pretensions* (Kluwer 1992) 145.
43 ibid 169.

Therefore, to deny the idea that in every "hard case" there is only one correct, unique answer is going against the purpose of the system. The assumption that judges *make* new law and they do it differently each time entails negative implications on adjudication. It also demythologises the vision of a judge as a just Hercules. Even, if we assume that some judges pass wrong judgments, we cannot condemn them as a social group. Arguably Sartre's ideas are applicable in the case of adjudication. Therefore, each judge is not 'fully determined'[44]: each judge has a moral choice.[45] In principle, the CLS' thesis of constant ideological questions which have to be answered is wrong. In my opinion, a judge, like any other being, in their 'human reality'[46] has the power 'to choose... [themselves]; nothing comes to...[them] either from the outside or from within it can receive or accept.'[47] This is more apparent in strictly legal writing:

> There is the universal conviction that something noble and fundamental is at stake when judges decide cases. It is doubtless platitudinous, but not less true, to observe that judges, unlike Rabelais' Judge Bridlegoose,[48] do not decide cases simply on the throw of a dice. Instead, judges strive conscientiously to reach conclusions which are manifestly explicable in terms of previous decisions.[49]

In order to pursue this, they provide reasons for their decisions. Their judgments are never made *in vacuo*. Northrop noted that despite the fact that most judges do not unequivocally express their method of reasoning, the

---

44 Jean P Sartre, *Being and Nothingness: An Essay on Phenomenological Ontology* (Methuen & Co 1972) 440.

45 Joseph Singer, 'The Player and the Cards', (1984) 94 Yale LJ 1, 13; William Lucy, *Understanding and Explaining Adjudication* (OUP 1999) 93; Philip Pettit, *The Common Mind* (OUP 1996), part II.

46 Sartre (n 44), 440.

47 ibid.

48 Francois Rebelais, *Gargantua and Pantagruel* (Penguin 1993), Chaps 37-43.

49 Hutchinson and Wakefield (n 6) 86.

functioning of some method is mandatory. In that sense, Filmer maintained that, 'in law, as in all other things, we shall find that the only difference between a person without a philosophy and someone with a philosophy is that the latter knows what his philosophy is'.[50] The idea of equal importance is apparent in adjudication. Practitioners are concurrent on the point that the judges bear the responsibility to 'maintain the law and apply it in deciding cases'.[51] Nevertheless, the judge, in order to resolve what *the law* on the contentious issue is, must first decide on the point what *law* is.[52] This point takes us back to the real beginning - that is - the dull reality of the incompleteness of law, governed by imperfect judges. However, it is important here to distinguish two different theoretical approaches,[53] notably that whose remit extends beyond mere description; their aim is to present a normative that:

> [t]here are...jurists, such as...Cross [54] ,...Levi [55] and...Murphy[56], whose principal concern is to describe the patterns of reasoning characteristically used by judges; their vantage-point is expository and analytical, rather than critical and evaluative...[Alternatively], there are jurists such as...Horwitz [57] and...Wasserstrom theory of how judges ought to decide cases and their stance is exhortatory.[58]

---

50 Filmer SC Northrop, *The Complexity of Legal and Ethical Experience* (Little, Brown & Co 1956) 6.

51 Hutchinson and Wakefield, (n 6) 86.

52 ibid.

53 Richard A Wasserstrom, *The Judicial Decision* (Stanford University Press 1961).

54 Rupert Cross, *Precedent in English Law* (3rd edn, Clarendon Press 1977).

55 Edward H Levi, *An Introduction to Legal Reasoning* (University of Chicago Press 1977).

56 Walter F Murphy, *Elements of Judicial Strategy* (University of Chicago Press 1977).

57 Donald L Horwitz, *The Courts and Social Policy* (The Brookings Institute 1977).

58 Hutchinson and Wakefield (n 6) 86.

I will focus now on the latter approach. My aim is to distinguish it from CLS and primarily to answer whether within this approach judges have discretion to apply their own political or moral values. Singer believed that the absence of a rational foundation to legal reasoning as advocated by CLS does not prohibit us from 'developing passionate moral and political commitments. On the contrary, it liberates us to embrace them.'[59] If none of the judges could stand above mankind, the purpose of the legal training and judicial career would vanish. The fact that some judges are unable to stand above mankind certifies the fact that they were wrongly selected. The adjudication is a too important social activity to be undertaken by ignoramuses.[60] A further point, flamboyantly expressed by Lucy, is the fact that judgments would no longer maintain such an important place in society if judges explicitly express the nature of the conditions that 'have influenced their reading of the law. And, even if judges are explicit in this way, their assessments can be set aside if determined by ideology (in the critical sense) or if judges are, as Kennedy would say, in "denial".'[61]

If we assume that, nevertheless, judges make new law in hard cases, we ought to consider the further issue, which is finality and infallibility of such law. In this respect, Hart made important observations. A supreme court, while deciding "hard cases", has the power to resolve disputes conclusively. It is irrelevant, whether it made it wrong or right. Nevertheless, such decisions can be denied legal effect by legislation.[62] The fact that judicial decisions in "hard cases" are final and infallible indicates that they form new law. However, since they are subject to a legislative change, they must be considered as inferior to statutory law. This is evidenced in the Snail Darter case[63]. Therefore, in "hard

59 Singer (n 45) 9.
60 HC 52-II (1995) 130.
61 ibid.
62 Hart (n 22) 153.
63 *Tennessee Valley Authority v Hill* (1978) 437 US 153.

cases", judges 'exercise a creative choice in interpreting a particular statute which has proved indeterminate.'[64] Hart supported this formalist approach by *Rex v Taylor*[65], where the court decided that it always has an inherent power to depart from a binding precedent. However, this rigid standpoint, as noticed by Hart, is always open to reconsideration by the simple fact that the choice in deciding whenever a particular statue is incomplete could always be considered as discovery.[66] This vein, apparent in the case law[67], could also support Dworkin's theory that even in "hard cases"; judges do not make new law.

Therefore, it could be argued that, '[i]f the judges make new law, the *power* to do so will be taken away from them'.[68] Such a standpoint was advanced by Lord Scarman in *Duport,*[69] where he said that if the general public and parliament come to the conclusion that the judicial power is only constrained by the judge's sense of what is right and appropriate '(or, as Selden put it, by the length of the Chancellor's foot)',[70] confidence in the judicial system will be substituted with the anxiety of it becoming not clear and biased in its applications. Society will then be prepared to apply parliamentarian powers to curb judicial powers. Their powers to do so will become more limited in a legal development than it should be or is currently.[71]

Sometimes, judges hypocritically support their wide discretion in the adjudication, which justifies Hart's theory.

---

64 ibid.

65 *Rex v Taylor* [1950] 2 KB 268.

66 Hart (n 22)153-154.

67 *Fisher v Bell* [1961] 1 QB 394 [1960] 3 WLR 919 [1960] 3 All ER 731 (1961) 125 JP 101 (1960) 104 SJ 981 1960 WL 18689; *Smith v Hughes* [1960] 1 WLR 830 [1960] 2 All ER 859 (1960) 124 JP 430 (1960) 104 SJ 606 1960 WL 18710.

68 John Snape and Gary Watt, *The Cavendish guide to mooting* (2nd edn Cavendish Publishing Limited 2000) 153.

69 *Duport Steels Ltd v Sirs* [1980] 1 All ER 529, 521.

70 ibid.

71 ibid.

This can be illustrated by the approach taken by Lord Denning. In *Congreve*:[72]

> when Roger Parker ... made a similar prediction [to this of Lord Scarman in *Duport*[73]] in his submissions to the ...[CA], Lord Denning ...stated: 'We trust that this was not said seriously, but only as a piece of advocate's licence.' Mr Parker subsequently apologised if anything he said sounded like a threat.[74]

Nevertheless, judges' hypocrisy can also have a different dimension. It can be aimed to cover judicial legislation. This is evidenced in *Royal College of Nursing*[75], where Lord Diplock departed from the literal meaning of the Abortion Act 1967 and adopted an interpretation inconsistent with law and Parliament's intention. He ruled that nursing staff who after 'the initial surgical intervention of the doctor in the abortion by prostaglandin'[76], actively involved in the remainder of the process came within the scope of "medical practitioner", anticipated by the 1967 Act. This interpretation, however philanthropic in intent, clearly reveals the hypocrisy of the judiciary, who while making new law in "hard cases", disingenuously claim that they merely apply existing rules. I think that such an approach, however duplicitous it may appear, supports Dworkin's theory, as judges strive to do the best of the legal system by the correct *application* of rules and principles.

---

72 *Congreve v Home Office* [1976] QB 629.

73 *Duport Steels Ltd v Sirs* [1980] 1 All ER 529, 521.

74 Stephen H Bailey and Michael J Gunn, *Smith and Bailey on The Modern English Legal System* (3rd edn, Sweet & Maxwell 1996), 256; *The Times* (London, 6 and 9 December 1925); Snape and Watt (n 68) 153; John Snape and Gary Watt, *How to moot: a student guide to mooting,* (OUP 2004) 177.

75 *Royal College of Nursing of the United Kingdom v Department of Health and Social Security* [1981] AC 800 [1981] 2 WLR 279 [1981] 1 All ER 545 (1981) 125 SJ 149 1981 WL 187265.

76 Michael Davies, *Textbook on Medical Law* (Oxford University Press 1998) 281.

Judges have always a moral choice.[77] Singer supports this theory by the reference to Checo:

> Man is the builder of a historical edifice: the House of man. He is the brick and the firm foundation of his own project...Man is the player and the cards; he is at stake but he repeats with Oedipus: "I will search out the truth."[78]

Although, Dworkin, due to imperfections of legal system, was incorrect in that there is always one correct answer, his thoughts about the judges' discretion and approaches are insightful. Even though, Hercules' supernatural attributes cannot be seen in every judge, there are judges who come close to this ideal. Therefore, although, I agree with the criticism of Dworkin's theses of his idealistic vision of the legal system, I also see the point advocated by Altman, who acknowledges strengths of Dworkin's jurisprudence in its potential to adopt 'the realist indeterminacy analysis to his advantage'[79]. It may be arguable that Dworkin's arguments are not that compelling, but it is plain that

> Hercules...adopts a cavalier attitude towards rules. However, merely to establish that he enjoys such a freedom does not provide an answer to the equally important question of what he ought to indulge in this freedom...[H]e must decide whether in the case before him the disagreement between the parties is genuine and, therefore, makes it into a "hard case". The only reason that Hercules...would wish to avoid a rule-dictated result is his anticipation that such a

77 Singer (n 45) 1.

78 Marcel Pallais Checa, 'Sketches on Hegel's Science iv' (June 1977) (unpublished manuscript on file in Sawyer Library at Williams Collage, Williamstown, Mass).

79 Andrew Altman, 'Legal Realism, Critical Legal Studies and Dworkin' (1986) 15(3) *Philosophy and Public Affairs* 205, 212, Andrew Altman, *Critical Legal Studies: A Liberal Critique* (Princeton University Press 1950).

> result would in some way be undesirable or unacceptable.[80]

Nevertheless, Altman in the cited article misunderstood Hart's theory, who implicitly in *The Concept of law* [81] and explicitly in its "Postscript" [82] accepts incorporationism.[83] Therefore, since Hart comes close to the Dworkin in the "Postscript", I believe that both theories provide a useful alternative to extremes advocated by CLS.

Unfortunately, in this more ambitious concept that judges do not make new law in "hard cases", we can find incoherence, which questions whether judges make new law. However, the incoherence is not the result of the adopted supernatural vision of the judge, but the imperfection of the concrete legal systems as such.

The only plausible criticism that can be aimed to the vision of the judge in "one answer model" is the structural incoherence in his moral convictions. In relation to Dworkin theory, Kennedy points out incoherence in the lack of "metacriterion" between the choice of political theories and the notion of coherence advocated. According to Kennedy, the only possible "metacriterion" is the judge's *personal* conviction, which is the only way to decide among the possible theories.[84] Such vision of this "metacriterion" significantly undermines Dworkin's notion that judges do not make new law.[85] Langemeijer advocates that the coincidence between "judicial intuition"[86] and "consensus value"[87] is to

---

80 Hutchinson and Wakefield (n 6) 99.

81 Hart (n 22).

82 ibid, 250-254; Brian Bix, 'Inclusive Legal Positivism and the Nature of Jurisprudential Debate' (1999) 12 Canadian J of Law and Jurisprudence 17.

83 Freeman (n 10) 334; Hart (n 22) 250-254; Joseph Raz, 'Authority, Law and Morality' (1985) 68 Monist 295, 295-324, Scott J Shapiro, 'On Hart's Way Out' (1998) 4(4) Legal Theory 469, 469-507; Jules L Coleman, (1998), 'Incorporationism, Conventionality, and the Practical Difference Thesis' (1998) 4(4) Legal Theory 381, 381-425.

84 ibid, 36.

85 Roland Dworkin, *Law's Empire* (Hart Publishing 1998) 6-11.

86 John Bell, *Policy Arguments in Judicial Decisions* (Clarendon Press 1983) 199.

only possible solution to "hard cases". Raz also spots incoherence in Dworkin's theory, namely the advocated by Dworkin postulate that '[r]ules that do not pass the test of integrity are not part of the law'[88] and the fact that 'the courts ... [cannot] compromise justice and fairness for the sake of integrity.' [89] Raz believes that this inconsistence of two principles shows that, probably all the time in "hard cases", courts cannot decide cases according to law.[90] A similar viewpoint is taken by MacCormick.[91] Conversely, Lucy, while referring to Kennedy's critique of Dworkin, says that the conclusion that "hard-case" adjudication 'turn upon considerations of fit and arguments that show the law in its best moral and political light ... is [itself] uncontested.'[92] I believe that this degree of uncertainly is acceptable. Nevertheless, an interesting trial in resolving this incoherence is given by McDowell[93] and Hurley[94], who indirectly referring to Wittgenstein's idea of a 'form of life'[95], persuade us 'that meaning does not come from self-interpreting entities, but that it derives in part from the practices, customs, and institutions in which the speaker participates.' [96] Furthermore, I think that the matter could be successfully resolved by the existential conception of 'a choice of being'[97]:

---

87 ibid.

88 Joseph Raz, (2004) 'Speaking with One Voice: On Dworkinian Integrity and Coherence' in Justine Burley (eds) *Dworkin And His Critics* ( Blackwell Publishing 2004), 287.

89 ibid.

90 ibid.

91 MacCormick, (n 22) ch 5-8.

92 Lucy (n 28) 299.

93 John McDowell, 'Non-Cognitivism and Rule-Following' in Steven Holtzman and Christopher M Leich (eds), *Wittgenstein: To Follow A Rule* (Routledge 1981), 160.

94 Susan L Hurley, 'Natural Reasons: Personality and Polity' (1990) 65(254) Philosophy 528, 528-530.

95 Wittgenstein (n 22), secs 185-7.

96 Brian Bix, (1993) *Law, Language and Legal Determinacy*, Clarendon Press, Oxford, 53-59.

97 Jean P Sartre, *Notebooks for an ethics* (The University of Chicago, 1992), 559.

'The master's caprice will be condemned by the virtuous slaveholder...In this moral hierarchy, perfection is to know one's place.'[98]

The greatest problem with Dworkin's theory is probably his vision of the legal system and his insistence that the law is determinate. This notion of determinacy is indispensable if we ever try to assume that judges merely apply the law in "hard cases".[99] He seems to misunderstand one fundamental issue - the difference between a pair of two *substantially* diverse concepts and the distinction between a pair of *logically contradictory* concepts (F and ˜F). The former occurs when two concepts

> are mutually exclusive...[but] they are not jointly exhaustive, so that there will be a logical gap between their boundaries. And it might be the case that some pairs of legal concepts are of that kind, so that, e.g., it is both false that a particular contract is valid and false that it is not, both false that a particular act constitutes a crime and false that it does not, etc.[100]

Dworkin always incorrectly considers the latter, when, in reality, judges, while judging "hard cases" usually face the former. Therefore, if we assume that most of the time, in hard-case adjudication, judges do face such a logical gap then it is a misunderstanding to maintain that they merely apply the law, because in such scenario there is no law at all. This problem cannot be successfully answered.

The second aspect that undermines Dworkin's thesis that judges do not make law in "hard cases" is the structure of precedent itself. Judges have discretion over the holdings in "hard cases"; therefore, as Altman argues they can generate

---

98 ibid, 565.

99 Joseph (n 45) 12.

100 Anthony D Woozley, 'No Right Answer' (1979) 29 The Philosophy Quarterly 25, p 26; Wilfred Hodges, *Logic: an introduction to elementary logic* (Penguin 1986), p 17-38; Jeffrie G Murphy, (1967) 'Law Logic', 77(3) Ethics 193, 193-201.

different rules of law capable of producing conflicting results in the same case.[101]

The issue of logical gap is interlinked with the doctrine of precedent.

Dworkin's notion that principles, alongside rules, are part of the legal system does not avail him a lot, because as pointed out, by Hart and subsequently by MacCormick, there is no rigid distinction between rules and principles. In relation to Dworkin's postulant that there is a difference of conclusiveness between principles and rules, namely that rules applies in an all-or-nothing fashion, whereas principles are non-conclusive[102], Hart advocates that Dworkin cannot be coherent and that 'a principle will sometimes win in competition with a rule and sometimes lose', which shows that rules do not operate in all-or-nothing mode, as postulated by Dworkin. Ironically, Hart illustrates this observation on the case used by Dworkin to illustrate the operation of the principles - *Riggs v Palmer*[103]. In this case,

> the principle that a man may not be permitted to profit from his own wrongdoing was held notwithstanding the clear language of the statutory rules governing the effect of a will to preclude a murderer inheriting under the victim's will... Even if we describe such cases (as Dworkin at times suggests) not as conflicts between rules and principles, but as a conflict between the principle explaining and justifying the rule under consideration and some other principle, the sharp contrast between all-or-nothing rules and non-conclusive principles disappears.[104]

---

101 Altman (n 79), 209.

102 Dworkin (n 5) chapter 2, Perry (n 29) 223.

103 *Riggs v Palmer* 115 NY 506, 22 NE 188 (1889); Dworkin (n 5); Dworkin (n 85) 15.

104 Hart (n 22) 262.

This point is also supported by Raz [105] and Waluchow[106]. Therefore, since Dworkin's notion for the distinction between rules and principles is at least incoherent, there is little justification in supporting his theory in "hard cases", where both rules do not have direct application and principles come into play. Even if we accept his inclusive theory, which in acknowledging principles does not differ a lot from 'inclusive positivism' [107], the problem of the interrelation between principles and rules is still present. His notion that rules operate in all-or-nothing fashion and that principles are non-conclusive simply does not work. Therefore, I cannot accept the idea that judges in cases such as *Riggs v Palmer*[108] do not make a new law is simply unrealistic. Nevertheless, I see Dworkin's point in his response to those criticism that when the judge is acting to achieve some "purpose" and 'his ambitions are complex and competing...[,] he must sometimes neglect one to serve another.' [109] Therefore, I would not incline to reject his theory completely on the ground of imperfection of the system.

## VI. Conclusions

Due to the incompleteness of the legal system, the fact that judges occasionally make law in "hard cases" is undeniable. The question that needs to be resolved is the actual process by which they engage in making new law. However, the fact that the judges make new law in "hard cases" is only the result of the indeterminacy and imperfection of a legal system, which has been aptly noticed in the logic account of CLS. Those observations significantly

105 Joseph Raz, 'Legal Principles and the Limits of the Law' (1972) 81 Yale LJ 823, 823-4.

106 Wilfred J Waluchow, 'Herculean Positivism' (1985) 5 Oxford Journal of Legal Studies 187, 189-92.

107 Freeman (n 10) 334.

108 *Riggs v Palmer* (1889) 115 NY 506, 22 NE 188; Dworkin (n 5) 23; Roland Dworkin (n 85) 15.

109 Justine Burley, *Dworkin And His Critics* (Blackwell Publishing 2004) 361.

undermine Dworkin's theory. However, the movement should never be justified empirically. The reverted nature and purpose of adjudication, even in "hard cases", cannot be based on *a priori* knowledge. The simplest notion of experience does not presuppose anything other than experience. That is why, '[a]s the principle of individuation Kant took time and space, for no object, he insisted, can be considered as existing of both or either. [sic.]' [110] This postulate can be aptly applied to the adjudication of "hard cases", where a superior aim is invoked. Simple observations of human nature will not suffice. Empirical scrutiny of the psychological judicial approach towards adjudication cannot be justified on firm scientific grounds. The consecrated nature of adjudication requires a more holistic approach, and it cannot be blemished by quasi-empirical generalisations. In this respect, Dworkin's theory of adjudication in "hard cases" remains firm and can be supported by existential accounts.[111] Lucy suggests that we need to draw a line between orthodoxy and heresy. Such a distinction could only be made if we make certain assumptions. Lucy refers as to the assumptions advocated by Dworkin on the face of knowledge implicit in the community's institutional political morality [112], that is firstly, that the kingdom of the 'political' surpasses party politics or interest group disagreements; secondly that the kind of higher degree or more abstract political preferences judges make does not trespass upon the law/politics divide or the prerequisite of judicial impartiality. [113] Therefore, it should not be surprising that Altman called Dworkinian jurisprudence a more advanced answer to realism than that

110 Olin McKendree Jones, *Empiricism and Intuitionism in Reid's Common Sense Philosophy* (Princeton University Press 1972) 1.

111 Georg Cohn, *Existenzialismus und Rechtswissenschaft* (1959) 53(3) *The American Journal of International* Law718, 718-719; Matthew L Williams, *Empty Justice: One Hundred Years of Law, Literature and Philosophy – Existential, Feminist and Normative Perspectives in Literary Jurisprudence*, (Cavendish Publishing 2002).

112 Lucy (n 28) 299.

113 ibid.

advocated by Hart.[114] Furthermore he acknowledged that Dworkin's "soundest theory of law" is the most justifiable ethical and political theory that fits together and explains the norms and choices adherent to the law already decided. The consistency does not have to be ideal, for Dworkin agrees that some of the decided legal judgments may be considered as mistakes. But consistency is necessary with a considerable dose of the decided case law. In the lack of a single, overarching theory that deals with the decided law – and Dworkin believes that there will often numerous theories in hard cases – then most appropriate theory is the one that is both fit and ethical adequate.[115] There is truth in the postulate that there is a social and moral need in the assumption that judges, even in "hard cases" merely apply the law and they are far from creating new law by referring to moral and political judgments, but since the law is indefinitely indeterminate, it is unfortunately true that, as Singer says, our legal system will never come close to this aim.[116] But it is too hasty to agree with him that 'the traditional goal is false or irrational.'[117] The judges' approaches, both truthful (*Duport*[118]) and hypocritical (*Congreve*[119], *Royal College of Nursing*[120]), show that such an unobtainable goal is right and sound. When, single judges get it wrong, it does not mean that the aim is hopeless, but merely that society has rather blemished.[121] Camus justified this goal in reference to 'the noble profession of lawyer'[122]. He believed that 'we ... [are] of the same species. Are we not all like, constantly talking

114 Altman (n 79) 207.

115 Altman (n 79) 35-36.

116 Singer (n 45) 13.

117 ibid 8.

118 *Duport Steels Ltd v Sirs* [1980] 1 All ER 529, 521.

119 *Congreve v Home Office* [1976] QB 629.

120 *Royal College of Nursing of the United Kingdom v Department of Health and Social Security* [1981] AC 800 [1981] 2 WLR 279 [1981] 1 All ER 545 (1981) 125 SJ 149 1981 WL 187265.

121 Camus (n 1) 6.

122 ibid.

and to no one, for ever up against the same question although we know the answer in advance?'[123] I think that those words clearly support Doworkin's vision of adjudication in "hard cases". In order to achieve this, judges while judging "hard cases" try to escape the bad faith of personal political and moral values by 'trying to make oneself nothing but the role demanded by society – to be *only* a waiter or a conductor or a mother, *only* an employer or a worker,'[124] only a judge.

123 ibid.

124 Jean P Sartre, *Existential Psychoanalysis* (Philosophical Library1953) 30.

## BIBLIOGRAPHY

### Articles

Altman A, 'Legal Realism, Critical legal Studies and Dworkin' (1986) 15(3) *Philosophy and Public Affairs* 205

Bix B, 'Inclusive Legal Positivism and the Nature of Jurisprudential Debate' (1999) 12 *Canadian J of Law and Jurisprudence* 17

Cohn G, 'Existenzialismus und Rechtswissenschaft' (1959) 53(3) *The American Journal of International Law* 718

Coleman JL, 'Incorporationism, Conventionality, and the Practical Difference Thesis' (1998) 4(4) *Legal Theory* 381

Fletcher GP, 'Two Modes of Legal Thought' (1981) 90 *Yale LJ* 970

Fuller LL, 'Forms and Limits of Adjudication' (1990) 92 *Harv L Rev* 353

——, 'Reason and Fiat in Case Law' (1946) 59 *Harv L Rev* 376

——, *Positivism and Fidelity to Law – A Reply to Professor Hart*, (1958) 71 *Harv L Rev* 630

Hart HLA, 'Positivism and the Separation of Law and Morals' (1958) 71 *Harv L Rev* 593

——, 'Law in the Perspective of Philosophy: 1776-1976' (1976) 51 *NYUL Rev* 538

Hurley SL, 'Natural Reasons: Personality and Polity' (1990) 65 (254) *Philosophy* 528

Hutchinson AC and Wakefield JN, 'A Hard Look at "Hard Cases": The nightmare of a noble dreamer' (1982) 2 *Oxford J Legal Studies* 86

Lucy W, 'What is Wrong with Ideology?' (2000) 20 *Oxford J L Studies* 283

MacCormick N, 'Reconstruction after Deconstruction: A Response to CLS' (1990) 10(5) *Oxford J L Studies* 539

Murphy JG, 'Law Logic' (1967) 77(3) *Ethics* 193

Pallais Checa M, 'Sketches on Hegel's Science iv' (June 1977) (unpublished manuscript on file in Sawyer Library at Williams Collage, Williamstown, Mass)

Perry SP, 'Judicial Obligation, Precedent and the Common Law' (1987) 7(2) *Oxford JL Studies* 215
Raz J, 'Authority, Law and Morality' (1985) 68 *Monist* 295
——, 'Legal Principles and the Limits of the Law' (1972) 81 *Yale L J* 823
Schlag P, 'No Vehicles in the Park' (1999) 23 *Seattle Uni L Rev* 381
Shapiro SJ, 'On Hart's Way Out' (1998) 4(4) *Legal Theory* 469
Sinclair MBW, 'Hercules, Omniscience, Omnipotence, And the Right Answer Thesis' (2003) 46 *New York Law School L Rev* 447
Singer J, 'The Player and the Cards', (1984) 94 *Yale L J* 1
Waluchow WJ, 'Herculean Positivism' (1985) 5 *Oxford J L Studies* 187
Woozley AD, 'No Right Answer' (1979) 29 *The Philosophy Quarterly* 25

## Books

Altman A, *Critical Legal Studies: A Liberal Critique* (Princeton University Press 1950)
Bailey SH and Gunn MJ, *Smith and Bailey on The Modern English Legal System* (3rd Edn, Sweet & Maxwell 1996)
Bell J, *Policy Arguments in Judicial Decisions* (Clarendon Press 1983)
Bix B, *Law, Language and Legal Determinacy* (Clarendon Press 1983)
Burley J, *Dworkin And His Critics* ( Blackwell Publishing 2004)
Camus A, *The Fall* (Penguin Books 1957)
Chambers JH, *Empiricist Research on Teaching: a Philosophical and Practical Critique of its Scientific Pretensions* (Kluwer 1992)
Cross R, *Precedent in English Law* (3rd Edn, Clarendon Press 1977)
Dworkin R, *Law's Empire* (Hart Publishing 1998)
——, *Taking Rights Seriously* (Duckworth 1978)

Finnis J, *Natural Law and Natural Rights,* (Clarendon Press 1980)

Freeman MDA, *Lloyd's Introduction to Jurisprudence* (7th edn, Thomson Sweet and Maxwell 2001)

Hart HLA, *The Concept of law* (2nd edn, OUP 1997)

Hodges W, *Logic: an introduction to elementary logic* (Penguin 1986)

Horwitz DL, *The Courts and Social Policy* (The Brookings Institute 1977)

Kennedy D, *A Critique of Adjudication* (Harvard University Press 1997)

Levi EH, *An Introduction to Legal Reasoning* (University of Chicago Press 1977)

Lucy W, Understanding *and Explaining Adjudication* (OUP 1999)

MacCormick N, 'On the "Internal Aspect" of Norms' in MacCormick N (ed) *Legal Reasoning and Legal Theory* (OUP 1994)

——, *Legal Reasoning and Legal Theory* (OUP 1993).

McCoubrey H and White ND, *Textbook on Jurisprudence* (3rd edn, Blackstone Press 1999)

McDowell J, 'Non-Cognitivism and Rule-Following' in Holtzman S and Leich CM (eds), *Wittgenstein: To Follow A Rule* (Routledge 1981)

McKendree Jones O, *Empiricism and Intuitionism in Reid's Common Sense Philosophy* (Princeton University Press 1972)

Michael Davies M, *Textbook on Medical Law* (OUP 1998).

Murphy WF, *Elements of Judicial Strategy* (University of Chicago Press 1977)

N Adams JN and Brownsword R, *Understanding law* (4th edn, Thomson Sweet and Maxwell 2006)

Northrop FSC, *The Complexity of Legal and Ethical Experience* (Little Brown & Co 1956)

Pettit P, *The Common Mind* (OUP 1996)

Raz J, (2004) 'Speaking with One Voice: On Dworkinian Integrity and Coherence' in Burley J (ed) *Dworkin And His Critics,* (Blackwell Publishing 2004)

Rebelais F, *Gargantua and Pantagruel* (Penguin 1993)
Rhenstein M, *Max Weber on Law in Economy and Society* (Harvard University Press 1954)
Sartre JP, *Being and Nothingness: An Essay on Phenomenological Ontology* (Methuen & Co 1972)
Sartre JP, *Existential Psychoanalysis* (Philosophical Library 1953)
—, *Notebooks for an ethics* (The University of Chicago 1992)
Snape J and Watt G, *How to moot: a student guide to mooting* (OUP 2004)
—, *The Cavendish guide to mooting* (2nd edn, Cavendish Publishing Limited 2000)
Twining W and Miers D, *How to do Things with Rules* (3rd edn, Weidenfeld and Nicolson 1991)
Unger R, *Knowledge and Politics* (Free Press 1976)
Wasserstrom RA, *The Judicial Decision* (Stanford University Press 1961)
Williams ML, *Empty Justice: One Hundred Years of Law, Literature and Philosophy – Existential, Feminist and Normative Perspectives in Literary Jurisprudence* (Cavendish Publishing 2002)
Winch P, *The Idea of a Social Science and its Relations to Philosophy* (Routledge 1958)
Wittgenstein L, *Philosophical investigations* (2nd edn, Blackwell 2003)

## Cases

*Tennessee Valley Authority v Hill* (1978) 437 US 153
*Rex v Taylor* [1950] 2 KB 268
*Fisher v Bell* [1961] 1 QB 394
*Smith v Hughes* [1960] 1 WLR 830
*Duport Steels Ltd v Sirs* [1980] 1 All ER 529
*Congreve v Home Office* [1976] QB 629
*Riggs v Palmer* (1889) 115 NY 506, 22 NE 188
*Royal College of Nursing of the United Kingdom v Department of Health and Social Security* [1981] AC 800

## Other Sources

HC 52-II (1995), 130
*The Times* (London, 6 and 9 December 1925)

# Corporate Takeovers And Shareholder Protection: UK Takeover Regulation In Perspective

Francis Okanigbuan

## Abstract

*This article examines the regulatory framework for shareholder protection during takeovers in the UK. It is aimed at ascertaining the extent that takeover regulations protect shareholders from company managements who may pursue objectives that are different from enhancing shareholder value. Pursuant to this, a narrative of the historical development of takeover regulation in the UK is provided to show why takeover regulation emerged. Using doctrinal legal analysis, the current takeover regulations The EU Takeover Directive [2004] and The City Code on Takeovers and Mergers (as amended in 2011) are examined. It emerged that these regulations were developed to protect only the shareholders of acquired companies. While this represents an improvement towards shareholder protection, the essay argues that the regulations do not provide a complete protection to shareholders of target companies as managements may promote their own interests using pre-bid defences. Aspects of directors' duties were examined to fill the regulatory gaps in this regard, but the duties appear not to provide any remedy to shareholders, since they do not specifically apply to takeovers. Further, in the absence of a specific regulation for protecting shareholders of acquiring companies, the derivative action procedure was examined as an alternative form of remedy. But since derivative actions are based on wrong done to a company, it is difficult for shareholders to rely on this. Thus, apart from strengthening the current takeover regulations that protect the shareholders of acquired companies, it is also imperative that a framework for protecting the shareholders of acquiring companies should be developed.*

## I. Introduction

The main objective of a company as a going concern[1] is to enhance its economic values through strategic investment decisions of its management team. Generally, an improvement in the economic value of a company is largely dependent on the extent to which the investments of its individual investors or shareholders are actually enhanced. A corporate takeover is one of the ways of enhancing the economic value of a company. It involves an auction in which prospective investors - usually a company - bid for the right to obtain control of another company.[2] A successful bid leads to a combination of the assets of the acquiring and the acquired companies, thereby leading to a higher economic value of the combined company than the individual values of the separate companies before the acquisition. But, the extent to which takeovers operate to promote the values of shareholders is unclear. Takeovers have different functions, including its disciplinary role[3] and synergy gains.[4] Takeovers may also lead to hubris,[5] when management pursue acquisitions that lead to an expansion of the corporate size, without corresponding economic gains to

---

1 'Going concern' is the ability of a company to make enough money so that it can continue to do business and avoid bankruptcy.

2 For a definition of takeovers, see: Nikhil P Varaiya, 'The 'Winner's Curse' Hypothesis and Corporate Takeovers' (1988) 9 Managerial and Decision Economics 209, 210.

3 Takeovers are a disciplinary tool for a poorly performing management board of a company. Often, a takeover leads to the dismissal of company managements of acquired companies who have failed to improve the performance of their companies, thus making their companies to be easy targets for takeovers. See: Richard A Brealey, Stewart C Myers et al, *Principles of Corporate Finance* (McGraw-Hill/Irwin New York 2008) at 887.

4 Takeovers have synergistic functions; they lead to a combination of the assets and operations of the acquired and acquiring companies. See generally: Lynn Hodgkinson and Graham L Partington, 'The Motives for Takeovers in the UK' (2008) 35 JBFA 102.

5 Takeovers may not lead to any economic gains to the acquired company, possibly caused by paying too much (over-payment) to acquire a company, which often leads to losses or zero-gains to the acquiring companies. See, Richard Roll, 'The Hubris Hypothesis of Takeovers' (1986) 59 JBF 197.

the company. The importance of these functions depends on whether a company is an acquiring company or a target company. While takeovers remain an important aspect of the market for corporate control,[6] company managements have largely played a central role in takeovers.

The relationship between company management and shareholders has been described as one of agent and principal respectively.[7] This relationship often gives rise to a conflict of interests, especially during takeovers. While management may be interested in retaining their positions in the company and extending the size of the company through acquisitions, shareholders are actually interested in an increase in the value of their shares. In recognition of these challenges, directors' duties have now been codified to generally enhance levels of accountability.[8] Also, the derivative action procedure provides an opportunity for shareholders to hold their company management to account for their role in enhancing the interest of the company.[9] More specifically, takeover regulations have been developed that seek to limit the extent to which company management influences the outcomes of takeovers.[10] However it remains doubtful whether the objectives of the codified duties of directors, derivative actions and takeover regulations have achieved these aims.

---

6 Corporate control includes measures through which the role of company managements is influenced. It has been suggested that the market for corporate control serves as an alternative means for enhancing the performance of company management where the internal control and corporate governance measures fail. See Henry G Manne, 'Mergers and the Markets for Corporate Control' (1965) 73 JPE 110, 112-120.

7 Shareholders as principals appoint the company management to manage the business as agents. See generally: Michael C Jensen, & William H Meckling 'Theory of the Firm: Managerial Behaviour, Agency Cost and Ownership Structure' (1976) 3 JFE 305.

8 Companies Act 2006 ss 171-177.

9 See Companies Act ss 260-269.

10 Council and European Parliament Directive 2004/25/EC of 21 April 2004 on Takeover Bids [2004] OJ L142/12 ('The Takeover Directive'), and The UK City Code on Takeovers and Mergers 2013 ('The Takeover Code').

In view of this, this paper examines corporate takeover regulation in the United Kingdom with a particular reference to shareholder protection. The objective here is to ascertain the extent to which the interests of the shareholders of the target and acquiring companies are effectively protected during takeovers. In section two, a brief historical development of takeovers in the United Kingdom will be highlighted. This is aimed at ascertaining the relationship between corporate managements and company shareholders during the takeovers that triggered the emergence of takeover regulation. Section three examines the extant takeover regulation in the United Kingdom. The combined effects of the Takeover Directive and the Takeover Code will be evaluated with particular reference to shareholder protection during takeovers as it affects shareholders in the target and acquiring companies. The duties of company directors under the Companies Act 2006 will be examined in section four. While these duties are not particularly directed at takeovers, they nevertheless describe the acceptable standard of behaviour that directors should conform to when executing their general functions. The extent to which company shareholders can effectively hold management to account for decisions that are made pursuant to takeovers through derivative actions will also be examined. Afterwards, section five concludes the paper.

## II. Historical Development of Takeover Regulations in the United Kingdom

Corporate takeovers in the UK have been a recurrent phenomenon from the early 1950s to recent times. Throughout these periods, takeovers have been fuelled by the desire of external investors to acquire and invest in companies which they believe could be better run to achieve a more desirable economic value than the present output of

the current managers.[11] Whilst the current management of a target company may become an obstacle to achieving this aim, this could be overcome by the dissatisfaction of the shareholders. Often, shareholders view the interests of external investors in acquiring their companies as an avenue to divest their holding and, more importantly, to profit from their investments in light of unfavourable investment decisions of their company managements. These considerations influenced the first successful takeover in the UK in 1953.[12] The shareholders of the company felt short-changed because the value of their investment was reflected not in the share price, but on dividends, and the rate of dividend payments was substantially low when compared to the company's profit margin. Consequently, a majority of shareholders at that time ignored promises made by the board to increase the rate of dividend and sold their shares to the investor, leading to what became the first successful corporate takeover in the UK.[13]

Later in the same year, investor Harold Samuel sought to acquire *Savoy Hotel Ltd.* The intention of the investor was to acquire the hotel and convert it to office premises. The management of the company were opposed to the takeover attempt and this led to a conflict that involved the management, the shareholders and the investor. The shareholders were disappointed that the management opposed the takeover bid and so they sought to sell their shares to the investor. However, to prevent the company from being acquired by the investor, the management of *Savoy Hotel* arranged for the hotel property to be sold to

---

11 Ronald W Moon, 'Business Mergers and Takeover Bids: A Study of the Postwar Patterns of Amalgamatons and Reconstruction of Companies' (Gee & Co, 1976) at 9-10.

12 The 'J Sears Holdings' Later acquired by *Charles Clore*.

13 A similar tactics used by the board of a company that was a takeover target by *Charles Clore* had earlier been successful. See John Armour and David A Skeel Jr, 'Who Write the Rules for Hostile Takeovers and Why? - The Peculiar Divergence of US and UK Takeover Regulation' (2006-2007) 95 The George Town Journal, 1727 at 1757 (fn 125).

*Worcester (London) Co. Ltd.* The company management made an agreement with the new company that the hotel should be leased back to them - the current management - on terms that the property should only be used for the purposes of a hotel business and not be converted into offices. This effectively frustrated the takeover attempt of Harold Samuel. The decision of the hotel management was made without reference to shareholder interest and this infuriated them further since they were powerless to change this decision.

Public concern led the Board of Trade to investigate the conduct of the directors, but as the report of the Board was not binding, company boards had an air of invincibility during takeovers.[14]

Conflict of interests during takeovers reached its peak when two different investors, *Reynolds Metal Company* in partnership with UK-based *Tube Investments ("TI-Reynolds")* and *Aluminium Company of America (ALCOA)* sought to acquire *British Aluminium Ltd.* The board of *British Aluminium* accepted the bid of one of the bidders and offered them a third of the shares in the company without the approval of their shareholders. The shareholders only became aware when the rival bidder informed the management of their intentions to deal with shareholders directly, and in an attempt to placate the shareholders the board offered to increase dividend payments, leading to an increase in shareholder value.[15] However, the shareholders became even more infuriated in view of the fact that the deal that the board had made with the other bidder was concluded at an undervalued price. This prompted shareholders to dispose of their shares to the rival bidder in anger. This sparked a widespread call for takeover regulations based on certain takeover principles,

14 ibid, 1757, citing 'Battle for the Savoy' *The Economist* (London, 12 December 1953).

15 ibid 1758 citing 'Dividend Raised to Counter Bid' *The Times* (London, 20 December 1958) 6.

such as board neutrality and shareholder primacy.[16] This led to the humble beginning of takeover regulation.

The historical development of takeovers in the UK revealed that company managements and their shareholders consistently disagreed during takeovers. It also revealed that managements have the capacity to determine the outcomes of takeovers independently of shareholders since as they occupy positions of authority that enable them to determine whether a takeover bid is accepted or rejected. However, it is arguable as to whether managements act in pursuit of their personal interests or for the interests of the company. Usually, a board may oppose a takeover bid to encourage the bidder to increase their offer. They may oppose bids to signal to the market that a bid has been made and encourage other bidders leading to a price war for the company.[17] These are clear examples of managements pursuing the interest of the company, particularly those of the shareholders. But this is not always the case. Management may oppose bids for fear of losing their positions. They may oppose a bid even where there are economic benefits that could be derived by the shareholders of their company. However, irrespective of the objective of the actions of management during takeovers, it is imperative that shareholders are actively involved in decisions leading to the outcome of a bid. The early period of takeovers was characterised by managerial decisions that disregarded the input of shareholders, even though their shares were the subject matter of takeovers. Consequently, it became imperative that shareholders must be protected from managerial excesses, meaning that takeover regulation was introduced to protect the interests of shareholders.

---

16 ibid 1759.

17 Thomas W Bates and David A Becher, 'Bid Resistance by Takeover Targets: Managerial Bargaining or Bad Faith? (2011) Chicago Meetings Paper, AFA/2012 1-49 at 29 <http://dx.doi.org/10.2139/ssrn.1786674> accessed 11 April 2013.

After the introduction of the regulation of takeovers in 1968,[18] it continues to be a prominent feature of the life of companies as a going concern, and it has indeed become an important function of the market for corporate control. Despite the existence of takeover regulations, the conflict of interests that characterised the relationship between corporate management and their shareholders during the early periods of takeovers remains a challenge. Although, the powers of management during takeovers have been reduced, they still possess significant discretionary powers that are capable of undermining the objectives of takeover regulation.[19] Thus, the extent to which the takeover regulations can effectively prevent company managements from promoting their personal interests during takeovers is largely unclear. The next section examines the extant takeover regulatory mechanisms.

## III. Shareholder Protection under the EU Takeover Directive and the UK City Code on Takeover

The events stated above led to demands for takeover regulations, which at the core contained the principles of shareholder protection and board neutrality.[20] The regulation took the form of the City Code on Takeovers and Mergers. This was supplemented with the EU Takeover Directive in 2004, which regulated takeovers within the EU. Thus takeovers in the UK are regulated by the combined effect of both sets of regulations. These restrict the roles of company management during takeovers to ensure that shareholders are not denied the opportunity of deciding on

18 The Takeovers Code is administered by The Panel on Takeovers and Mergers.

19 This discretionary powers can be used as takeover defences measures, especially where management are not favourably disposed to a takeover bid.

20 The *board neutrality principle* and *shareholder primacy* is meant to ensure that the management board of target companies do not use their positions of authority to influence a takeover bid. They are to ensure that the shareholders of their companies give prior approval to their decisions concerning a bid. See the UK Takeover Code 2013, *General Principles B1* (3); The Takeover Directive, Art 9 (2).

the merits of a takeover bid. Although these regulations promote the interests of company shareholders during takeovers, they have a restricted scope of application. This gives room for corporate management to unduly interfere with takeovers at the expense of the interests of their shareholders.

The other issue is that, despite takeovers involving both the acquiring and target companies, debate on takeover regulations has largely considered the effect on shareholders and the conduct of the management of the target company. This has been at the expense of considering takeovers from the perspective of the acquiring company, especially their shareholders. In light of this, the extent to which shareholders are protected from managerial excesses during takeovers will be discussed from the perspective of both target and acquiring companies. In pursuance to this, the combined effects of the EU Takeover Directive and the UK City Code on Takeovers will be examined.

### *A. Shareholder Protection in Target Companies*

Largely, the scope of shareholder protection appears to be restricted to the shareholders of target companies. This can be deduced from the principal objectives that underlie the EU Takeover Directive[21] and the UK City Code on Takeovers.[22] Under these regulations, the management of target companies are prevented from making any decision that is capable of indicating that a takeover bid has been accepted or rejected by the company without the approval of the shareholders of the company.[23] The essence of this non-frustration rule is to prevent managements from doing

21 The Takeover Directive, Art 3 *(General Principles).*

22 The purpose of the code is indicated to protect the shareholders of the offeree (target) company: 'The Code is designed principally to ensure that shareholders in an offeree company are treated fairly and are not denied an opportunity to decide on the merits of a takeover and that shareholders in the offeree company of the same class are afforded equivalent treatment by an offeror.' See: Nature and Purpose of the Code, Introduction 2 (a).

23 The Takeover Directive, Art 9 (2 & 5), The Takeover Code, rule 21.

anything that will discourage a bidder from continuing with a takeover attempt without the authority of shareholders. Managements are meant to play advisory roles only in the determination of whether a takeover bid should be accepted or rejected by the company. While this represents the objective of the regulatory functions of takeovers, it remains to be seen whether this objective can be achieved in light of the operative provisions of the regulatory mechanisms. Management can actually 'go round' the restrictions contained in takeover regulations to enhance their own objectives in the following ways.

Firstly, managements are required not to conduct themselves or carry out any action that may enhance or mitigate the chances of a bid, they are also required not to carry out any positive action towards enhancing the interest of shareholders, other than making available competent independent advice on the fairness of a bid.[24] As such, they may misrepresent the true state of affairs and mislead shareholders based on advice provided in bad faith. But, they are not likely to be held responsible for any unpleasant outcome based on such advice.[25] While it may not be inferred that company management deliberately refuses to promote the interests of their shareholders during takeovers, the fact that management are keenly interested in the outcome of a takeover bid[26] provides a sufficient reason to supervise their role during takeovers. But since their opinion regarding a takeover bid is not subject to any external review, and their shareholders are highly likely to rely on their opinions, they may subtly influence the outcome of takeovers.

Secondly, the non-frustration rule of takeovers, which restricts the role of management during takeovers, may

24 The Takeover Code, rule 3 (1).

25 Blanaid J Clark, 'The Takeover Directive: Is a little Regulation Better than No Regulation?' (2009) 15 ELJ 174 at 188.

26 Managements are interested in the outcome of takeover bids because they are likely to be dismissed post-takeover.

not effectively prevent takeover bid defences, especially pre-bid defences. This rule adds nothing to the existing obligations of directors, since they are already required under company law to act in good faith in fulfilling their obligations.[27] In anticipation of the restrictive roles which company management are expected to play during takeovers for which they are aware, they can devise a means of limiting the extent to which the rule applies to them. Since the non-frustration rule only applies when takeover bids have been made, certain pre-bid defensive mechanisms can be adopted by the board that take effect when a takeover bid is made.[28] These include lucrative compensation packages based on contracts of employment, which serve to compensate company management should their positions be terminated by a change of control. This operates to make a takeover more expensive for the offeror, as they would incur the cost of these compensation packages. Also, a staggered board appointment procedure[29] or dual class voting stock[30] may be adopted. Although most of these measures appear to have been largely restricted by corporate governance principles[31]

---

27Paul Davies and Edmund-Philip Schuster et al, 'The Takeover Directive as a Protectionist Tool' (2010) ECCG Working Paper Series in Law 141/2010, 1 at 4. <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1554616> accessed 10 April 2013.

28 See particularly: The Takeover Directive, introductory paragraph 2, and Article 9, Rule 2; UK Takeover Code B1 General Principles 3 see also: William Magnuson, 'Takeover Regulation in the United States and Europe: An Institutional Approach' (2009) 21 PILR 205 at 221.

29 Staggered boards procedure prevents a bidder from gaining immediate control of the board of directors of the target company if the takeover is successful. It discourages a bidder from continuing with a takeover. See generally; Lucian A Bebchuk, John C Coates IV & Guhan Subramanian, 'The Powerful Antitakeover Force of Staggered Boards: Theory, Evidence and Policy' (2002) 54 SLR 887.

30 Dual class voting stocks frustrates takeovers. Two or more classes of shares may be created by a company. While one class (class A) can be publicly traded and carries a single vote per share, another class (class B) is not publicly traded and may carry as much as five or ten votes per share. Thus when there is a threat of a takeover, the holders of class B shares would have enough votes to defeat the takeover bid.

31 Shareholders' approval is required to issue new shares and dual-class voting stock is largely unsupported by institutional shareholders.

and company law,[32] they may not be effective to actually restrict pre-bid managerial-entrenchment practices since they operate to only regulate the relationship between company shareholders and management as principal and agents respectively.[33] The need to develop specific takeover regulations is indicative of the fact that the effectiveness of company law and corporate governance principles in regulating takeovers and restricting the role of management cannot be guaranteed.

Furthermore, since companies are required to disclose the identities of owners of shares which carry voting rights after a takeover bid has been made,[34] they are usually allowed a certain period of time after a bid has been made to collate and document such information. This has the side effect of granting extra time to company managements to explore more defensive means to subtly frustrate a takeover bid.[35] For example, the management can solicit 'white knights'.[36] Although this is not explicitly forbidden by the rules, the mere invitation by management may effectively mean that they are interfering with the takeover bid. Nevertheless, an invitation may be justified in view of the fact that their participation in takeovers operates to promote competition, which may enhance the value of the bid. In fact, when a bid has been announced and made public, other bidders may become interested in acquiring the target company as well. In either case, bid prices are likely to rise as competition amongst bidders intensifies. Thus any attempts to prevent or discourage managements from inviting 'white knights' to become involved in the bid process could

32 Staggered boards mechanisms are rendered ineffective in view of the fact that shareholders can remove directors at anytime. See Companies Act s 168.

33 Reinier Kraakman and John Armour et al, *The Anatomy of Corporate Law: A Comparative and Functional Approach* (OUP, 2009) at 247.

34 Apparently to prevent false markets, see The Takeover Directive, Article 8.

35 Kraakman and Armour (n 33) 236.

36 Han W Liu 'The Non Frustration Rule of the UK City Code on Takeovers and Mergers and Related Agency Problems: What are the Implications for the EC Takeover Directive?' (2010-2011) 17 CJEL, 5-10 at 9.

have an adverse effect on the competitive nature of a bid. Also, it would be difficult to establish whether a rival bidder has been invited by management or whether a rival bidder was genuinely introduced as a result of the bid announcement.

The emergence of takeover regulation has largely restricted the scope of the powers of managements during takeovers. Shareholders now have an active role particularly with reference to deciding whether to accept a bid on its merit. But since management may find a way to subvert these protective measures through their advisory roles and pre-bid defences, it appears that takeover regulations have not effectively restricted the managements of target companies from interfering with takeover bids. To provide effective protection to shareholders of target companies, a further restriction and supervision of the roles of managements during takeovers is required. However, further restrictions on the roles of managements during takeovers may interfere with their ability to perform their functions as agents of the shareholders. Thus it is important to provide the needed balance between preventing company managements from interfering with the interests of their shareholders during takeovers and allowing them to perform the functions that they owe to their company and the shareholders. In light of this, the shareholders may rely on the general duties of directors to make managements more accountable.[37] But whether company shareholders can actually rely on these duties during takeovers is unclear and this will be examined in section four.

### *B. Shareholder Protection in Acquiring Companies*

As stated above, the regulatory mechanisms on takeovers mainly focus on protecting the interests of the shareholders of target companies. However, during takeovers, the economic interests of the shareholders of

37 Companies Act, ss 171-177.

acquiring companies are also affected as the benefit that they derive from takeovers depends on the net economic value of their investment following the takeover.[38] Usually, whether the economic interests of shareholders are enhanced post-takeover depends on the underlying motive of the particular takeover. Since takeovers form part of investment decisions of corporate management, the decision to acquire another company should be based on the economic benefits that should accrue to the combined company post-takeover.

The inherent conflict of interests between directors and shareholders in acquiring companies means that the extent to which corporate acquisitions enhance shareholder value of the acquiring company remains contentious. As decisions to engage in takeovers are not specifically regulated, with only the broad duties on a board to operate the company with the shareholders' interests applying, a board can take advantage of their position at the expense of their shareholders. Firstly, a board may be driven by the desire to enlarge the size of the company that they control, to enhance the value of their positions and for the purpose of prestige among other reasons.[39] Also, they may pursue acquisitions as a pre-bid takeover defensive mechanism, undertaken by management fearing being acquired themselves, because takeovers which lead to the expansion of the acquiring company, make that company less likely to be a target of a takeover itself.[40]

Where acquisitions made by corporate managements are not driven by any of the above factors, it is expected that such acquisitions would invariably lead to gains for shareholders of acquiring companies. When this is not the case, it is difficult to determine the actual motive of management in their decisions to pursue acquisitions. When

38 Roll (n 5) at 201-03.

39 Michael Firth, 'Corporate Takeovers, Shareholder Returns and Executive Rewards' (1991) 12 MDE 421 at 425-27.

40 See generally: Gary Gorton Matthias Kahl, and Richard J Rosen, 'Eat or Be Eaten: A Theory of Mergers and Firm Size' (2009) 64 TJF, 1291.

an acquisition fails to confer benefits on shareholders, it may be caused by managerial overestimation of the synergistic gains that were expected from such acquisitions, leading to payment of higher bid premiums than necessary.[41] Although the hubris hypothesis does not suggest that management deliberately pay too much, the fact that management fail to diligently focus on realistic gains by being overconfident, may suggest that they deliberately made such acquisitions based on a motive which is different from the pursuit of shareholder value.[42] The effect of such acquisition is a transfer of wealth from the shareholders of the acquiring company to the shareholders of the target company. Where such decisions are not deliberate, they are likely to have been negligently made, especially as management suffers no loss in any event. Hence it was argued that managements that wish to maximise their private benefits bid for larger targets and pay premiums that are greater that the values of synergies. Whereas managements that are more concerned with enhancing shareholder value would seek smaller targets with which they can achieve corporate synergies.[43]

41 Managerial hubris. See generally Roll (n 5).

42 Ulrike Malmendier and Geoffrey Tate, 'Who Makes Acquisitions? CEO Overconfidence and the Market's Reaction' (2008) 89 JFE 20-43 at 36-42.

43 Paul Draper and Krishna Paudyal, 'Acquisitions: 'Private versus Public' (2006) 12 EFM 57 at 73. See also, Mara Faccio, John J McConnell, and David Stolin, 'Returns to Acquirers of Listed and Unlisted Targets' (2006) 41 JFQA 197.

Table III (b)

| | | High relative size ratio | | | Low relative size ratio | | | |
|---|---|---|---|---|---|---|---|---|
| | Holding period | Cash | Shares | All | Cash | Shares | All | dummy |
| All bidders | Pre-event | 0.8545 | 1.0707 | 0.7332 | 0.2850 | 1.5101 | 0.5077 | 0.3542 |
| | Around event | 0.3550 | 1.4077 | 0.4513 | 0.8704 | -0.5477 | 0.9376 | -0.4683 |
| | Post-event | 0.7165 | 1.5160 | 0.4106 | 0.4906 | 0.1729 | 0.3904 | -0.0225 |
| | Entire event | 1.9496 | 6.0820 | 1.7428 | 1.7394 | 1.4935 | 2.0543 | -0.0283 |
| | N | 2685 | 267 | 4271 | 1943 | 450 | 4271 | 8542 |
| Bidders for listed firms | Pre-event | 0.1175 | 0.6900 | 0.1582 | -1.1349 | 2.2520 | 0.2448 | 0.1179 |
| | Around event | -0.5408 | -0.5590 | -0.0691 | -0.0556 | -2.6681 | -0.5273 | 0.4065 |
| | Post-event | 0.3232 | -2.9590 | -0.2283 | -0.1679 | -0.2587 | 0.1441 | -0.4507 |
| | Entire event | 0.3198 | -2.0170 | 0.0425 | -1.3696 | -0.3940 | -0.0908 | 0.1509 |
| | N | 214 | 16 | 344 | 237 | 174 | 745 | 1089 |
| Bidders for private firms | Pre-event | 0.9068 | 1.0919 | 0.7542 | 0.4748 | 0.9609 | 0.5626 | 0.3332 |
| | Around event | 0.4346 | 1.4930 | 0.5011 | 0.9993 | 0.6835 | 1.2578 | -0.7261 |
| | Post-event | 0.7425 | 1.8410 | 0.4483 | 0.5883 | 0.3497 | 0.4504 | -0.0219 |
| | Entire event | 2.0956 | 6.6230 | 1.8594 | 2.1698 | 2.7430 | 2.5259 | -0.3180 |
| | N | 2471 | 251 | 3927 | 1706 | 276 | 3526 | 7453 |

Table III (b) [44] indicates that the acquirers of large targets (listed firms) fail to gain from the announcement of bids while the acquirers of privately held companies (small companies) benefit significantly. This is less likely to occur if management invested in their companies, as they would then have an incentive to take more care in making investment decisions. [45] Generally, managements have a duty to promote the interests of a company; and this is often indicated by its share price, payment of dividends and profits, meaning that managements are expected to make decisions that promote these objectives and other corporate values.

In summary, it appears that takeover regulation, designed to curb managerial excesses, has not completely achieved the desired objective of promoting shareholders' interests. This is particularly obvious when considering the shareholders of acquiring companies. Also, even though management may 'go round' takeover regulations, by deliberately soliciting 'white knights', they have an underlying duty to their companies, particularly to promote shareholder value. This makes directors duties crucial to ascertain the extent to which they may be responsible to shareholders during takeovers. Also, since takeovers directly affect the shareholders of a company, the extent to which they can hold directors to account for unproductive acquisitions through derivative actions is unclear. These two issues form the focus for the next section.

---

44 ibid 74-75.

45 Michael Firth, 'Takeovers, Shareholder Returns and the Theory of the Firm (1980) 94 QJA 235 at 255-58. See also, Yakov Amihud, Baruch Lev, and Nickolaos G Travlos, 'Corporate Control and the Choice of Investment Financing: The Case of Corporate Acquisitions' (1990) 45 TJF 603 at 611-15.

## IV. Directors duties and Derivative Action under the Companies Act

### *A. Directors Duties*

Following the codification of the duties of company directors, their responsibilities and scope of functions have been clearly defined. Although the duties as outlined under the Companies Act 2006[46] do not make reference to specific corporate matters, they nevertheless serve as a collective touchstone for determining whether company directors have used their powers for the purpose for which they exist. In view of the fact that company directors play active roles in making a takeover bid or in accepting or rejecting one, the duties of directors apply in relation to takeovers.[47] Whilst all the duties apply, the most relevant duties during a takeover are the duties to promote the success of the company, to exercise reasonable care, skill and diligence and the duty to avoid conflict of interests.

Firstly, directors are required to promote the interests of the company 'for the benefit of its members'.[48] Directors are meant to focus on enhancing the interests of the shareholders of their company through the company itself,[49] since the interests of a company as an artificial person, cannot be distinguished from the interests of the persons who are interested in it.[50] Although directors are expected to consider the interests of certain stakeholders, these stakeholders are to be considered only to the extent that the consideration of these interests enhances the interests of the members of the company.[51] This means that

46 Companies Act ss 171 -177.

47 The duties are stated to be the general duties of directors; hence they apply to all investment decisions including takeovers.

48 See Companies Act s 172 (1).

49 Paul L Davies (ed), *Gower and Davies Principles of Modern Company Law* (8th edn, Sweet & Maxwell, 2008) at 508.

50 *Brady v Brady* [1988] BCLC 20 at 40 Nourse LJ.

51 Companies Act s 172 (1).

the duties of directors towards the company are to be measured by reference to the extent to which the interests of members of the company are enhanced. This duty applies in relation to the interests of the shareholders of the target and acquiring companies during takeovers.

When a company becomes a target of a takeover, the directors of the company may accept the bid if it will enhance the value of their shareholders, or oppose the takeover bid if it appears to them that the takeover will not enhance the value of the shareholders of their company. With regards to acquiring companies, the directors should not make acquisitions unless such acquisitions are highly likely to enhance shareholder value of their company. However, it may be difficult to hold directors to account for investment decisions by reference to this duty, because the duty is subjective. Directors are required to act 'in the way that they consider' to be in good faith to promote the success of the company, this has been held to effectively leave business decisions to be made by directors.[52] Although, directors are required to act in good faith in performing this obligation, when they do not act in good faith, as 'an honest' business person would reasonably be expected to act, it is unlikely that they can be held liable for failure to act in good faith. Hence it was observed that the proof that the decisions of the directors of a company had caused substantial harm to the company was evidence against the contentions of the directors that they had acted in good faith rather than an absolute proof that the directors have not acted in good faith.[53] But the decisions of directors can be rejected when such decisions clearly show that they had failed to consider the interests of the company for the benefit of its members.[54] Also, the decisions of the director may be questioned by

52 *Re Smith & Fawcett Ltd* [1942] Ch 304 at 306, Lord Green MR See also, *Cobden Investments Ltd v RWM Langport Ltd* [2008] EWHC 2810 Ch.

53 *Regentcrest Plc (in liquidation) v Cohen* [2001] 2 BCLC 80 (Jonathan Parker J.)

54 *Extrasure Travel insurances Ltd v Scattergood* [2003] 1 BCLC 598.

considering whether an intelligent and honest man in a position of the director of the company could have reasonably believed that the transaction was for the benefit of the company.[55] While directors can actually be made to account for their investment decisions, it is difficult to prove that directors have actually breached their duty to act in the interest of the company, except in cases where they have left a clear record of their thought processes leading up to the challenged decisions.[56]

In light of this, it has been rightly contended that the duty of directors to promote the success of the company as contained in the Companies Act provides little or no guidance either to the directors of a company in making investment decisions or to the courts in reviewing the decisions.[57] Hence, it may be difficult for the shareholders of a company to rely on this duty to hold their directors to account for takeovers that did not promote the interests of their company.

Similarly, the duty to avoid a conflict of interests applies directly to target companies during takeovers. When a takeover bid is made, the decision of directors of target companies to accept or reject the bid should be based only on the extent to which their decisions will enhance the value of the shareholders of their company. They are not to unduly interfere with the bid and they are to ensure that their personal interest does not influence the outcome of the takeover bid.[58] The duty to avoid a conflict of interests is meant to ensure that persons who discharge fiduciary duties should not enter into business negotiations on behalf of a

---

55 *Charterbridge Corp Ltd v Lloyds Bank Ltd* [1970] Ch 62.

56 Davies (ed) (n 49) 510.

57 Andrew Keay 'The Duty to Promote the Success of a Company: Is it fit for Purpose?' in Joan Loughery (eds), *Directors Duties and Shareholder Litigation in the Wake of the Financial Crisis,* (Edward Elgar Publishing, 2013) at 85.

58 By reference to this duty, directors are required not to oppose a takeover bid to protect their positions in the company. See Tilton L Willcox, 'The Use and Abuse of Executive Powers in Warding off Corporate Raiders' (1988) 7 JBE 47.

company if they have or can have personal interests in the outcome of such negotiations.[59] Yet, as directors who act as company management cannot be excluded from their managerial roles during takeovers, their roles can only be restricted. This is what the current takeover regulations seek to achieve by providing that shareholders should be allowed to decide on the merit of a bid without the interference of management except in relation to their advisory role.[60] Since the purpose of the duty is to exclude directors from acting on behalf of their company when there is a possibility of conflict of interests, it is not exactly clear whether the shareholders of a company can hold directors to account for managerial decision during takeovers by relying on this duty. This is because the managerial role of directors during takeovers has only been restricted, and when the directors use their advisory role to gain personal benefits during takeovers, it may be difficult to prove that they promoted their personal interests over the interests of the company even when this is clearly the case. This duty may have had a clear sense of application if directors' managerial responsibilities were excluded when their companies were subject to a takeover attempt but, unfortunately, this is not the case. The exclusion of the roles of directors during takeovers may be thought to stifle entrepreneurial activities of companies, but as was rightly contended, this argument ignores the fact that the duty merely prevents a director in a situation of conflict of interest from exploiting that situation.[61]

Also, the duty of directors to exercise reasonable care, skill and diligence applies to takeovers. Particularly, this duty applies to the directors of acquiring companies. One of the functions of takeovers is to enhance the corporate

---

59 See *Aberdeen Railway Co v Blaikie Bros* [1854] 1 Macq 461 at 471. During takeovers, directors are interested in the outcome of a takeovers bid, because, they may be dismissed post-takeover.

60 The Takeover Directive, Art 9 (5).

61 See similar argument in: Brenda Hannigan, *Company Law* (3rd edn, OUP, 2012) at 241.

and economic value of a company. This means that takeovers should be aimed at enlarging not only the corporate size of an acquiring company, but also its economic value.

Meanwhile, it is difficult to establish the motives of the management of an acquiring company when making acquisitions, but the fact that acquisitions put managements in a better position than they were in before the acquisitions – even when such acquisitions may not lead to significant gains –[62] raises the presumption that they may have cared less whether or not the economic values of a company is enhanced by an acquisition. Usually, the expansion of the company leads to an increase in the responsibilities of the company management; this effectively leads to an increase in their allowances and benefits. Although, when acquisitions lead to losses or zero gains for an acquiring company, it may not be categorically contended that management deliberately ignored the high possibility of losses from such acquisitions but it can be argued that such losses may have been caused by lack of care and diligence. This is because such overpayments may have been averted had the management exercised restraint in deploying their skill and managerial expertise when pursuing acquisitions. In light of this, it is important that shareholders of acquiring companies are protected from managerial hubris[63] during takeovers. Since the current takeover regulations do not contain provisions that are meant to protect the shareholders of the acquiring companies, it appears that regard may be had to shareholder remedies through derivative actions.

---

62 Elazar Berkovitch, M P Narayanan, 'Motives for Takeovers: An Empirical Investigation' (1993) 28 Journal of Finance and Quantitative Analysis, 347 at 352

63 See generally: Cathy M Niden, 'An Empirical Examination of White Knight Corporate Takeovers: Synergy and Overbidding' (1993) 2 Financial Management 28. M Raj & M Forsyth, 'Hubris Amongst UK Bidders and Loses to Shareholders' (2003) 8 International Journal of Business 2.

### *B. Derivative Actions*

A derivative action[64] offers company shareholders the opportunity to commence legal proceedings against directors when they breach their duties that are owed to the company. Although directors' duties as provided in the Companies Act are to be owed to their companies, the capacity of the company shareholders to commence proceedings against directors in breach of their duties shows that the shareholders are the beneficiaries of the duties that directors owe to the company. However, in spite of the derivative action procedure, shareholders of an acquiring company face a difficult task in persuading the courts to rule in their favour. This is because the losses that are suffered by shareholders of an acquiring company after a takeover has been concluded arise as a result of the negligent conduct of the directors in the fulfilment of their duty towards the company. As such, it has been held in *Prudential Assurance Co. Ltd v Newman Industries Ltd* (No. 2)[65] that when a company suffers loss as a result of breach of duty by the directors, the loss to shareholders through a reduction in the value of their shares, or loss of dividend, merely reflected the loss suffered by the company, and shareholders cannot recover damages.[66] Respectfully, the decisions of the court with regards to reflective loss suffered by shareholders are unclear, especially when applied to takeovers. During takeovers, gains or losses made by companies are actually measured with reference to the value of shares that are held by the shareholders. It may be thought that when making acquisitions, the directors seek to enhance the economic value of the shareholders of their companies through corporate expansion leading to synergies, so that gains or losses suffered by the company are to be measured by reference to the extent that share prices increase or

64 As provided under the Companies Act, ss 260-269.

65 [1982] 1 Ch 204.

66 See also, *Stein v Blake (No 2)* [1998] 1 BCLC 573.

diminishes post-takeover. In light of this, an attempt was made to distinguish the loss that is suffered directly by a shareholder from the loss that is suffered by the company. It was stated that the rule that a shareholder cannot bring an action when they suffer loss as a result of losses that had also been suffered by the company has nothing to do with a shareholders' right of action for a direct loss caused to his own pocket as distinct from a loss caused to the value of a company in which he holds shares.[67] In furtherance of this, the principle set down in *Prudential Assurance* was reviewed in *Johnson v Gore Wood & Co* thus:[68]

> Where a company suffers loss caused by a breach of duty to it, and a shareholder suffers a loss separate and distinct from that suffered by the company caused by a breach of a duty independently owed to the shareholder, each may sue to recover the loss caused to it by breach of the duty owed to it, but neither may recover loss caused to the other by breech of the duty owed to that other.

While this may effectively seek to grant the shareholders of an acquiring company the right to commence actions against directors, they must first show that the action is brought in their own right. As indicated in *Johnson v Gore Wood,* the shareholders must show that the loss that they have suffered arises from a breach of the directors' duties, which is owed to them. This decision deviates from the earlier decision in *Prudential Assurance* and it represents an improvement in the protection of the interests of company shareholders especially from the negligent conduct of directors in making acquisitions. But since directors owe the general duties to the company and not to the shareholders particularly,[69] it may be difficult to establish that certain

---

67 *Heron International Ltd v Lord Grade* [1983] BCLC 244 at 262 (Lawton LJ).

68 [2002] 2 AC 1 Lord Bingham.

69 Companies Act s 170 (1).

duties of directors are owed to the shareholders of a company during takeovers. However, it is arguable whether directors can be said to be acting on behalf of shareholders during takeovers in view of the fact that gains or losses are measured by reference to share prices post-takeovers. Likely, this may be determined by reference to the conduct of the directors and the circumstances of each particular case.

Clearly, the emergence of takeover regulations became necessary since takeovers during the early periods were characterised by the conflict of interests between company management and their shareholders despite the common law duty of care, which was owed to the company by directors. But it has been revealed here that the management of a target company may still undermine the current takeover regulations to promote their personal interests, and the shareholders of the acquiring companies remain unprotected in the absence of any specific regulation to restrain the management of an acquiring company from making negligent acquisitions.

## V. Conclusion

Shareholder protection during takeovers has been an issue of much debate. Although, the duties of directors appear to be applicable during takeovers, the ability of directors to undermine these duties when they make specific investment decisions resulted in the establishment of a specific regulatory framework for takeovers. Despite the fact that the objective to protect shareholders from managerial excesses during takeovers influenced the emergence of takeover regulation, the extent to which shareholders are protected during takeovers appears less satisfactory. While current takeover regulation largely focuses on shareholder protection, it emerged that corporate managements still have the capacity to promote their personal interests during takeovers. This was showed to be particularly possible through pre-bid defence strategies that may be used by managements in pursuit of their objectives. Also, it was revealed that the scope of protection that the regulatory

framework of takeovers offers to shareholders is actually limited to the shareholders of target companies. Shareholders of acquiring companies are also affected by the outcome of takeovers. This was examined by reference to the hubris hypothesis of takeovers which occurs as a result of overpayments which is made by management of the acquiring company to the target company in pursuit of a takeover. This in itself represents a transfer of wealth from the shareholders of the acquiring companies to shareholders of the target companies. Irrespective of whether the overpayment made by the management of acquiring companies was motivated by empire building or synergistic gains, management would be expected to be more careful if they have an incentive to engage in careful and scrupulous acquisitions. If they have such an incentive, they would exercise restraints in making acquisitions and reduce the possibility of a loss. Although they may be merely performing their obligations as company management they would also be less likely to make hasty acquisitions where there is a mechanism to regulate their conduct and hold them accountable for careless and avoidable losses which occurs as a result of acquisitions. With respect to this, derivative actions by the shareholders of acquiring companies were revealed to be possible only to the extent that they can show that the directors of their company made acquisitions to enhance the value of their shares specifically and the size of the company generally.

Whether management of acquiring companies should be regulated by direct legal means or supervised by legal institutions depends on the extent to which the interests of the shareholders can be protected while managements can still perform their functions effectively. With respect to target companies, shareholder protection during takeovers can only be reasonably guaranteed by strengthening the current regulatory framework that was created for this purpose.

## BIBLIOGRAPHY


### Cases

*Aberdeen Railway Co v Blaikie Bros* [1854] 1 Macq 461
*Brady v Brady* [1988] BCLC 20
*Charterbridge Corp Ltd v Lloyds Bank Ltd* [1970] Ch 62
*Cobden Investments Ltd v RWM Langport Ltd* [2008] EWHC 2810 Ch
*Extrasure Travel insurances Ltd v Scattergood* [2003] 1 BCLC 598
*Heron International Ltd v Lord Grade* [1983] BCLC 244
*Johnson v Gore Wood & Co* [2002] 2 AC 1
*Prudential Assurance Co Ltd v Newman Industries Ltd (No 2)* [1982] 1 Ch, 204 CA
*Regentcrest Plc (in liquidation) v Cohen* [2001] 2 BCLC 80
*Smith & Fawcett Ltd Re* [1942] Ch 304
*Stein v Blake (No 2)* [1998] 1 BCLC 573 CA

### Statutory Material

Companies Act 2006
Council and European Parliament Directive 2004/25/EC of 21 April 2004 on Takeover Bids [2004] OJ L 142/12
The UK City Code on Takeovers and Mergers 2013 (20th May)

### Books

Brealey R A, Myers S C *et al, Principles of Corporate Finance* (McGraw-Hill/Irwin New York, 2008)
Davies P L (ed), Gower and Davies Principles of Modern Company Law (8th edn, Sweet & Maxwell London 2008)
Dignam A, *Hicks & Goo's Cases & Materials on Company Law* (7th edn, OUP 2011)
Girvin S, Frisby S & Hudson A, *Charlesworth's Company Law* (18th edn Sweet & Maxwell, 2010)
Hannigan B, *Company Law* (3rd edn, OUP, 2012)
Keay A, The Duty to Promote the Success of a Company: Is it fit for Purpose? in Joan Loughery (ed), *Directors Duties*

*and Shareholder Litigation in the Wake of the Financial Crisis* (Edward Elgar Publishing, 2013)

Kraakman R, Armour J *et al, The Anatomy of Corporate Law: A Comparative and Functional Approach* (OUP, 2009)

Moon R W, *Business Mergers and Takeover Bids: A Study of the Postwar Patterns of Amalgamatons and Reconstruction of Companies* (Gee & Co, 1976)

Rubach R S An Overview of Takeover Defences in Alan J Auerbach (ed) *Mergers and Acquisitions* (University of Chicago Press, 1987)

Sealy L & Worthington S, *Cases & Materials in Company Law* (9th edn, OUP, 2010)

## Academic Articles

Amihud Y, Lev B & Travlos N G, 'Corporate Control and the Choice of Investment Financing: The Case of Corporate Acquisitions' (1990) 45 The Journal of Finance 603

Armour J and Skeel D A Jr, 'Who Writes the Rules for Hostile Takeovers and Why? - The Peculiar Divergence of US and UK Takeover Regulation (2006-2007) 95 The George Town Journal, 1727

Bates T W and Becher D A, 'Bid Resistance by Takeover Targets: Managerial Bargaining or Bad Faith?' (2011) Chicago Meetings Paper, AFA/2012 1-49 <http://dx.doi.org/10.2139/ssrn.1786674>

Bebchuk L A, Coates J C IV & Subramanian G, 'The Powerful Antitakeover Force of Staggered Boards: Theory, Evidence and Policy' (2002) 54 Stanford Law Review 887

Berkovitch E & Narayanan M P, 'Motives for Takeovers: An Empirical Investigation' (1993) 28 Journal of Finance and Quantitative Analysis, 347

Clark B J, 'Takeover Regulation - Through the Regulatory Looking Glass' (2007) 8 German Law Journal 381

Clark B J, 'The Takeover Directive: Is a little Regulation Better than No Regulation?' (2009) 15 European Law Journal 174

Cotter J F & Zenner M, 'How Managerial Wealth Affects the Tender Offer Process' (1994) 35 Journal of Financial Economics 63

Cotter J F, Shivdasani A & Zenner M, 'Do Independent Directors Enhance Target Shareholder Wealth during Tender Offers? (1997) 39 Journal of Financial Economics 3

Davies P, Schuster E P, et al, 'The Takeover Directive as a Protectionist Tool' (2010) ECCG Working Paper Series in Law 141/2010 1 <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1554616> accessed 10 April 2013

Draper P & Paudyal K, 'Acquisitions Private Versus Public' (2006) 12 European Financial Management, 57

Eddey P H & Casey R S 'Directors Recommendations in Response to Takeover Bids: Do they Act in their Own Interests?' (1989) 14 Australian Journal of Management 1-28

Darren H, 'Directors' Recommendations in Takeovers: An Agency and Governance Analysis' (2005) 32 Journal of Business Finance & Accounting 129-159

Faccio M, McConnell JJ, & Stolin D, 'Returns to Acquirers of Listed and Unlisted Targets' (2006) 41 Journal of Financial and Quantitative Analysis 197

Firth M, 'Takeovers Shareholder Returns and the Theory of the Firm (1980) 94 The Quarterly Journal of Economics 235

Firth M, 'Corporate Takeovers, Shareholder Returns and Executive Rewards' (1991) 12 Managerial and Decision Economics 421

Goergen M, Martynova M, & Renneboog L, 'Corporate Governance Convergence: Evidence from Takeover Regulation Reforms in Europe (2005) 21 Oxford Review of Economics Policy 243

Gorton G, Kahl M & Rosen R J, 'Eat or Be Eaten: A Theory of Mergers and Firm Size' (2009) 64 The Journal of Finance 1291

Hodgkinson L & Partington G L, 'The Motives for Takeovers in the UK' (2008) 35 Journal of Business Finance & Accounting 102

Holl P & Demetris K, 'The Determinants of Outcome in UK Takeover Bids' (1996) 3 International Journal of Economics and Business 165

Jensen M C & Meckling W H, 'Theory of the Firm: Managerial Behaviour, Agency Cost and Ownership Structure' (1976) 3 Journal of Financial Economics 305

Liu H W, 'The Non Frustration Rule of the UK City Code on Takeovers and Mergers and Related Agency Problems: What are the Implications for the EC Takeover Directive?' (2010-2011) 17 The Columbia Journal of European Law 5

Magnuson W, 'Takeover Regulation in the United States and Europe: An Institutional Approach' (2009) 21 Pace International Law Review 205

Malmendier U & Tate G, 'Who Makes Acquisitions? CEO Overconfidence and the Market's Reaction (2008) 89 Journal of Financial Economics 20

Manne H G, 'Mergers and the Markets for Corporate Control' (1965) 73 Journal of Political Economics 110

Niden CM, 'An Empirical Examination of White Knights Corporate Takeovers: Synergy and Overbidding' (1993) 2 Financial Management 28

Raj M & Forsyth M, 'Hubris Amongst UK Bidders and Loses to Shareholders' (2003) 8 International Journal of Business 2

Roll R, 'The Hubris Hypothesis of Takeovers (1986) 59 Journal of Business Finance 197

Subramanian G, 'Takeover Defences and Bargaining Power' (2005) 17 Journal of Applied Corporate Finance 85

O'Sullivan N & Wong P, 'Board Composition, Ownership Structure and Hostile Takeovers: Some UK Evidence' (1999) 29 Accounting and Business Research 139

Varaiya N P, The 'Winner's Curse' Hypothesis and Corporate Takeovers (1988) 9 Managerial and Decision Economics 209

Walking R, 'Predicting Tender Offer Success: A Logistic Analysis' (1985) 20 Journal of Financial and Quantitative Analysis 461

Willcox T L, 'The Use and Abuse of Executive Powers in Warding off Corporate Raiders (1988) 7 Journal of Business Ethics 47

## Newspaper Articles

*The Economist*, 'Battle for the Savoy' (London, 12 December 1953)

*The Times,* 'Dividend Raised to Counter Bid' (London, 20 December 1958)

# What has the state got to do with healthcare?

Malcolm Oswald

## Abstract

*How should healthcare resources be allocated? Who should pay for it? What is the role of the state? There is little sign of agreement on these questions because differences are fundamental and often inter-disciplinary. Some writers, typically philosophers and ethicists, begin with a human right to health or healthcare, whilst some pursue equality of capability or procedural justice. Economists tend to look to maximise health yield from scarce resources. These analyses often rely heavily on state involvement and state funding. Many libertarians would reject these claims and seek to mininise the involvement of the state. They would argue that, so far as possible, individuals should be responsible for choosing and paying for the healthcare cover that they want.*

*In this article I draw on thinking from several academic disciplines including: political philosophy, law, bioethics, economics and psychology, in order to consider what the minimum involvement of the state should be, from the perspective of an ethical libertarian seeking to mininise state involvement and maximise individual autonomy and responsibility. In the story that follows, set in a fictitious democracy, I argue that even for an ethical libertarian there is much for the state to do including:*

- *Funding basic healthcare and many public health activities;*
- *Subsidising (or making the market cross-subsidise) insurance cover for more-than-basic-healthcare for certain people who would otherwise, through no fault of their own, have high-cost insurance premiums; and*
- *Creating law and policy on how decisions are made about health care entitlements, how procedural justice is provided, and devising a regulatory framework, governing providers of healthcare products and services.*

## I. Beginning

"Great news, John. I have organised a big campaign speech entitled: 'what has the state got to do with healthcare?' The TV crews and national newspapers will all be there."

"Good work, Barney." The candidate paused. "And what is my policy on healthcare, Barney?"

"Our usual message John. It's 'let's get government out of healthcare, because we know what's best for ourselves and our families'. We want consumer-driven healthcare – we each buy the health insurance we want.[1] It's a fashionable message, in keeping with our other policies, and our supporters will love it. You could add in a bit of nudge policy[2] to encourage people to do the sensible thing – but be careful we are not accused of telling people that government knows best."

The candidate was smart enough to know that the market for healthcare was not as straightforward as the market for soap. "OK let me talk to a few people."

But Barney knew what he was thinking. "Not those academics again John. If you must talk to them, I'm coming along".

## II. Middle

John welcomed them as they arrived: the economist, the political philosopher, the clinician, the bioethicist, and the historian. It was one thing to persuade those who already distrusted the state, and quite another to persuade other more thoughtful and sceptical voters. These academics made him think more deeply about difficult policy questions. He also consulted them because he cared about doing the right thing.

***Candidate:*** Ladies, thank you for coming to meet me. We are a refreshingly pluralist democracy, full of people

---

1 Regina E Herzlinger, *Consumer-driven health care: Implications for providers, payers, and policy-makers* (Jossey-Bass 2004).

2 Cass R Sunstein and Richard H Thaler, 'Libertarian Paternalism Is Not an Oxymoron' 70 U Chicago L Rev 1159.

who express freely their differing opinions. But many people share my beliefs that government is much too big and that each of us is responsible for choosing and finding our own way in the world, and making a success of our own life.[3] What counts as a good life for me might not be a good life for you. That adds to the richness and diversity of our society. We must respect others, let them make their choices – good and bad - and live with the results. To interfere with those freedoms more than we must is wrong; it is unethical. I recognise that we do not all have the same opportunities, and we may be able to do something to level the playing field, especially when people are young, but we cannot legislate away good and bad luck. In general, we should step back and respect the autonomy of individuals.

I recognise that others have alternative ethical convictions. They talk of rights to healthcare, and of equality of one thing or another. We might disagree, but I must listen to their arguments. I am a politician who looks to govern a pluralist democratic state, so I must look to govern those who agree with me and those who do not. I must lay out my thinking and let people judge me on my values and my policies. If they vote for me they must know what they are getting.

Few things matter more to people than their own health. So when it comes to healthcare I want people to make their own choices, according to their own priorities, and for the state to interfere as little as possible. But what is as little as possible? When it comes to healthcare, what are the minimum responsibilities of a state? I know that some libertarians see no role for the state in healthcare[4], but I am open to persuasion. Let us leave what we can to the market, but where are we morally bound to intervene? I have brought you here today to ask you these questions.

---

3 Ronald Dworkin, *Is democracy possible here?: principles for a new political debate* (Princeton University Press 2006) 17.

4 Robert Nozick, *Anarchy, state, and utopia* (Basic Books 1974) 297.

So let us begin from the position that the state has got nothing to do with healthcare, and identify the minimum that it must do in any decent democratic society. What are the general responsibilities of the state?

***Political philosopher***: Well, of course there is much disagreement amongst scholars, especially about the characteristics and responsibilities of an ideal democracy,[5] [6] [7] but almost all political philosophers and political scientists would agree that a government in a working democracy[8] has a responsibility amongst other things to:

- Protect the safety of the people, an idea dating back to Thomas Hobbes;[9]
- Devise laws to clarify what is right and wrong, and interpret and apply that law, resolving disputes in the courts, as argued by John Locke;[10]
- Show equal concern for the lives of everyone, even though it is inevitable that laws and policies will affect different people differently;[11]
- Listen to, respect, and be responsive to, the preferences of citizens,[12] although most would agree that that this does not mean that politicians must always follow the will of the majority.

***Candidate***: OK, but let's not forget John Stuart Mill who said that the only justification for the state interfering

---

5 Robert A Dahl, *On democracy* (Yale Univ Pr 2000).

6 Dworkin (n 3).

7 Benjamin R Barber, *Strong democracy: participatory politics for a new age* (University of California Press 1984).

8 Robert A Dahl, *A preface to democratic theory* (University of Chicago Press 1971) 63.

9 Thomas Hobbes, *Leviathan* (Second edn, Cambridge University Press 1996).

10 John Locke, *The second treatise of government: and, A letter concerning toleration* (Dover Pubns 2002) 57.

11Dworkin 144 (n 3).

12 Robert A Dahl, *Polyarchy: participation and opposition*, vol 54 (Yale Univ Pr 1971) 1.

with our liberty is to prevent harm to others.[13] Do the responsibilities you mention imply that the State has to get involved in the healthcare of individuals?

***Political philosopher***: When we think about the state's role of protecting safety, we tend to think about national defence and perhaps the police service. Yet many of us today face a greater threat from viruses, diseases and accidents.[14] Often such threats have been deadly and come from foreign shores.[15] How can the state protect our safety without addressing these threats?

***Candidate***: I agree we all want to be protected from danger. But I want people to take responsibility for their *own* health and their *own* lives. It is important that they insure themselves against threats to their health. Markets work - this we know. Let us leave the state out of this and let the market insure citizens against these threats.

***Economist***: Ideally, the best way to run the economy is to let individuals work, play, and consume what they want without restrictions. The interaction of supply and demand in the market naturally leads to equilibrium in which marginal benefits equal marginal costs. The prices that arise from the exchange in the market direct individuals to work in jobs where their skills provide the most value to society, to find efficient means of production, to limit the consumption of goods that are most scarce, and to save and invest for the future. Under ideal conditions, the entire economy functions without any central control or direction from the government. However, perfect market conditions...do not occur in the real world. Imperfect market conditions justify government intervention to protect the public's health. A "public good" is a good or service that does not lend itself to

---

13 John S Mill, 'On Liberty' in S Collini (ed), *On Liberty and Other Essays* (Cambridge University Press 1989).

14 C A Erin and J Harris, 'AIDS: ethics, justice, and social policy' 10 Journal of applied philosophy 165, 166.

15 Lincoln C Chen, Tim G Evans and Richard A Cash, 'Health as a global public good' 1 Global Public Goods 284.

market allocation because it costs nothing for an additional individual to enjoy its benefits, and it is generally difficult or impossible to exclude individuals from consuming it. The institutional and technical capacity to respond to disease outbreaks and prevention research are examples of public goods. A fundamental problem with public goods is the difficulty of motivating people to pay for them.[16]

***Political philosopher***: That suggests that in order to protect the safety of the people, the government has to intervene and pay for "public goods" like preventing and controlling epidemics. Otherwise viruses and diseases will develop and spread. These activities cannot be left to the market.

***Barney***: Why? Let us make it a criminal offence to fail to buy healthcare insurance to pay for this protection.

***Political philosopher***: That would hardly signal individual autonomy and small government. It would be the state coercing the individual to pay for something the state wants the citizen to have.

***Barney***: The state need not fund everyone. Those who can afford it can pay for themselves.

***Economist***: Means testing will mean that some will buy cover and some will not. Cover will not be universal. Public health works by protecting whole populations. Economists call some public health activities "public goods", and some like vaccination against infectious disease, we call "merit goods" where there are "externalities" - benefits or costs to others from our economic choices. When I am vaccinated you benefit from my protection against disease. Externalities prevent markets from working efficiently where consumers or producers are not compensated for these effects. They can apply to individual healthcare and public health. For example, choosing to see the doctor when I am ill is likely to have a positive impact on other people, such as

16 Vilma G Carande-Kulis, Thomas E Getzen and Stephen B Thacker, 'Public goods and externalities: a research agenda for public health economics' 13 Journal of Public Health Management and Practice 227, 227.

the people I meet (who otherwise might become infected), my employer who needs me at work, and the economy as a whole. The knock-on benefits of my doctor's visit are not reflected in market prices, and as a result some people will be deterred from visiting the doctor even though the overall benefits justify a visit.

***Candidate***: I accept that the state has to fund, or at least subsidise, public health activities like vaccination where the population has to be protected so that each individual is protected. Advocates of consumer-driven healthcare like John Goodman also accept your arguments about externalities:

> We don't want a parent to choose not to have her child vaccinated, or an at-risk expectant mother to avoid prenatal care, or a heart patient to eschew aspirin or beta-blockers. The reason: there is overwhelming evidence that the social benefits of the care exceed the social cost. Yet instances where we can be absolutely sure that we know which alternative is the right choice are rarer than one might suppose. At the other extreme, there are literally thousands of cases where only the patient can make the right choice.[17]

He goes on to argue that whether to spend an extra $800 on a brand-name drug is a decision that can only be made by an individual. Drugs affect people differently, and people have different attitudes toward risk. Only when individuals spend their own money will they reveal their preferences. Therefore, one person cannot make an informed choice for another.

***Lawyer***: That of course ignores children and adults lacking the capacity to make decisions for themselves. The state must make law to say who can make decisions on their

17 John Goodman, 'Consumer-directed health care' (SSRN, 2006) 4 <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=985572#PaperDownload> accessed 18 March 2012

behalf to protect their interests. Also, when two identical patients are offered two different levels of care, one of superior quality to the other, this could raise some difficult legal questions. For instance, is it acceptable to offer "sub-standard" care to one of the patients? Could that be construed as negligence?[18]

***Clinician***: Furthermore, the patient will not know in advance how the drug will affect her. And how well will she know her preferences for chemotherapy treatment if she has never experienced it before? When I buy many goods, like eggs for instance, I know my preferences but that does not hold true for much healthcare. Thus the individual patient is not especially well placed to judge either risk or her own preferences.

***Economist***: Research does suggest that human beings tend to be poor at making decisions that involve the assessment of risk,[19] and indeed at making rational choices in general. Our choices are shaped by how problems are framed.

***Candidate***: Whatever the evidence of our failings as rational actors, you are not going to convince me that anyone other than me is best placed to make the important choices that affect my health and life. However, I accept that the State has a role to play where there are significant wider social benefits from healthcare. But other than that, the consumer buys insurance to protect her own health. Agreed?

***Economist***: There are difficulties with insurance. Economists know that many people would rather consume today than insure for tomorrow. Younger people, especially those on lower wages, are likely to allocate an insufficient

---

18 MA Hall, 'Paying for What You Get and Getting What You Pay for: Legal Responses to Consumer-Driven Health Care' 68 Law & Contemp Probs 159, 176.

19 Amos Tversky and Daniel Kahneman, 'The framing of decisions and the psychology of choice' 211 Science 453.

portion of their wages to future healthcare.[20] So they are likely to under-insure. That preference also leads people to consume too much food today, despite the negative impact that this will have on their future health and life expectancy.

***Barney***: This is where individuals have to take responsibility for their actions. We have options, we make choices, and we must live with the consequences. The state is not there to bail us out for our dumb choices.

***Lawyer***: But we can at least encourage people to avoid bad choices so that they won't need bailing out. Responsible governments around the world make laws and policies that create incentives for good behaviours, and disincentives for unhealthy or unsafe behaviours... like smoking.

***Bioethicist***: Yes, the state can be seen as a steward, with a responsibility to guide people towards good choices, and to reduce health inequalities.[21]

***Barney***: The state is not a shepherd guiding us through life!

***Candidate***: I accept there is a role even for a libertarian government to "nudge" people towards sensible choices,[22] but not to tell people what to do.

***Bioethicist***: An ethical government has to do more than nudge people. It has to be there to bail out some people even when they have made bad choices.

***Barney***: Oh save me from bleeding heart liberals. The state is not a big cash cow to be milked dry by people who make dumb choices and get themselves in a fix.

***Bioethicist***: We know there will be people who will not buy healthcare insurance. Imagine Al... he is an alcoholic, homeless, and with very little money. Drunk one

---

20 Mark A Hall, *Reforming private health insurance* (American Enterprise Institute 1994) 66.

21 Nuffield Council on Bioethics, 'Public Health: ethical issues' (2007) 18 <http://www.nuffieldbioethics.org/public-health>.

22 Cass R Sunstein and Richard H Thaler, 'Libertarian paternalism' 93 American Economic Rev 175.

night he crosses the road in front of a hospital and is hit by a passing car. He requires basic medical care to clean his wounds and stem the profuse bleeding from his leg. With no money, and no insurance, should he be left to die?

***Barney***: Al should have stayed off the booze and off the streets. No one forced him to become an alcoholic. He must live or die with the consequences of his decisions.

***Bioethicist***: How can we know it was Al's fault that he became an alcoholic? Maybe he had a tendency in his genes. Maybe he was abused as a child and ran away from home. Maybe he failed to get a job after years of trying. Maybe he became depressed because his wife left him and took the kids. How are we in practice to sort out if Al is to blame? And even if he is to blame, are we all to stand by and watch him die? Could we ever call that ethical behaviour? Many would argue there is a moral rule of rescue[23] that means we cannot walk on by.

***Political philosopher***: Whether or not we accept the moral rule of rescue, we are compassionate beings. As Jean-Jacques Rousseau said, "it is this compassion that hurries us without reflection to the relief of those who are in distress" [Discourse on the Origin of Inequality, 76]. He maintained we are naturally sympathetic to others, and are upset by their suffering. Mencius, an early scholar from the Confucian tradition, argued that humans find suffering in others unbearable, are naturally benevolent, and that benevolence is the strongest motive to moral action.[24]

***Barney***: So what? Why should we listen to some long-dead Frenchman and a prehistoric Chinese guy?

***Clinician***: Because modern science has proven Mencius and Rousseau to be right. Brain research tells us that when a human detects pain in another person, it triggers a response in the observer's brain in the same area of brain circuitry as that of the sufferer – a "compassionate" response.

---

23 John McKie and Jeffrey Richardson, 'The rule of rescue' 56 Social Science & Medicine 2407.

24 Din C Lau, *Mencius* (Chinese Univ Pr 2003) xviii.

Not only does an observer's brain "mirror" activity in that of the pain sufferer, but his or her empathy varies directly with pain intensity.[25]

***Candidate***: I accept that people are compassionate. But why not leave it to individual compassion? Let each of us choose to give to charities that can help Al.

***Barney***: Yeah, leave the state out of it.

***Political philosopher***: We have said already we cannot reliably judge who is to blame for a person's ill health. If we left it to charity the state would be failing to protect not only Al's safety, but also the safety of children who suffer harm and disease through no fault of their own, and the disabled or genetically unfortunate who are burdened with chronic ill-health, disability or loss of life. Like Al, they too may have great need but may have little money. The duty to protect safety cannot be abdicated and left to individual philanthropy. Why should I abide by the coercive laws of the state when it does not protect me? Furthermore, the scale of the philanthropy you envisage would be considerable. Individual autonomy comes with responsibility, and that should not be shirked. The compassionate and generous should not have to pay for Al because the selfish would like to see him saved but would prefer to keep their money for themselves. Each must pay their fair share to the state, so the state can be fair.

***Candidate***: Remember that the state has already stepped in to protect Al's safety with traffic laws, speeding fines, road signs and so on. There are limits to the state's responsibilities. Nevertheless, I accept the state should fund these catastrophic cases. I am not persuaded that it is our moral duty, but I am persuaded that my voters are compassionate. But state funding should cover the very minimum necessary to prevent serious harm and protect human life, and only for those cases where basic care brings great benefit. In these cases our compassion is strong. A

25 Miiamaaria V Saarela and others, 'The compassionate brain: humans detect intensity of pain from another's face' 17 Cerebral Cortex 230.

clear line needs to be defined and drawn - the state can afford basic care for Al, but coercive state taxes should not be imposed in order to pay for Al to have expensive cancer drugs. Beyond basic care it is for each of us to decide how we spend our money. We might choose to spend less on houses and hobbies, so that we can spend more on healthcare. No one else can make those trade-offs for us. It would be wrong for the state to prevent some of us from choosing better healthcare.[26] Each of us chooses and pays for our own healthcare insurance cover.

***Bioethicist***: That sounds fair on libertarian grounds but pause a moment. You believe in equal opportunities for all - let each of us be given the chances and then make our own luck. So then what do you say to those who are dealt the poor cards: born to a deprived family with a poor diet, or with damaged genes, or with a chronic illness? That makes them unlucky enough to expect poor health, and their ability to earn may be diminished. Are we to add to that by making them pay double or triple the health insurance premiums of the rest of us? We said earlier that when making law and policy, the state has a responsibility to show equal concern for everyone. Equal concern must mean that the state makes the healthy subsidise the unhealthy.

***Barney***: The insurance companies can look after that if they want. It's not for the state to interfere.

***Economist***: Unless the state intervenes, the market will charge according to risk. So the unhealthy will pay handsomely. If an individual firm offered to cross-subsidise as you suggest a rational healthy consumer will simply move to another company with cheaper premiums. Furthermore, if obliged to cross-subsidise by law and fix prices in favour of one or more groups of consumers, the market will not function efficiently, because the market would not be setting prices according to cost.

---

26 H Tristram Engelhardt, 'Health care reform: A study in moral malfeasance' 19 J Medicine and Philosophy 501.

***Historian***: These cross-subsidies may not be efficient economically, but they have been a common way that healthcare has been funded in the past. They have occurred not only because different people are blessed or burdened with different health characteristics, but because not everyone has the same ability to pay:

> Under ancient Roman law and in Renaissance England, physicians, like barristers, were legally precluded from enforcing ordinary contracts for their fees because this was seen as inconsistent with their status as noble, learned professionals. Instead, physicians and barristers received voluntary honoraria and were expected to serve patients regardless of their ability to pay.[27]

***Barney***: But this is not ancient Rome!

***Candidate***: We have already said that the state must provide funding to ensure that basic healthcare is accessible to all to protect their safety, as long as it is not too costly. I do not accept that the state has also to be concerned about ability to pay for insurance for healthcare that goes beyond that basic minimum. However, I do accept that there is an argument for subsidies for "more-than-basic" healthcare for those who inherit or are afflicted by serious health problems. I envisage two conditions... firstly, it must be absolutely clear that ill health is through no fault of their own. If they in any way caused their own ill health, for example by smoking or eating too much, then they must live with the consequences. Secondly, it normally should apply only to children, because adults can decide for themselves to buy insurance cover before they are struck down with illness or disability. But I accept the state might intervene in some cases, like say for those children born with a disability, either by providing direct subsidies to those affected, or by regulating the

27 Hall, 'Paying for What You Get and Getting What You Pay for: Legal Responses to Consumer-Driven Health Care' 164.

insurance market to enable cross-subsidies. But that sounds complex. Can it be made to work?

***Economist***: Several European systems, including Dutch healthcare, operate with cross-subsidies. The consumers choose their healthcare insurer and insurance package, and consumers who have been assessed as high-risk, high-cost cases are subsidised from a risk equalization fund.[28] Furthermore, The Netherlands is considered to have one of the most successful healthcare systems.[29]

***Clinician***: However, your distinction of people "at fault" and "not at fault" of causing their ill health, and of children and adults, will be very difficult to apply in practice. For example, is an adolescent who is brain damaged after falling from a tree "at fault"?

***Candidate***: I can see difficult policy decisions are required there, but they can be confronted. We have accepted that the state has to intervene to fund public health activities and basic healthcare for individuals, to subsidise (or make the market cross-subsidise) more-than-basic insurance cover for certain individuals who would otherwise, through no fault of their own, have high-cost insurance premiums. Is there anything else the state has to do?

***Lawyer***: It must make laws and policies. You may not agree that health or healthcare is a human right, or with the role of healthcare in securing equality of capability,[30] [31] [32] or even that healthcare has a special moral significance

---

28 Gwyn Bevan and Wynand Van de Ven, 'Choice of providers and mutual healthcare purchasers: can the English National Health Service learn from the Dutch reforms' 5 Health Economics, Policy and Law 343.

29 Karen Davis and others, 'Mirror, mirror on the wall: How the performance of the US health care system compares internationally: 2010 update' (*Commonwealth Fund*, 2010) <http://www.commonwealthfund.org/Publications/Fund-Reports/2010/Jun/Mirror-Mirror-Update.aspx?page=all > accessed 28 March 2012.

30 Amartya Sen, 'Why health equity?' 11 Health Economics 659.

31 Martha C Nussbaum, *Women and human development: The capabilities approach,* vol 3 (Cambridge Univ Press 2001) 77.

32 Cécile Fabre and David Miller, 'Justice and Culture: Rawls, Sen, Nussbaum and O'Neill' 1 Political Studies Rev 4.

because it protects our equal right to opportunity.[33] Nonetheless, I am sure you accept that healthcare is a very important good – more important than motors and mowers and movies. We may die through lack of it. Who should get it and who should decide who gets it? Those are very important questions and because they may be matters of life and death, they are ones that the state cannot ignore. The state must make laws and policies to answer these difficult questions, or at least to explain how, and by whom, these questions are to be answered. For example, our discussion today suggests that we need to decide which public health activities should be funded by the state. Similarly, we have said that basic healthcare will be funded by the state – but how and by whom are decisions made about what constitutes "basic healthcare"? These are complex questions on which people will disagree depending on their values.[34] And if I am ill, who decides in my particular case whether some or all of my treatment fits within whatever has been defined as "basic healthcare"? As my life may depend on it, justice demands an appeals procedure. The state must provide, or regulate to stipulate who provides, for procedural justice.[35]

***Clinician:*** Yes, psychological research shows that procedural justice engenders trust and legitimacy, so that people are prepared to accept decisions as fair even when they go against them.[36]

***Barney:*** My head hurts. It was already starting to sound like socialised medicine. Now you are suggesting British death panels![37]

---

33 Norman Daniels and James E Sabin, *Setting Limits Fairly - Can we Learn to share Medical Resources?*, vol 1 (Second edn, OUP 2007) 14.

34 ibid.

35 ibid.

36 Tom R Tyler, 'Psychological perspectives on legitimacy and legitimation' 57 Annual Review of Psychology 375, 379.

37 Andy Barr, 'Palin doubles down on 'death panels'' (2009) <http://news.yahoo.com/s/politico/20090813/pl_politico/26078>.

***Candidate***: The lawyer is right. There are difficult choices to be made, and because they could be about life and death, the government must either make them, or stipulate who can make them. And individuals must be able to appeal against decisions.

***Bioethicist***: And what about the insurance companies who provide the more-than-basic healthcare cover? Are they the right people to decide what is covered and what is not? Whether a particular cancer drug is covered by my policy could also be the difference between my life and death.

***Economist***: The insurance companies will be able to respond to demand and consumers will be able to choose the insurance cover they want, based on what is included and excluded, and on price.

***Lawyer***: Nevertheless, the importance of these policies justifies regulation of the insurance companies too. How policy cover is decided, and what action I can take to challenge a decision that my treatment is outside the remit of my policy cover - these are questions of public concern. There are also other complex regulatory issues.[38]

***Clinician***: Patients are often particularly vulnerable when seriously ill, open to exploitation by those who might profit from that vulnerability, and thus in need of protection.

***Candidate***: Yes, I accept there is a need for regulation of the insurance market too.

***Lawyer***: And then there are the healthcare providers and clinicians themselves – who is to regulate them? And what about medications and medical devices that are used to treat us?

***Economist***: There is an asymmetry of information at work here and so another type of market failure. The manufacturer knows a lot more than we can about the effectiveness and efficacy of their device or drug. We have relatively little information on which to judge the competence of a doctor and the value of the healthcare that they offer.

---

38 Timothy S Jost and Mark A Hall, 'Role of State Regulation in Consumer-Driven Health Care, The' 31 Am JL & Med 395.

What is more, because of their expertise and authority, the patient is vulnerable to being exploited.[39] For example, the doctor may sell the patient more services than she needs. Nevertheless, the capability and reliability of a doctor, a medical device, or a drug, to deliver a good outcome is of great importance to us as individuals – it could be the difference between life and death. Asymmetry of information is one important reason to regulate.[40]

***Candidate***: An interesting explanation. Few would doubt the importance of regulating healthcare so that we can have competent, qualified clinicians and can trust that medications and medical devices will do us more good than harm. Either it should be self-regulation, overseen by government with the ground rules laid down in law, or it should be state regulation.

***Bioethicist***: You have shown a touching faith in the reliability of the market to provide healthcare. What happens to people if market mechanisms break down and we have no healthcare provided? How then could the state protect our safety?

***Candidate***: I know she has talked about market failure, but I am sure the economist would tell us there is sound theory and empirical evidence that demand for goods and services generates supply. But that is unnecessary because I recognise that in principle to protect the people, the government has a responsibility to ensure that healthcare services are made available. In the unlikely event that the markets were to fail, the government would have to step in, and do something to rectify the problem.

Let us finish here. I am sure there is more that could be said, but I think we have identified the main responsibilities of the state. The state has much to do. It should:

---

39 Robert A Berenson and Christine K Cassel, 'Consumer-driven health care may not be what patients need—caveat emptor' 301 JAMA: The Journal of the American Medical Association 321, 321.

40 ibid.

- Fund and ensure the provision of many activities necessary to protect public health;
- Ensure basic healthcare i.e. lower-cost care that protects safety - is accessible to all, with the state funding either everyone or just those with insufficient means to pay for themselves;
- Subsidise, or make the market cross-subsidise, insurance cover for more-than-basic-healthcare for certain people who would otherwise, through no fault of their own, have high-cost insurance premiums;
- Oversee and ensure that there is continuing provision of a wide range of healthcare;
- Create laws and policies which centre on:
    - how decisions about who is entitled to healthcare are made,
    - systems of procedural justice enabling, for example, appeals by those denied healthcare,
    - the regulatory framework (either self-regulation or state-regulation) governing insurance companies, healthcare professionals, medical devices, and medication.

***Barney***: John, that message is political suicide.

## III. End

The crowd were raucous and rowdy; this was no tea party. People were chanting: "We hate government, we love John!" Many wore T-shirts declaring: "What has the government got to do with healthcare? Nothing!" A woman, presumably from the religious right, held up a sign proclaiming: "John stands firm against Johnnies!"

The candidate stood before his faithful crowd and began: "So...what has the State got to do with healthcare?" A huge roar came from his expectant audience, each one a believer in individual freedom and small government. "My answer is..." Another pause and another roar. "Quite a lot!"

The gasps were audible as shock spread across the faces of the crowd. Barney had his head in his hands. He was already thinking about his next job.

## BIBLIOGRAPHY

### Books

Ariely D, *Predictably Irrational: The Hidden Forces That Shape Our Decisions* (HarperCollins 2009)

Barber BR, *Strong Democracy: Participatory Politics for a New Age* (University of California Press 1984)

Dahl RA, *On Democracy* (Yale Univ Pr 2000)

——, *Polyarchy: Participation and Opposition*, vol 54 (Yale Univ Pr 1971)

——, *A Preface to Democratic Theory* (University of Chicago Press 1971)

Daniels N and Sabin JE, *Setting Limits Fairly - Can We Learn to Share Medical Resources?*, vol 1 (Second edn, Oxford University Press 2007)

Dworkin R, *Is Democracy Possible Here?: Principles for a New Political Debate* (Princeton University Press 2006)

Garoupa N, *Regulation of Professions in the Us and Europe: A Comparative Analysis* (bepress 2004)

Hall MA, *Reforming Private Health Insurance* (American Enterprise Institute 1994)

Herzlinger RE, *Consumer-Driven Health Care: Implications for Providers, Payers, and Policy-Makers* (Jossey-Bass 2004)

Hobbes T, *Leviathan* (Second edn, Cambridge University Press 1996)

Lau DC, *Mencius* (Chinese Univ Pr 2003)

Locke J, *The Second Treatise of Government: And, a Letter Concerning Toleration* (Dover Pubns 2002)

Mill JS, 'On Liberty' in Collini S (ed), *On Liberty and Other Essays* (Cambridge University Press 1989)

Nozick R, *Anarchy, State, and Utopia* (Basic Books 1974)

Nussbaum MC, *Women and Human Development: The Capabilities Approach*, vol 3 (Cambridge Univ Press 2001)

## Journal Articles

Berenson RA and Cassel CK, 'Consumer-Driven Health Care May Not Be What Patients Need—Caveat Emptor' 301 JAMA: The Journal of the American Medical Association 321

Bevan G and Van De Ven W, 'Choice of Providers and Mutual Healthcare Purchasers: Can the English National Health Service Learn from the Dutch Reforms' 5 Health Economics, Policy and Law 343

Carande-Kulis VG, Getzen TE and Thacker SB, 'Public Goods and Externalities: A Research Agenda for Public Health Economics' 13 Journal of Public Health Management and Practice 227

Chen LC, Evans TG and Cash RA, 'Health as a Global Public Good' 1 Global public goods 284

Engelhardt HT, 'Health Care Reform: A Study in Moral Malfeasance' 19 Journal of Medicine and Philosophy 501

Erin CA and Harris J, 'Aids: Ethics, Justice, and Social Policy' 10 Journal of applied philosophy 165

Fabre C and Miller D, 'Justice and Culture: Rawls, Sen, Nussbaum and O'Neill' 1 Political Studies Review 4

Gruskin S and Daniels N, 'Process Is the Point: Justice and Human Rights: Priority Setting and Fair Deliberative Process' 98 American Journal of Public Health 1573

Hall MA, 'Paying for What You Get and Getting What You Pay For: Legal Responses to Consumer-Driven Health Care' 68 Law & Contemp Probs 159

Goodman J, 'Consumer-Directed Health Care' (*SSRN*, 2006) <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=985572#PaperDownload> accessed 18/03/2012

Jost TS and Hall MA, 'Role of State Regulation in Consumer-Driven Health Care, The' 31 Am JL & Med 395

Mckie J and Richardson J, 'The Rule of Rescue' 56 Social Science & Medicine 2407

Saarela MV and others, 'The Compassionate Brain: Humans Detect Intensity of Pain from Another's Face' 17 Cerebral Cortex 230

Sen A, 'Why Health Equity?' 11 Health Economics 659

Sunstein CR and Thaler RH, 'Libertarian Paternalism' 93 American Economic Review 175

——, 'Libertarian Paternalism Is Not an Oxymoron' 70 University of Chicago Law Review 1159

Tversky A and Kahneman D, 'The Framing of Decisions and the Psychology of Choice' 211 Science 453

Tyler TR, 'Psychological Perspectives on Legitimacy and Legitimation' 57 Annual Review of Psychology 375

## Websites

Barr A, 'Palin Doubles Down on 'Death Panels'' (2009) <http://news.yahoo.com/s/politico/20090813/pl_politico/26078>

Davis K and others, 'Mirror, Mirror on the Wall: How the Performance of the Us Health Care System Compares Internationally: 2010 Update' (*Commonwealth Fund*, 2010) <http://www.commonwealthfund.org/Publications/Fund-Reports/2010/Jun/Mirror-Mirror-Update.aspx?page=all > accessed 28 March 2012

Nuffield Council on Bioethics, 'Public Health: Ethical Issues' (2007) <http://www.nuffieldbioethics.org/public-health>